TESIS DOCTORAL

SELECCIÓN DE CARACTERÍSTICAS PARA EL RECONOCIMIENTO DE PATRONES CON DATOS DE ALTA DIMENSIONALIDAD EN FUSIÓN NUCLEAR

Augusto Pereira González

PROGRAMA DE DOCTORADO EN INGENIERIA DE SISTEMAS Y CONTROL

UNIVERSIDAD NACIONAL DE EDUCACION A DISTANCIA
DEPARTAMENTO DE INFORMÁTICA Y AUTOMÁTICA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

DIRECTOR: Dr. D. Jesús Antonio Vega Sánchez
CO-DIRECTOR: Dr. D. Sebastián Dormido Canto

PH.D. THESIS 2015

Agradecimientos

La presente investigación no hubiera sido posible sin la permanente ayuda de mi director principal de tesis, el Dr. Jesús Vega Sánchez. Su dedicación, experiencia, consejos y comentarios me han sabido guiar para que este trabajo pudiera salir satisfactoriamente adelante. De la misma forma, me gustaría también agradecer y destacar la ayuda y el apoyo recibido por parte del Dr. Sebastián Dormido Canto.

No quiero olvidarme del buen hacer de todos los compañeros y colegas de trabajo que, desde el año 2004, han formado parte o integran actualmente la Unidad de Adquisición de Datos del Laboratorio Nacional de Fusión del CIEMAT. De una forma u otra, todos ellos han contribuido y aportado su valiosa colaboración en la realización del presente trabajo.

Resumen general de la tesis

La mejora tecnológica en los sistemas de adquisición de datos, su abaratamiento y la creciente capacidad computacional de los ordenadores, facilita que la dimensionalidad de los datos y su almacenamiento crezca continuamente. Los métodos de reconocimiento de patrones y de selección de características tienen que adaptarse al incremento cada vez mayor de la información adquirida. La reducción de dimensionalidad se convierte en crucial para hacer no solo manejable los datos sino también para poder obtener información inmediata en el análisis de los mismos y su presentación al usuario final.

En esta tesis se propone como principal contribución, un método rápido de selección de características caracterizado por la combinación de técnicas de algoritmos genéticos, clasificación supervisada basada en predictores probabilísticos y una función de comparación fundamentada en la diferencia entre tasas de aciertos y falsos positivos, para determinar los atributos más impactantes y relevantes en la predicción de disrupciones del plasma del dispositivo experimental de fusión termonuclear JET.

A continuación, se facilita una tarea que aporta una solución mejorada en la localización, extracción y recuperación de sub-patrones similares en series temporales de señales digitalizadas. Se proponen búsquedas alternativas y flexibles que incrementan el reconocimiento de formas de onda muy largas. Se presentan estrategias de búsqueda muy precisas en la detección de sub-patrones largos. Las técnicas investigadas facilitan la posibilidad de localizar formas de onda muy similares y con elevada diversidad, independientemente de la longitud y la cantidad de los mismos. Todos estos métodos fueron implementados y están operativos en el estellerator TJ-II y en el tokamak JET por medio de una herramienta de exploración que permite la localización inmediata de formas de onda similares en señales de evolución temporal.

Finalmente y como tercera aportación más relevante, se examinan y solucionan los problemas de comunicación de eventos que se producen durante el transcurso de la operación experimental del TJ-II. Se automatizan tareas de aprendizaje y clasificación que se ejecutan en equipos muy diferentes y remotos. Se aportan técnicas de sincronización entre procesos para plataformas y entornos muy heterogéneos, como pueden ser entre sistemas de tiempo real (OS9, VxWorks), sistemas de tiempo compartido (Unix, Linux) y aplicaciones JAVA y se da solución a las dificultades de sincronización entre procesos remotos que corren en diferentes sistemas operativos y los procesos generados por los servidores de eventos que los distribuyen.

La presente tesis doctoral es el resultado aglutinador de la aplicación de una serie de técnicas originales y procedimientos analíticos relevantes, concernientes a la selección de características y la búsqueda de patrones que se han venido estudiando e investigando en la Unidad de Adquisición de Datos del Laboratorio Nacional de Fusión del CIEMAT y haciendo uso de la enorme cantidad de información disponible en las bases de datos de los dispositivos experimentales de fusión TJ-II y JET.

Índice general

Abreviaturas

Las palabras que aparecen en **negrita** se utilizan para resaltar un término que se considera importante dentro de la tesis. Las palabras en "*letra cursiva*" pueden representar, una expresión en lengua inglesa, un símbolo o una fórmula matemática. Las referencias a cualquier artículo o revista están formadas por el apellido del autor principal, seguido del año de publicación y todo ello incluido entre corchetes en color azul, por ejemplo [Chatterjee, 2006]. Si dicha referencia aparece en color verde [Pereira et al., 2014], significaría que en dicha publicación ha participado el autor de la tesis, bien sea como autor principal o como co-autor en la misma a modo de colaborador, [Vega et al., 2014b]. Las referencias guiadas mediante hipervínculos aparecen resaltadas en color gris.

	Significado	**Uso o procedencia**
ACP	Análisis de Componentes Principales	Método informático
AG	Algoritmos Genéticos	Algoritmo informático
AGA	Algoritmo Genético Adaptado	Algoritmo informático
AGS	Algoritmo Genético Simple	Algoritmo informático
ANOVA	ANalysis Of Variance	Método estadístico
API	Application Programming Interface	Modelo de programación
APODIS	Advanced Predictor Of DISruptions	Clasificador de disrupciones
ASDEX	Axially Symmetric Divertor EXperiment	Dispositivo experimental tipo Tokamak
CIEMAT	Centro de Investigaciones Energéticas Medio Ambientales y Tecnológicas	Centro de investigación español
CMP	Centroide Más Próximo	Taxonomía del centroide más próximo
DIII-D	Doublet III Dshape tokamak	Dispositivo experimental tipo Tokamak
DNS	Domain Name System	Familia de protocolos de Internet
ESRF	European Synchrotron Radiation Facility	Acelerador de electrones
ETB	External Transport Barrier	Confinamiento del plasma
FA	False Alarms	Falsas alarmas o falsos positivos
FN	False Negatives	Falsos negativos. Alarmas perdidas
FP	False Positives	Falsos positivos o falsas alarmas
GPU	Graphics Processing Unit	Circuito electrónico de procesamiento
HL	High Low plasma transition	Transición del plasma
HPC	High-Performance Computing	Computación paralela de alto rendimiento
HTTP	HyperText Transfer Protocol	Protocolo informático
IEEE	Institute of Electrical and Electronics Engineers	Asociación internacional
iid	Independiente e idénticamente distribuido	Aleatoriedad estadística de los datos
IoT	Internet of Things	Término informático
ITB	Internal Transport Barrier	Confinamiento del plasma
ITER	International Thermonuclear Experimental Reactor	Dispositivo experimental tipo Tokamak
JAC	JET Analysis Cluster	Granja de ordenadores en el JET para analizar señales
J2EE	Java 2 Enterprise Edition	Paquete informático
J2EE	Java 2 Enterprise Edition	Librería Java cliente y servidor

JDBC	Java DataBase Connectivity	Modelo de programación
JDBC	Java Data Base Connection	Conector Java hacia una base de datos
JET	Joint European Torus	Dispositivo experimental tipo Tokamak
JMS	Java Message Service	Servicio de mensajería JAVA
JNLP	Java Network Launching Protocol	Especificación para la gestión de distribuciones en JWS
JWS	Java Web Start	Servicio de instalación de aplicaciones JAVA
L/H	Low High	Transición del plasma
MAE	Mean Absolute Error	Media de los errores en valor absoluto
MCC	Mathew's Correlation Coefficient	Coeficiente de correlación de Mathew. Métrica de clasificación
MHD	MagnetoHydroDynamics	Estado del plasma
MPI	Message Passing Interface	Protocolo estándar informático
MSE	Mean Square Error	Error cuadrático medio
MSR	Mean Square Regression	Regresión cuadrática media
NFS	Network File System	Protocolo informático
NMSE	Normalized Mean Square Error	MSE normalizado. Métrica de regresión
NPV	Negative Predictive Values	Valores predictivos negativos. Métrica de clasificación
PAPI	Point of Access to Providers of Information	Infraestructura de autenticación y autorización
PCI	Peripheral Component Interconnect	Bus de datos informático
PLA	Primitivas de Longitud Adaptable	Término informático
PLC	Primitivas de Longitud Constante	Término informático
POJO	Plain Old Java Objects	Modelo de programación
POSIX	Portable Operating System Interface for uniX	Estándar en sistemas operativos
PPV	Positive Predictive Values	Valores predictivos positivos. Métrica de clasificación
PXI	PCI eXtensions for Instrumentation	Estándar bus de datos informático
RBF	Radial Basis Function	Función matemática
RMSE	Root Mean Square Error	Raiz cuadrada de MSE. Métrica de regresión
ROC	Receiver Operating Characteristic	Método estadístico
RR	Regresión Ridge	Regresión estadística
SDEA	Sistema de Distribución de Eventos Asíncronos	Arquitectura informática
SR	Success Rates	Tasas de acierto. Métrica de clasificación
SSE	Sum Square Errors	Suma de los cuadrados de los errores. Métrica de regresión
SSR	Sum Square Regression	Suma de los cuadrados de la regresión. Métrica de regresión
SST	Sum Square Total	Suma total de los cuadrados. Métrica de regresión
SVM	Support Vector Machine	Algoritmo de aprendizaje
T3	Tokamak 3 (Kurchatov Institute)	Dispositivo experimental tipo Tokamak
TCP	Transmission Control Protocol	Protocolo informático
TFD	Transformada de Fourier Discreta	Transformada matemática
TFR	Transformada de Fourier Rápida	Transformada matemática
TFTR	Tokamak Fusion Test Reactor	Dispositivo experimental tipo Tokamak
TJ-II	Torus JEN II.	Dispositivo experimental tipo Stellarator
TN	True Negatives	Verdaderos negativos. Descargas no disruptivas
TP	True Positives	Verdaderos positivos. Descargas disruptivas
URL	Uniform Resource Locator	Formato para identificar recursos en internet
VME	Versa Module Europa bus	Bus de datos informático
VXI	Versa module europa eXtension bus	Bus de datos informático

Símbolos

	Significado	Uso o procedencia
A	Matriz de coeficientes	Regresión
b	Valor del término independiente en una ecuación lineal	Regresión
C	Límites de confianza	Leyes de escala
$\complement_{m,n}$	Combinación de elementos	Combinación de m elementos tomados de n en n
D_α	Señal de radiación emitida por el divertor en su cara interior	Tokamak JET
F	Cociente de significación ANOVA	Métrica de regresión
f	Salida de un clasificador	Valor usado como entrada en regresión logística
H	High	Modo de configuración del plasma
I	Matriz identidad	Estadística
k	Valor del exponente	Leyes de escala
K	Función kernel	Regresión
l	Número de generaciones	En algoritmos genéticos
L	Low	Modo de configuración del plasma
p	p-valor	Predictores conformales
P	Probabilidad	Casos favorables entre casos posibles
Q	Factor de potencia	Potencia producida / Potencia suministrada
q	Parámetro de configuración	Regresión logística
q_{95}	Señal factor de seguridad	Tokamak JET
r	Correlación de Pearson	Medida de correlación
R^2	Coeficiente de determinación	Métrica de regresión
s	Parámetro de regularización Ridge Regression	Regresión
s^2	Error cuadrático medio	Regresión
Sxy	Covarianza	Medida de covarianza
t	Inversa de la distribución Student-t	Estadística
v	Peso de cada característica o atributo para datos normalizados	Regresión
w	Peso de cada característica o atributo para datos brutos	Regresión
x	Vector de características o atributos	Regresión
y	Vector de etiquetas o valores	Regresión
Y	Clases	Usado en los predictores Venn
α	Puntuación de no conformidad	Predictores conformales
δ	Codificación delta, letra minúscula	Diferencia entre valores consecutivos
Δ	Codificación delta, letra mayúscula	Diferencia entre valores consecutivos
Δx	Periodo de muestreo	Distancia entre dos muestras
ε	Nivel de significancia	Predictores conformales
η	Parámetro de la tangente hiperbólica	Estadística
λ	Tamaño de la población	En algoritmos genéticos
σ	Desviación típica	Medida de dispersión
τ	Taxonomía	Predictores Venn

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La mayoría de los dispositivos experimentales de fusión, investigan el comportamiento del plasma, isótopos de hidrógeno ionizados a muy elevadas temperaturas, mediante su confinamiento con trampas magnéticas. Multitud de dispositivos experimentales y diagnósticos, exploran el comportamiento del plasma midiendo magnitudes físicas, presión, temperatura, campos eléctricos, campos magnéticos, radiación emitida, partículas emitidas, etc. Estos sistemas de medida transforman las observaciones físicas en señales eléctricas, que se digitalizan y, las series temporales resultantes se transfieren y almacenan después de finalizar la descarga de operación. La duración de una descarga del estellerator TJ-II está limitada a menos de 500 ms y su base de datos almacena unas 38000 descargas. En el tokamak JET, el mayor dispositivo mundial actualmente en operación, con una duración de descarga de 40 s, sus diagnósticos producen alrededor de 8Gbytes/descarga[1] de datos brutos, habiéndose realizado hasta el momento cerca de 87000 descargas. El dispositivo experimental ITER, que está en fase de construcción, contendrá un volumen de plasma casi 10 veces más grande que el del JET, va a ser un dispositivo de pulso largo (1000 s), y generará millones de señales de muy elevado tamaño, calculándose que puede llegar a alcanzar los 100 Pbytes/año[2] de almacenamiento de datos (1 Tb/descarga según [Greenwald et al., 2005]). Dado el gran volumen de datos con el que nos enfrentamos, la búsqueda de información y su análisis, bien sea, señales completas, imágenes estáticas o patrones característicos dentro de ellos, se hace intratable sin unos mecanismos de acceso a ellos efectivos y eficaces, tanto en la calidad de la información recuperada como en el tiempo empleado en la ejecución de dichas tareas.

La extracción de conocimiento oculto, atributos más relevantes, reconocimiento de patrones, generación de leyes de escala, etc., en bases de datos masivas, requiere el uso de herramientas y técnicas automáticas de algoritmos de aprendizaje que faciliten la generación de modelos predictivos eficientes y con elevado poder interpretativo. Los modelos resultantes, tienen que ser capaces de generalizar con el mínimo error posible frente a nuevas entradas de datos para que sean efectivas no solo, en tareas críticas de control sino también para la identificación y predicción de fenomenología física que atañe a diferentes comportamientos del plasma.

[1] Diagnósticos. Euro-fusion.org.
[2] How to handle the Petabytes. http://www.iter.org/newsline/230/1239

1.1 Motivación y objetivos

Esta investigación se encuadra en el estudio y desarrollo de métodos avanzados de acceso a datos y el procesamiento ágil en bases de datos de dispositivos experimentales de fusión. En este campo existe una problemática clara, que se pretende resolver mediante técnicas novedosas, de alto rendimiento computacional y que puedan ser compartidas por la comunidad internacional. Aborda un problema importante en el desarrollo experimental en grandes instalaciones de fusión nuclear, donde el volumen de datos adquiridos es tan elevado que analizarlos de forma exhaustiva se convierte en una tarea casi imposible, salvo que se encuentren y mejoren, como es el objetivo de esta investigación, técnicas adecuadas para la identificación de patrones importantes y extracción de características relevantes.

Durante la evolución temporal del plasma se observan fenómenos físicos muy característicos y comportamientos análogos que se repiten en diferentes descargas. Las señales e imágenes adquiridas en estas descargas reproducen este comportamiento por medio de formas de onda estructurales y regiones en imágenes muy similares que responden a patrones equivalentes. El análisis de datos requiere la búsqueda automatizada de estas subregiones gráficas, de forma que puedan ser recuperadas de entre todas las descargas que identifiquen a un patrón característico y que evite así el procedimiento manual e intratable de tener que examinar y visualizar cada señal o imagen individualmente. La realización de una herramienta tecnológica que implemente funciones inteligentes de **búsqueda y localización de patrones** y que pueda ser de utilidad a la comunidad científica, ha sido una de las motivaciones y justificaciones principales de la presente tesis.

La búsqueda y selección de características más relevantes haciendo uso de enormes bases de datos, fue otro de los objetivos tratados en esta investigación. En el tokamak ITER, el elevado número de señales y de datos esperados, hará necesario desarrollar métodos novedosos que permitan analizarlos. Es más, existen eventos físicos muy peligrosos en estos dispositivos que deben ser detectados y mitigados, como son las **disrupciones**. Se trata de inestabilidades repentinas y no controladas que originan una pérdida de confinamiento del plasma y que provocan que toda la energía almacenada en el mismo, se transmita violentamente hacia la estructura del dispositivo, pudiendo llegar a ocasionar daños irreparables en sus componentes. La **predicción de disrupciones** es de extrema importancia en JET, el conocimiento que se tiene hasta el momento sobre las disrupciones de este dispositivo es fundamentalmente empírico, se basa en la experiencia pasada de los datos adquiridos por los diagnósticos y no tanto en el entendimiento de la física subyacente de dichas inestabilidades. En ITER, no se puede esperar a la información previa de centenas de descargas para poder extraer modelos significativos y válidos que puedan predecir dichas disrupciones futuras. Es más, la información disponible en los inicios de la operación estará fuertemente sesgada o desbalanceada y el conocimiento a priori será pobre y escaso de juicio diferenciador. Esto es debido a que no se operará a máxima potencia, entre otros muchos motivos, el de intentar evitar disrupciones catastróficas y peligrosas para el dispositivo. Las técnicas ágiles de selección de características y la búsqueda de señales más significativas que están involucradas en la predicción y detección de disrupciones, ha sido otro de los objetivos importantes y que se ha perseguido en esta investigación.

Además, durante el desarrollo de esta tesis doctoral se han generado un conjunto de algoritmos y herramientas software que potencian la técnica de minería de datos y del aprendizaje automático en otros problemas del área de la fusión nuclear, como son las **transiciones** de confinamiento del plasma **L/H**. Existe un umbral de potencia de calentamiento que, una vez superado, mejora el tiempo de confinamiento de la energía y produce un alto gradiente de presión en el borde. El plasma transita desde un estado de bajo confinamiento L, a un estado de alto confinamiento H mejorado. Por sus buenas propiedades de confinamiento, este es el modo de operación que se pretende para los futuros dispositivos, y su identificación y estudio son considerados de gran relevancia en la investigación en fusión.

1.2 Contexto general

Antes de describir detalladamente las principales aportaciones de esta investigación doctoral, se presenta a continuación, el contexto en el que se sitúan los diferentes temas tratados e investigados:

- **Dispositivos experimentales de fusión termonuclear TJ-II, JET**: El dominio de aplicabilidad de la presente tesis discurre en dichos dispositivos experimentales de investigación. Se utiliza información y datos pertenecientes a ellos, bien sea mediante el uso de **señales de evolución temporal** almacenados durante el transcurso de la operación de los mismos, archivos de **imágenes** adquiridos por diferentes diagnósticos o grandes ficheros pertenecientes a películas de **vídeo** que monitorean la radiación del plasma en diferentes espectros electromagnéticos. Así mismo, todos estos datos se utilizan con varias finalidades como son, la optimización de técnicas que mejoren la búsqueda de información compleja y muy abundante, el reconocimiento de patrones o su clasificación y, la predicción de fenomenología concreta del plasma (disrupciones, transiciones de confinamiento, etc). Parte de las investigaciones y técnicas realizadas con estos datos tienen también como finalidad el que puedan ser aplicadas al futuro dispositivo experimental **ITER**, bien sea extrapolando resultados o demostrando su conveniencia de aplicación futura.

- **Procesamiento de datos masivos**: La recopilación y almacenamiento de datos masivos se ha simplificado y abaratado. Hoy en día el almacenamiento de información ya no se considera ni un problema ni un coste. En el campo de la fusión nuclear ocurre lo mismo. Los dispositivos experimentales señalados anteriormente generan cantidades ingentes de datos. El valor de la información ya no reside en estos datos concretos y sí en la forma ingeniosa de cómo se van a correlacionar y procesar para poder descubrir patrones que ni siquiera se habían imaginado ni, por supuesto, buscado de forma intencionada. Bien sea para su estudio y análisis previo o para su uso en tiempo real, dicho procesamiento incluye necesariamente el uso de técnicas optimizadas que puedan reducir la dimensionalidad de las variables implicadas y el tamaño de las enormes bases de datos. Mejoras y aportes en técnicas de **selección y eliminación de características** y de **registros**, la **normalización** de datos, la optimización y mejora de **algoritmos genéticos** o el pre-procesamiento de información mediante funciones matemáticas de transformación (**wavelet**, **fourier**), que reducen los datos de entrada en componentes de tiempo-frecuencia, aportando así si cabe, aún más información desconocida y relevante, sin perder ni descartar al mismo tiempo, la caracterización bruta y el origen de los mismos. Éstas son algunas de las técnicas llevadas a cabo y utilizadas en la presente tesis. Por tanto, la preparación y el acondicionamiento de grandes masas de datos son necesarios para que los métodos de búsqueda a aplicar sean más eficaces tanto en calidad de resultados como en el tiempo empleado para conseguirlo.

- **Aprendizaje en enormes bases de datos**: Si tan importante son las transformaciones y los procesos que se le aplican a los datos, aún lo es más la eficiencia y optimización en el aprendizaje que podemos hacer de ellos. A veces, merece la pena tolerar la imprecisión, la duplicidad, aceptar la inexactitud de los datos, si a cambio se obtiene

un sentido más completo de la realidad y una modelización más generalista de ellos, en otras ocasiones, interesa introducir el nuevo valor pronosticado para mejorar el modelo y la siguiente predicción, quizás otras veces, existan situaciones en los que el aprendizaje deba ser más directo, debido a la enorme cantidad de datos a manipular y no tener que esperar ni demorarse en la obtención inductiva de un modelo concienzudamente entrenado y que ocasione la ralentización de todo el proceso deductivo. Estas estrategias se alcanzan a través de las predicciones aportadas por sistemas inteligentes y algoritmos de aprendizaje, capaces de extraer auténtico significado de la información previa disponible y poder ser aplicada en diferido a grandes bases de datos o en tiempo real a flujos de datos en continuo. El reconocimiento y análisis de **patrones morfológicos**, diferentes algoritmos y **métodos de clasificación** así como funciones de aproximación basadas en **regresión** estadística, son utilizadas en esta investigación. Técnicas basadas en **predicción conformal** también son aplicadas, las cuales incluyen información cualitativa en las predicciones basadas en medidas de confianza y credibilidad, aportando así más datos acerca de cómo de precisos y cuanto de exactos son los pronósticos realizados.

- **Utilización de diferentes tecnologías y entornos de computación**: Las técnicas investigadas, tanto en el procesamiento de los datos, como en el aprendizaje de los mismos, son implementadas en entornos de computación remota. Con ello se proporciona, su usabilidad a la comunidad científica, visibilidad y publicidad a las mismas, así como facilidad para su **portabilidad**. Se desarrollan herramientas tecnológicas que son implementadas en arquitecturas distribuidas y capaces de dar soporte en plataformas muy heterogéneas. Se hace uso de marcos embebidos de software que facilitan la **recuperación masiva de datos** y su **visualización** mediante aplicaciones de usuario gráficas. Se aportan y aplican técnicas de **compresión de datos** y de **sincronización de procesos** y se hace uso de algoritmos que utilizan técnicas de paralelización de datos y que se ejecutan en entornos de **supercomputación de alto rendimiento**, así como otros que dan servicio en **entornos de participación remota** y segura.

1.3 Principales aportaciones de la investigación

Las principales aportaciones realizadas en esta tesis, que han permitido su divulgación a nivel internacional y, en las que se ha participado, bien sea como autor principal o colaborador en las mismas, son las que se recogen en los siguientes trabajos:

1. En [Pereira et al., 2014] se aporta una técnica combinada mediante algoritmos genéticos y clasificadores probabilísticos, capaz de encontrar las características más relevantes involucradas en la predicción de disrupciones del tokamak JET, haciendo uso de 1237 descargas disponibles y 14 señales diferentes, sin necesidad de utilizar procesos de búsqueda combinatorios de fuerza bruta muy costosos computacionalmente. Se evalúan y exploran diferentes medidas de rendimiento. Se concluye que es realmente importante el uso de funciones de ajuste ponderadas para evaluar las tasas de aciertos y poder descubrir las variables predictoras más significativas en el menor tiempo posible de computación. Este trabajo se fundamenta en otro previo y en el cual también se ha colaborado, [Vega et al., 2014], donde se utiliza por primera vez los clasificadores probabilísticos mediante predictores Venn aplicados a la predicción de disrupciones y con elevadas tasas de acierto, caracterizándose éstos por aportar información añadida a la clasificación acerca del grado o exactitud con que se determinan dichas predicciones.

2. En la publicación [Pereira et al., 2010b] se facilita y aporta una solución mejorada en la localización y extracción de sub-patrones similares en series temporales de señales digitalizadas. Se proponen búsquedas alternativas y flexibles que incrementan el reconocimiento de formas de onda muy largas. Se presentan estrategias de búsqueda más finas en la detección de sub-patrones largos, debido a que éstos son más difíciles de identificar. Estas técnicas, facilitan la posibilidad de localizar sub-patrones de formas de onda muy similares, independientemente de la longitud de los mismos. Todos estos métodos fueron implementados y están operativos en el TJ-II y en el JET por medio de una herramienta que permite la localización de formas de onda similares en señales de evolución temporal. En [Vega et al., 2009], se resumen muchas de las técnicas empleadas, tanto las aplicadas a series temporales de formas de onda, como las también llevadas a cabo y las investigadas dentro de imágenes [Vega et al., 2008b]. Inicialmente, métodos de búsqueda automáticos de señales para formas de onda completas se publicaron en [Vega et al., 2008] y aplicadas al TJ-II, aportándose aquí una interfaz de usuario gráfica para la selección y visualización de señales. Seguidamente, se atacó el problema de la búsqueda de sub-patrones más pequeños dentro de señales, aportándose varias metodologías que fueron contrastadas y recogidas en [Dormido-Canto et al., 2008], e implementadas mediante varias aplicaciones, instaladas en el JET y dadas a conocer en la publicación [Vega et al., 2008c]. Así mismo, dichas herramientas se utilizaron en [Rattá et al., 2008] para la investigación de fenomenología física mediante el estudio de patrones morfológicos con señales del

JET. Parte de todos estos trabajos y resultados fueron también recopilados y explicados en [Vega et al., 2007].

3. En la publicación [Pereira et al., 2010] se da a conocer la implementación de una herramienta software multipropósito, no solo para señales del JET, multiplataforma y válido para sistemas Windows, Linux, Mac. Se hace uso de una arquitectura distribuida, en un entorno de participación remota mediante internet, para dar visibilidad, complementar y publicitar el trabajo principal realizado en el punto anterior. Se hace uso de marcos embebidos de software que evitan configuraciones e instalaciones complejas y ajenas de otros servicios que son muy necesarios en la utilización de la herramienta. Un motor de base de datos y un servidor web son encapsulados en una única aplicación de escritorio tanto para la parte cliente como para la parte servidora. Una dificultad importante a solventar en este trabajo fue la recuperación de grandes cantidades de información y el envío de las señales brutas originales para poder ser visualizadas entre diferentes ordenadores muy distantes y remotos. Para ello, se han aplicado diferentes técnicas de compresión de datos sin pérdida de información, recopiladas en [Vega et al., 2007b] y que algunas fueron revisadas para su utilización en tareas de tiempo real y orientadas a descargas de pulso largo, consiguiendo migrar satisfactoriamente dichos algoritmos hacia librerías JAVA. De esta manera, el envío de todas las señales se realiza de forma comprimida y en la recepción de los mismos es descomprimida por los diferentes clientes remotos, haciendo así más transparente y ágil el trasiego de grandes cantidades de información.

4. En [Pereira et al., 2006] se aportan técnicas de sincronización de procesos para plataformas y entornos muy heterogéneos, como pueden ser entre sistemas de tiempo real (OS9, VxWorks), sistemas de tiempo compartido (Unix, Linux) y aplicaciones JAVA y se da solución a las dificultades de sincronización entre procesos remotos que corren en diferentes sistemas operativos y los generadores por los servidores de eventos que los distribuyen. El objetivo principal es que estos subprocesos locales, lanzados en máquinas muy diferentes, puedan darse cuenta de los sucesos que se generan en un entorno de red de área local, sin tener que estar preguntando por ellos constantemente. Se facilitan librerías de propósito general que aportan sincronización no solo entre procesos remotos, sino también sincronización entre procesos locales en una misma computadora. Por ejemplo, se han utilizado para sincronizar programas de usuario de algoritmos de aprendizaje, ejecutados en entornos escalables de supercomputación basados en Linux y que tienen que esperar por la finalización de otras aplicaciones locales para que puedan ser lanzados nuevamente en ciclos repetitivos. Estas técnicas se han utilizado igualmente en la actualización que se hizo del sistema de análisis automático para el reconocimiento de imágenes del diagnóstico de esparcimiento Thomson del TJ-II, que se publicó en [Makili et al., 2010], basado a su vez en un trabajo previo, utilizando técnicas de clasificación inteligentes mediante máquinas de vectores soporte (SVM, siglas procedentes del inglés) y publicado en [Vega et al., 2005]. En estos dos trabajos se hace uso del sistema de sincronización, para automatizar el proceso de clasificación de imágenes entre una máquina Solaris-Unix y el sistema de distribución de eventos asíncronos de la red de área local del TJ-II. Igualmente y para estas publicaciones, se contribuyó también con la implementación cliente del sistema de control del reconocimiento de patrones, una herramienta en diferido para el ordenamiento de las imágenes y otra más de depuración para las imágenes

clasificadas erróneamente por el sistema. Del mismo modo, en [Sánchez et al., 2006] se aporta una aplicación de mensajería basada en JAVA y que utiliza también el presente sistema de sincronización de eventos, viniendo así a complementar el seguimiento de la operación en el entorno del TJ-II, que se fundamenta sobre todo en servidores web y aplicaciones de visualización JAVA que hacen uso de este servicio de mensajería.

1.4 Otras contribuciones originales

Otras contribuciones aportadas muy relevantes, relacionadas con temas de la presente investigación, presentadas también a la comunidad científica internacional y en las que el autor de la tesis ha participado en diferentes trabajos pero solamente como coautor de las mismas, aportando su colaboración en diversos proyectos de investigación en fusión, han sido las siguientes publicaciones:

5. En [Vega et al., 2014b] se contribuye con unos análisis estadísticos preliminares sobre la física subyacente en el estudio de las disrupciones del JET. En esta publicación, se recoge también la evolución en el tiempo y un resumen de todas las aportaciones llevadas a cabo acerca de la predicción de disrupciones y en las que también se ha colaborado; por ejemplo, en [Dormido-Canto et al., 2013] por primera vez se desarrolla un clasificador de disrupciones '*from scratch*' o sea empezando desde el principio, con un solo ejemplo de clase disruptiva y otro de clase no disruptiva, las descargas son procesadas en orden cronológico y el clasificador tiene que ir aprendiendo sin apenas información relevante desde sus inicios. En [Vega et al., 2013c] se analizan los resultados obtenidos mediante el predictor APODIS en la red de tiempo real de JET pero aportando líneas de investigación de cara a ITER y en [Vega et al., 2013b] se aportan técnicas avanzadas de análisis de datos, no sólo las llevadas a cabo en los análisis de datos disruptivos sino también los concernientes y aplicados al estudio de transiciones de confinamiento. Los estudios más relevantes realizados en estos trabajos por el autor de la tesis fueron la posibilidad de aplicar clasificadores bayesianos con técnicas de aprendizaje activo en la predicción de disrupciones y el estudio de la combinación de clasificadores basados en regresión logística con otros basados en SVM haciendo uso del predictor APODIS.

6. Siguiendo con el estudio de la transición L/H en el Tokamak JET, diversas publicaciones se han presentado en las que se ha participado como colaborador. La colaboración más relevante realizada en [González et al., 2012], consistió en la estimación paramétrica del umbral de potencia en el instante de la transición L/H, haciendo uso de diferentes variables explicativas, aplicadas a distintos métodos de regresión y comparando los puntos del instante real de la transición con el obtenido automáticamente por el clasificador. Este clasificador, basado en SVM con predicción conformal, fue aplicado también en [González et al., 2012d], precisamente para estimar la frontera de separación de la transición H/L con aportaciones añadidas a la clasificación de valores de confianza y credibilidad. Anteriormente se presentó en [González et al., 2010] un trabajo que se basa en la extracción de características o señales más importantes que están involucradas en la clasificación para el instante de la transición L/H haciendo uso de un hiperplano SVM lineal. Se aplica un método novedoso para la selección de características mediante la eliminación de los coeficientes con menor peso estadístico en la

ecuación paramétrica de dicho hiperplano. Se contribuye en este trabajo con la implementación de un algoritmo que extrae dichos coeficientes significativos a partir del modelo SVM generado en el proceso de entrenamiento. Esta aportación también se utilizó en [Farias et al., 2012], aplicando el mismo método de selección de características a un algoritmo de reconocimiento de patrones morfológico y un modelo multicapa basado en SVM, con el objetivo igualmente de determinar los instantes de tiempo de la transición L/H en el tokamak DIII-D. Contribuciones menores concernientes al análisis de datos y la reducción de dimensionalidad aplicando la transformada wavelet se realizaron en las publicaciones [González et al., 2010b] y [González et al., 2012c].

7. La usabilidad de predictores conformales es de gran importancia en problemas, no solo de clasificación, sino también en problemas de regresión, sobre todo cuando se presentan dominios donde existen errores, tanto en los datos como en sus predicciones. En temas de clasificación, los predictores establecen niveles de confianza y credibilidad, aportando información añadida a dicha predicción. Para temas de regresión, se establece una barra de error estimada, que es tanto más grande, cuanto mayor es la confianza en el ajuste de la predicción. En la publicación [Vega et al., 2012], no solo se sintetizan las aplicaciones realizadas utilizando dichos predictores conformales, sino que se contribuye con una implementación conformal y no paramétrica para ajustes de regresión. Predicciones multiclase con un alto nivel de fiabilidad y significancia se aplicaron también en [Vega et al., 2010] en modo diferido, al reconocimiento de patrones en imágenes, pertenecientes al diagnóstico de esparcimiento Thomson del estellerator TJ-II y utilizando el algoritmo del vecino más próximo de forma inductiva y conformal. Con el mismo propósito de clasificación de imágenes, pero utilizando medidas de no conformidad mediante el algoritmo SVM, se colaboró en [González et al., 2012b] y en [Vega et al., 2013] para un estudio en la localización espacial de perturbaciones producidas en plasmas del JET.

8. Una aplicación de monitorización, perteneciente a un sistema de distribución de datos en tiempo real, se aportó en el trabajo [Castro et al., 2010b], basado en conceptos de experimentación remota, instalado para un diagnóstico del JET y orientado a descargas de pulso largo. En el sistema de participación remota del TJ-II se colaboró también en varios trabajos, parte de los mismos fueron recopilados en la publicación [Vega et al., 2005b], el más relevante consistió en la implementación de una aplicación de mensajería, englobado dentro de una arquitectura orientada a mensajes, que funciona mediante un protocolo de publicación-subscripción y dado a conocer en [Sánchez et al., 2007b]. Las aplicaciones de participación remota instaladas para controlar el sistema de adquisición de datos del TJ-II han permitido comandar y seguir la operación de descargas del stellerator TJ-II desde Cadarache en Francia, dicho seguimiento fue publicado en [Vega et al., 2006], y en la que también se ha participado. En [Sánchez et al., 2008] se participó en el diseño e implementación de una base de datos orientada a eventos en el entorno de la red de área local del TJ-II y en [Castro et al., 2009] en una arquitectura de distribución de datos en tiempo real. Finalmente, se aportaron pequeñas tareas menores de gestión y configuración realizadas en la integración del sistema de participación remota del TJ-II como identidad federada de autenticación y autorización entre diferentes laboratorios europeos de investigación en Fusión [Castro et al., 2008b], [Castro et al., 2010].

1.5 Organización de la tesis

La memoria de la presente tesis está dividida en ocho apartados. El tema primero abarca la introducción de la misma. Seguidamente, se hace un repaso a los temas más novedosos sobre la investigación en fusión nuclear. Los temas tres y cuatro están dedicados a la fundamentación teórica de los métodos y desarrollos aplicados, tanto los relacionados con el reconocimiento de patrones, como los utilizados para la selección de características y metodologías de evaluación en algoritmos de aprendizaje. Los temas cinco, seis y siete corresponden a la implementación práctica de lo investigado en la tesis, con aplicación a datos del TJ-II y del JET. Esto es, la optimización de recursos en procesos de aprendizaje, aplicación práctica del reconocimiento morfológico de señales y la búsqueda de los atributos más relevantes en la predicción de disrupciones. Finalmente, en el tema ocho, se exponen las conclusiones finales más relevantes e importantes de la presente tesis doctoral. Para los temas cinco, seis y siete se van exponiendo conclusiones parciales y muy específicas del tema investigado y un resumen de las publicaciones donde fueron presentados dichos trabajos. En un anexo adjunto, se muestran los algoritmos utilizados para la obtención de la ecuación del hiperplano SVM a partir de un modelo de entrenamiento lineal generado por el software *libsvm*. El anexo principal de la memoria, recoge la recopilación completa de todas las publicaciones científicas, revistas de divulgación, informes técnicos y aportaciones a congresos internacionales.

Capítulo 2

La investigación en fusión nuclear

Probablemente, la aplicación práctica más conocida de la energía nuclear es la generación de energía eléctrica para su uso civil, en particular mediante la fisión de átomos de uranio enriquecido. En un reactor nuclear de fisión se colocan barras de uranio dentro de masas de bloques de grafito o de cadmio. Éstos hacen de moderadores en la reacción y controladamente, los neutrones emitidos por el material radiactivo, crean nuevas fisiones en los otros átomos de uranio circundantes mediante una reacción en cadena, desprendiéndose calor en dicho proceso. Esta energía se utiliza para calentar agua, cuyo vapor resultante se aprovecha para mover una turbina y así poder producir energía eléctrica. El concepto físico y su aplicación técnica son muy simples. Enrico Fermi en 1942 construyó el primer reactor de fisión y desde entonces se sigue utilizando con notable éxito. La fusión nuclear es infinitamente más compleja y problemática, no en su concepción teórica pero sí en su aplicación técnica. De hecho, se requirió más de medio siglo para que un dispositivo experimental, que no reactor, el británico JET, pudiera producir energía solamente durante un segundo. La energía que se produce en las reacciones de fusión se libera cuando dos iones ligeros se fusionan, ésta se produce de forma natural en las estrellas a una temperatura de 10 millones de grados centígrados. Su combustible, el plasma de hidrógeno, es confinado mediante la fuerza de la gravedad. En la Tierra, la reacción más sencilla se produce entre iones de dos tipos de hidrógeno como son el deuterio y el tritio. El confinamiento debe ser logrado mediante mecanismos más complejos, requiriéndose unas condiciones muy difíciles, como son, una densidad de materia del orden de 1 mg/m^3, una temperatura de cerca de 100 millones de grados centígrados y un tiempo de confinamiento de algunos segundos.

La fusión por confinamiento magnético utiliza un plasma caliente confinado por campos magnéticos, esto es, las partículas eléctricas, iones y electrones, giran y permanecen atrapadas alrededor de las líneas de campo magnético toroidal generado por diferentes solenoides situados alrededor de la cámara de vacío que contiene el plasma y, para ciertas clases de dispositivos, como son los **tokamaks**, mediante corrientes eléctricas inducidas que circulan por el interior del propio plasma confinado. Para que el plasma caliente no se dilate y entre en contacto con la cámara que lo recubre, se añaden campos poloidales colocados tanto en la parte superior como en la inferior del dispositivo y a lo largo de la dirección axial del plasma. Si el campo toroidal y poloidal son generados conjuntamente y de forma totalmente externa entonces hablamos de dispositivos experimentales de tipo **estellerator**. Una de las características más importantes del dispositivo español TJ-II, el mayor estellerator actualmente en funcionamiento en Europa, es su configuración helicoidal flexible. Su transformada rotacional, que es el ángulo promedio en dirección poloidal girado por las líneas de campo de una superficie magnética determinada, al dar una vuelta alrededor del dispositivo, puede variar entre 0.9 y 2.2, obteniendo volúmenes de plasma desde 0.3 m^3 hasta 1.2 m^3. Esta flexibilidad proporciona a los científicos un amplio rango de diferentes configuraciones para poder desarrollar sus experimentos[3].

Inicialmente, en un dispositivo tokamak, el plasma se calienta debido a la corriente eléctrica que circula por él. Sin embargo, a medida que la temperatura aumenta, la resistencia del plasma disminuye, lo que reduce la eficacia del calentamiento. Cuando la temperatura del plasma alcanza 10 millones de grados, los electrones se mueven tan rápidamente en ese medio poco denso que dejan de interactuar con los iones. El efecto Joule que resulta de las colisiones entre electrones e iones desaparece. Se puede suponer entonces que el medio se vuelve casi superconductor. La temperatura del plasma se satura. Se necesitan entonces sistemas adicionales de calentamiento adicional, como la inyección de haces de neutros y el calentamiento por radiofrecuencia, para permitir alcanzar las temperaturas termonucleares que producen la fusión de los núcleos. El funcionamiento de un dispositivo experimental de fusión precisa además de otros muchos sistemas auxiliares muy complejos para poder controlar la reacción del plasma y que permitan su funcionamiento en régimen continuo y estacionario.

Se sabe desde los comienzos de la investigación en fusión nuclear, que los plasmas a alta temperatura que se tratan de confinar usando campos magnéticos son terriblemente inestables y están sujetos a inestabilidades magneto-hidrodinámicas (MHD). Esta disciplina de la física, estudia la dinámica de fluidos conductores de electricidad en presencia de campos eléctricos y magnéticos. Se trata, en efecto, de mecanismos disipativos por medio de los cuales un sistema trata de expulsar la energía que contiene, facilitando su transporte. Por ejemplo, las corrientes de convección en una olla de agua caliente, la formación de corrientes ascendentes en la atmósfera, los torbellinos marginales en la punta de las alas de los aviones y los regímenes turbulentos en fluidos líquidos, son fenómenos disipativos similares pero con la diferencia de que en todos ellos no aparece la interacción de la fuerza electromagnética. En los plasmas, los problemas se vuelven tremendamente más complejos por el hecho de que regiones distantes del dispositivo quedan instantáneamente acopladas por el campo electromagnético. En ingeniería aeronáutica en cambio, cuando se crea una turbulencia en alguna parte externa del ala de un avión, parte de ella se disipa automáticamente al medio ambiente gaseoso externo del mismo y su incidencia y propagación al resto del aparato es muy débil y atenuada. En dinámica de fluidos, cuando se quiere modelar el comportamiento de una partícula dentro de un fluido en estado líquido deben de tomarse en cuenta seis

[3] Overview of TJ-II experiments. [Sánchez et al., 2007], [Sánchez et al., 2009], [Sánchez et al., 2011], [Sánchez et al., 2013], [Hidalgo et al., 2005]

parámetros, tres para la posición y tres para la velocidad. Dichas partículas, en consecuencia, viven en un espacio de seis dimensiones sometidas a la interacción mecánica de otras partículas y se hace realmente difícil e impreciso predecir su comportamiento temporal cuando pasamos de un régimen laminar a otro régimen turbulento de mayor energía. En física de plasmas, las interacciones mecánicas entre partículas no solo son más numerosas y energéticas sino que aparecen otras más, como pueden ser las fuerzas columbianas de atracción y repulsión sometidas a muy elevadas temperaturas. Teóricamente, este medio de transporte casi caótico, debería de poder ser descrito mediante un sistema de ecuaciones diferenciales de Boltzmann[4] fuertemente acopladas por campos electromagnéticos, derivando en las ecuaciones de Navier-Stokes y de Maxwell [Alfven, 1942], pero la posibilidad de hacer simulaciones numéricas de millones de partículas para predecir el comportamiento global del plasma, en un sistema de interacción tan endiabladamente complejo, se hace intratable. Inclusive, bajo el supuesto de considerar el plasma virtualmente en constante equilibrio termodinámico y con interacciones no cuánticas entre sus partículas. Las actuales leyes de escala que modelan el comportamiento del plasma son puramente empíricas y están basadas en la experiencia adquirida, observada y estudiada en otros dispositivos experimentales y solamente pueden ser contrastadas de forma aproximada con los conocimientos teóricos y simulados de la física real que los modela. En particular, no existe ningún modelo empírico del comportamiento del plasma en un dispositivo funcionando en régimen estacionario y en ignición, estado en el cual la energía del plasma es auto-mantenida sin aporte externo, y por tanto, ninguno demostrado ni contrastado que permita realizar extrapolaciones en ese sentido.

El proyecto más avanzado en fusión nuclear por confinamiento magnético es el ITER (siglas procedentes de su acepción inglesa, International Thermonuclear Experimental Reactor), prototipo basado en el concepto tokamak, y en el que se espera alcanzar la ignición. En la máquina ITER no se producirá energía eléctrica, se probarán las soluciones a los muchos problemas que necesitan ser resueltos para hacer viables los futuros reactores de fusión nuclear. Producir más energía de la que se suministra, a sabiendas de que calentar un plasma es costoso energéticamente, es otro de los objetivos marcados para ITER. El cociente, potencia térmica producida entre potencia suministrada, se designa por la letra Q y en ITER se espera que éste tenga un valor entre 5 y 10. En JET[5], se consiguió en 1997 un factor $Q = 0.6$ produciendo $16\ MW$ de potencia y habiéndose empleado para ello $24\ MW$ de potencia suministrada. No obstante, desde que en 1968 el tokamak T3 de la URSS mostrase los primeros plasmas calientes, se ha mejorado la ganancia Q en un factor de 10000, demostrando así que, aun estando lejos de conseguir un reactor nuclear produciendo energía, si se han mejorado muchos conceptos y avances tanto en el campo de la ingeniería como en el de la física. Cuando se tiene un funcionamiento con un factor Q mucho mayor que la unidad, conocido en lengua inglesa como *breakeven*, la producción de energía por fusión aparece. Nada se sabe de cómo se comportará el plasma en una máquina en la que se den esas condiciones y durante muchos minutos de confinamiento. La combustión de un plasma turbulento, fusionando sus núcleos en régimen continuo y en presencia de fuertes campos magnéticos todavía no se ha presenciado y supone un verdadero desafío a ser afrontado e investigado.

[4] Boltzmann equation. (Pag. 39), [Miyamoto, 2011]
[5] https://www.euro-fusion.org/wpcms/wp-content/uploads/2012/01/jeteuropeansucess.pdf

2.1 El grave problema de las disrupciones

Cuando ocurre una disrupción[6], la temperatura del plasma cae extremadamente rápido, en pocas milésimas de segundo, hasta en un factor de 10000, pasando de 100 millones de grados centígrados a sólo unas decenas de miles de grados. La energía se disipa por conducción térmica turbulenta en las paredes y por radiación. La pérdida de confinamiento durante una disrupción causa que toda la energía almacenada en el plasma, tanto térmica como magnética, se pierda repentinamente. Debido a una temperatura tan baja, el plasma se vuelve resistivo. El efecto Joule reaparece. Típicamente, la energía se deposita en el desviador y en la primera pared del tokamak, lo que puede originar grandes flujos de energía sobre estas superficies, pudiendo llegar a derretirse o evaporarse. Por tanto, la motivación principal para estudiar las disrupciones y su mitigación, es el efecto secundario y dañino que pueden tener sobre los componentes estructurales del dispositivo.

La pérdida de confinamiento lleva a una pérdida de la corriente de plasma. La rápida supresión de la corriente produce corrientes inducidas en el recinto de vacío del tokamak. Además de la corriente inducida en el recinto del tokamak, si hay contacto entre el plasma y las paredes del recinto entonces la corriente que fluye en el plasma circulará a través de las paredes conductoras del recinto. Las corrientes que fluyen en las paredes, conocidas como corrientes en halo[7], interactúan con el campo toroidal y dan lugar a esfuerzos de Laplace que se transmiten a la estructuras mecánicas de todo el dispositivo. La rápida supresión de la corriente genera también una fuerza electromotriz inducida que puede servir para acelerar los electrones en el plasma hasta energías relativistas. Estos electrones de alta energía, conocidos como electrones fugitivos o desacoplados, cuya intensidad es del mismo orden de magnitud que la corriente de plasma, pueden llevar a la producción de rayos X cuando el haz de electrones interactúa con los componentes de la cámara de vacío.

Para poder operar futuros tokamaks, como es el ITER, en buenas condiciones de confiabilidad, seguridad, integridad y desempeño[8], se hace cada vez más necesario controlar las disrupciones del plasma. Puesto que el contenido energético de los futuros tokamaks y reactores es varios órdenes de magnitud mayor que el de las máquinas actuales, las consecuencias de las disrupciones pueden llegar a ser mucho más graves y peligrosas.

Las causas de las disrupciones son múltiples y variadas, consistiendo a menudo en una secuencia de eventos[9] tales como, incremento moderado de la actividad MHD y de la densidad dentro de una fase pre-disruptiva, aumento abrupto de la actividad MHD y de la potencia radiada que conlleva una caída repentina de la temperatura del plasma, en lo que se conoce como enfriamiento térmico y seguidamente, caída repentina de la corriente del plasma junto con un desplazamiento vertical del mismo, reflejando tasas más altas de

[6] Instabilities. (Pag. 96), [McCracken, 2005]
[7] Halo currents. (Pag. 58), [Reux, 2010]
[8] Disruption researchers investigate design options. www.iter.org/newsline/252/1448
[9] Physics of a disruption. (Pag. 20-21), [Thornton, 2011]

actividad MHD a medida que se reduce paulatinamente la potencia radiada y la densidad. El enfriamiento repentino del plasma es la fase más peligrosa, debido a que toda esa energía térmica que contenía el plasma debe ser expulsada, transportada y depositada en algún otro lugar, que no es otro que la primera pared de la cámara de vacío del dispositivo experimental. Se hace indispensable el poder detectar las disrupciones a tiempo en la fase pre-disruptiva, con el objeto de poder aplicar acciones mitigadoras y/o correctoras para que dichas disrupciones no lleguen a desarrollarse.

2.2 Identificación de eventos físicos relevantes

Desde el inicio de una descarga hasta el final de la misma, diferentes eventos tienen lugar durante la operación del plasma. Algunos son muy peligrosos, como las disrupciones térmicas y de confinamiento explicadas anteriormente, pero hay otros fenómenos físicos que benefician el transcurso de la operación y son relevantes para mejorar el tiempo de confinamiento y las propiedades físicas del plasma. La identificación correcta y precisa de estos fenómenos es un tema de especial relevancia en los dispositivos actuales y también de cara a ITER.

Como ya se ha descrito anteriormente, el confinamiento en dispositivos por confinamiento magnético está dominado por el denominado transporte anómalo de origen turbulento. Existen modos de confinamiento mejorado que multiplican por factores de hasta 4 veces el tiempo de confinamiento. En 1982 el Tokamak ASDEX (Garching, Alemania), descubrió[10] el modo H, basado en una barrera de transporte situada en el borde (conocido como "pedestal" o ETB, por sus siglas en inglés), que da lugar a una mejora del confinamiento en un factor dos. Durante el calentamiento del plasma, mientras la válvula del gas estaba cerrada, se encontró que había un incremento de la densidad y temperatura, elevando también el gradiente de presión, causado por una repentina mejora del confinamiento de las partículas. En contradicción con predicciones de modelos de transporte basados en colisiones de Coulomb ya que el tiempo de confinamiento se reduce al incrementar la temperatura del plasma (modo L). Este fenómeno era completamente inesperado dado que no había sido predicho por ninguno de los modelos teóricos. La transición al modo H aparece por una cizalladura en la rotación del plasma que rompe las estructuras turbulentas creando así la barrera y actualmente es el escenario principal de operación previsto para ITER[11]. Tan pronto como se entendió el modo H se planteó la posibilidad de provocar barreras adicionales más hacia el interior del plasma. Esto se consiguió en los dispositivos americanos DIII-D y TFTR (San Diego, Princeton) en los años 1995 y 1996. Controlando el perfil de corriente del plasma se lograba una inversión del perfil radial de iota o transformada rotacional y, aproximadamente en la zona del máximo de iota, aparecía una barrera interna de transporte (ITB, del inglés) que daba lugar a abruptos gradientes y a una mejora significativa del confinamiento con factores de 3-4 veces. Los escenarios con ITB producen excelentes valores de confinamiento y son muy propicios para mantener corrientes autogeneradas muy intensas, no procedentes del transformador. La desventaja principal es que están limitados a densidades mucho más bajas que las obtenidas en el modo H. No obstante, aunque las ventajas del modo H son extremadamente importantes para alcanzar un reactor de fusión, también tiene algunos aspectos negativos a solventar como son, el incontrolado incremento de la densidad y la generación de impurezas que se depositan en el plasma. La operación a alta densidad también es un requisito básico para un reactor, pero los tokamak tienden a entrar en disrupción cuando la densidad supera un cierto límite, debido a un enfriamiento del borde como consecuencia de que el gas inyectado tiene problemas

[10] How Fritz Wagner "discovered" the H-Mode. http://www.iter.org/newsline/86/659
[11] Confinamiento. (Pag. 215), [García et al., 2001]

para alcanzar el centro del plasma, provocando una contracción del mismo y la desestabilización repentina de modos MHD. Por tanto, se hace imprescindible la coordinación y el control de todos estos eventos y regímenes de confinamiento para poder conducir el plasma sin ningún contratiempo hasta las condiciones de ignición.

2.3 Sistemas de control, diagnósticos y adquisición de datos

El control de la reacción de fusión, su optimización y los estudios científicos necesarios para entender y predecir el comportamiento de los plasmas, requieren el conocimiento de multitud de parámetros de los mismos: temperaturas y densidades, velocidad del plasma y campos eléctricos, parámetros de configuración (corriente, voltaje, geometría de las superficies de flujo, campos magnéticos), composición del plasma (impurezas, pérdidas por radiación, etc.), condiciones del borde (partículas, deposición de material, etc.).

Para la medida de todos estos parámetros se utilizan un conjunto de sensores llamados diagnósticos. Estos dispositivos traducen la magnitud física o variable de instrumentación en una tensión eléctrica. Se requiere posteriormente una etapa de acondicionamiento, que adecua la señal analógica a niveles compatibles con el elemento que hace la transformación a señal digital. El elemento que hace dicha transformación es el módulo de digitalización o tarjeta hardware de medición. La adquisición de datos, consiste en la toma de estas señales, su acondicionamiento y transformación para que puedan ser manipulados y almacenados en un ordenador. Otras tarjetas hardware de control realizan el proceso inverso, transforman una señal digital en una tensión eléctrica analógica para poder comandar diversos actuadores de instrumentación (eléctricos, neumáticos, hidráulicos, etc.). Un ejemplo básico de control, mediante hardware de medición y actuación, similares a las utilizadas en el dispositivo TJ-II, interactuando con un transductor de posición en tiempo real se facilita en [Pereira, 2001]. El concepto tiempo real también es importante. Comúnmente se utiliza para definir a un sistema que debe ejecutarse sin fallo en un intervalo de tiempo acotado. El error más común es pensar que tiempo real significa en realidad, rápido; cuando de hecho, muchas aplicaciones de adquisición de datos y control tienen ciclos muy lentos, pero deben de responder de forma segura en el tiempo en que se han programado, independientemente de los eventos secundarios que se produzcan en el sistema. En sistemas operativos de tiempo real, las interrupciones y eventos son jerarquizados y los eventos con la mayor prioridad se ejecutan antes que los eventos de prioridad menor. En las grandes instalaciones científicas se utilizan estos sistemas operativos para poder realizar desarrollos de control no estándares que garantizan respuestas del sistema ante eventos inesperados y con capacidad de maniobra y actuación.

En todas las grandes instalaciones de investigación existe multitud de componentes mecánicos e instrumentación científica. Se necesita poder actuar en tiempo real y medir con precisión y exactitud muchas señales frente a entornos electromagnéticos muy hostiles. En la contribución que se hizo al diseño del sistema de control y adquisición de datos de la línea española del sincrotrón ESRF (Grenoble, Francia) [Pereira et al., 2004], [Pereira et al., 2005b], más de 200 canales de medida y de control permiten de forma precisa el acondicionamiento del haz incidente de rayos X (45 keV de energía), actuando sobre diferentes elementos ópticos como son espejos, rendijas, atenuadores, etc, y

permitiendo el control de otros elementos experimentales, difractómetros, monocromadores, analizadores, de forma remota y segura.

En los dispositivos de fusión por confinamiento magnético, la medida se realiza en condiciones, si cabe de más dificultad: altísimas temperaturas de plasma (no todos los sensores aguantan dichas temperaturas), dificultad de acceso (grandes cámaras de vacío, sistemas de bobinas) y elevados campos magnéticos (perturbaciones y mucho ruido en las medidas). En estas condiciones, los diagnósticos deben de explorar el comportamiento del plasma intentando medir en todo el espectro electromagnético. Por ejemplo, el diagnóstico de esparcimiento Thomson del Stellarator TJ-II es utilizado como el sistema básico para la medición de la densidad y de la temperatura electrónicas locales en dicho plasma, en el rango de la radiación visible, basándose en la medida de la frecuencia e intensidad de la luz dispersada de un láser de alta potencia que atraviesa el plasma. En el tokamak JET, existen más de 90 diagnósticos funcionando y hay más de 20 en fase de diseño. Los diagnósticos de JET también incluyen cámaras del espectro visible y del infrarrojo que obtienen grabaciones de video e imágenes del plasma durante un pulso.

Pero un sistema de adquisición de datos abarca también muchos más conceptos como son el procesamiento, almacenamiento, recuperación y visualización posterior de toda la información. El ciclo de adquisición de datos que siguen los dispositivos de fusión por confinamiento magnéticos suele ser a intervalos cíclicos temporales, conocido también como ciclos pulsados. En cada ciclo de adquisición, pequeños sistemas autónomos electrónicos, donde residen las tarjetas de adquisición, recolectan datos de muchos diagnósticos. El número de estos sistemas depende lógicamente de la cantidad de señales a adquirir. Es habitual la presencia de varios centenares de canales digitalizadores. En el estellerator TJ-II hacen uso de más de 1000 canales de medida, aunque en las máquinas más grandes futuras puede escalar este valor en más de un factor diez. La coexistencia de varios sistemas de adquisición se debe a que no existe un único medio capaz de manejar eficientemente la gran diversidad de señales. Por este motivo, es usual que en la diagnosis de un plasma de fusión convivan diferentes configuraciones para la recogida de datos, como son los sistemas VME, VXI, PCI, PXI, etc. Los sistemas VME y VXI han sido los sistemas más utilizados en centros de investigación en fusión, permiten adquirir más canales y alojar más tarjetas de adquisición que cualquier otro, en cambio tienen un volumen y tamaño mayor que otros sistemas pero juegan también en su contra el hecho de que están siendo relegados por los fabricantes, a favor de PCI y compactPCI, donde se pueden encontrar más variedades comerciales en tarjetas de adquisición. El reemplazo de sistemas VME[12] se ha valorado en otros centros de investigación, sobre todo porque muchas fabricantes de tarjetas procesadoras modernas, como son las tarjetas Intel-Pentium y Motorola-PowerPC, han migrado sus viejas arquitecturas, Motorola-68000 sobre VME, a buses PCI más baratas, dentro de la propia tarjeta y por tanto, es necesario no solo hacer un paso intermedio hardware[13], sino desarrollos de software adicionales de conversión [Pereira et al., 2004b][Pereira et al., 2004c], entre el bus PCI y el bus VME, para que el procesador pueda comunicarse con las tarjetas de adquisición alojadas en el bus VME, no existiendo tampoco los drivers necesarios para que las tarjetas de adquisición puedan cohabitar con dos arquitecturas diferentes, haciendo necesario el tener que poder programarlos e implementarlos para ciertas tarjetas [Pereira et al., 2004d], [Pereira et al., 2004e]. En el sistema de adquisición de datos del TJ-II cohabitan sistemas basados en bus VXI con otros más modernos basados en compactPCI de National Instruments. Tradicionalmente, los sistemas de adquisición de datos se encargaban de

[12] VME replacement. (Pag. 13-20), [Pereira et al., 2005]
[13] PCI bridge, http://www.idt.com/products/interface-connectivity/vme/pci-vme-bridge

digitalizar señales y almacenarlas en bases de datos. Hoy en día se está fomentando la innovación en este campo haciendo que estos sistemas jueguen un papel más activo e inteligente. Nuevos elementos hardware como las unidades de procesamiento gráficas GPU, aumentarán la capacidad de cálculo en los propios dispositivos de adquisición, facilitando su procesamiento en tareas de tiempo real, como pueden ser el acondicionamiento digital de señales, técnicas de optimización, técnicas de compresión de datos, etc. El objetivo, mejorar lo presente para poder conseguir controladores más rápidos de adquisición, muchas mega-muestras por segundo, que puedan dar servicio a procesamiento y almacenamiento de formas de onda en continuo y orientado a entornos con pulsos más largos previstos en ITER. La recuperación posterior de los datos y su presentación visual ha sido otro de los temas relevantes en la investigación en fusión. Igualmente, la innovación en esta área es de especial relevancia de cara al futuro. Sistemas menos pasivos y más inteligentes de acceso y visualización se hacen también muy necesarios, orientados a pulso largo y en tiempo real.

2.4 Herramientas de análisis y supercomputación

En la medida que se van facilitando los medios para almacenar datos, toma importancia el siguiente paso, su análisis. Reconocer los patrones de comportamiento y de relación entre los datos es una tarea tediosa en tiempo y forma, pero puede ser más fácil si se tiene un objetivo claro en las metodologías a utilizar y en las herramientas necesarias a explotar.

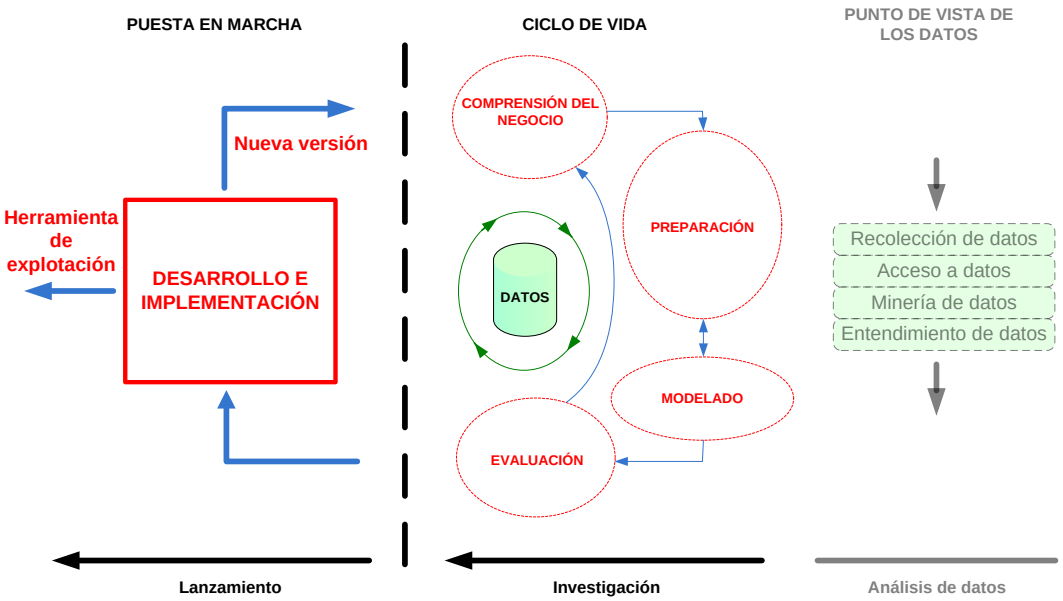


Figura 2. 1. Ciclo de vida en minería de datos

En el ciclo tradicional de minería de datos, representado en el diagrama de la Figura 2. 1, subyace paralelamente a las etapas indicadas, la realización de un análisis exhaustivo de los datos que pueda permitir comprender no solo la lógica general de los mismos, sino también el descubrimiento de información puntual muy relevante, valiosa y difícil de extraer. El sistema de adquisición de datos facilita la recolección de la información desde diferentes dispositivos electrónicos, como se ha descrito en la sección anterior. El acceso a esos datos comprende el uso de diferentes protocolos de comunicación y archivado de datos, así como tareas de compresión/descompresión de información, métodos de búsqueda e indexación, etc. La minería de datos es el campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. A veces, no solo por la gran cantidad de información disponible sino también por la mala calidad de la misma, es muy difícil extraer conocimiento útil que pueda ayudar en tareas precisas de clasificación o en tareas de decisión en control de procesos. Bien sea desde un punto de vista estadístico como la utilización de métodos descriptivos, inferenciales, análisis de varianza, de regresión, etc., o desde un punto de vista de la informática, mediante el uso de algoritmos genéticos, redes neuronales,

técnicas de agrupamiento, algoritmos de clasificación, etc., se hace necesario el analizar los datos concienzudamente para poder realizar tareas futuras automáticas de predicción y/o de control. La innovación de nuevas técnicas en el análisis de datos o la combinación precisa y ocurrente de los métodos ya existentes, se hace fundamental en la investigación en fusión. La evaluación de muchos modelos, que hacen uso a su vez de diferentes métodos de transformación y acondicionamiento de los datos, es una constante en la investigación en el análisis de datos de fusión. Tanto los datos de entrada, como los modelos resultantes, pueden estar más correctos o incorrectos, pueden ser más precisos o menos precisos y efectivamente, no existen los modelos finales perfectos, pero lo más importante es que sean útiles y valiosos en tareas críticas de predicción o que sepan generalizar suficientemente para la mayoría de los casos, comportándose así de una forma más robusta en tareas menos críticas de clasificación. Si el modelo final no superara el proceso de evaluación, los pasos anteriores se repiten continuamente. Esta retroalimentación es una constante habitual y necesaria hasta obtener un modelo válido. Una vez validado el modelo, si resulta ser aceptable y proporciona salidas adecuadas y/o con márgenes de error admisibles, éste ya está listo para su explotación. El desarrollo posterior de una herramienta más transparente hacia el usuario final, que implemente la abstracción investigada y pueda ser distribuida en diferentes entornos de computación, complementa el proceso final de investigación y añade más valor y publicidad al trabajo desarrollado.

Cuando el análisis de los datos se hace intratable debido a la gran cantidad de información a procesar, es necesario hacer uso de recursos computacionales que aceleren los procesos en la extracción de información. Al igual que en muchas otras disciplinas, los datos generados por los dispositivos experimentales de fusión se procesan paralelamente de forma distribuida en granjas de ordenadores. La computación distribuida o paralela es un nuevo modelo para resolver problemas de computación masiva utilizando un gran número de ordenadores organizados en clústeres incrustados en una infraestructura de telecomunicaciones de alta velocidad también distribuida. En el CIEMAT, institución donde reside el dispositivo TJ-II, se cuenta con un sistema de supercomputación formado por un conjunto de nodos de cálculo organizados en forma de cluster de alto rendimiento (HPC, siglas del inglés) con 240 nodos para cálculo y 1920 núcleos. El sistema de almacenamiento, está formado por 6 servidores, conectados también a las redes de fibra óptica y de *ethernet*, y con una capacidad de espacio en disco de 120 TB brutos. El software que gestiona este sistema es Lustre, que está organizado como un sistema de ficheros paralelo, proporcionando un alto ancho de banda en el acceso a los mismos. La gestión de los trabajos y procesos se realiza mediante MPI o interfaz de pasos de mensajes. MPI es un protocolo de comunicación entre computadoras. Es el estándar para la comunicación entre los nodos que ejecutan un programa en un sistema de memoria distribuida. Las implementaciones en MPI consisten en un conjunto de bibliotecas y rutinas que pueden ser utilizadas en programas escritos en lenguajes de programación C/C++. En esta tesis se ha implementado y se ha hecho uso de programas basados en lenguaje C que hacen a su vez uso de librerías MPI y que implementan algoritmos de aprendizaje basados en SVM e inferencia bayesiana para realizar tareas de clasificación.

2.5 La innovación en el tratamiento masivo de datos

El uso del tratamiento masivo de datos se convertirá en una base clave de la competencia y el crecimiento futuro en todas las organizaciones. En el sector de los dispositivos experimentales de fusión nuclear esta información se genera constantemente pero apenas se trata, debido también en parte a su dificultad para su análisis y comprensión. El análisis de grandes conjuntos de datos, el llamado '**análisis masivo de datos**', apunta como uno de los negocios de mayor fututo. Consiste en analizar y explotar masas de datos demasiado grandes y complejos para manipular e interrogar con métodos y herramientas estándares. Mediante la innovación, trata de crear nuevos productos y mejorar la competitividad y la productividad. Este término plantea básicamente tres retos[14] sobre el flujo de datos:

- Volumen: saber cómo gestionar e integrar grandes volúmenes de datos, procedentes de fuentes heterogéneas.
- Velocidad: poder acceder a la plataforma desde cualquier lugar, de forma autónoma por cualquier usuario de negocio, para mejorar y agilizar la toma de decisiones mediante la automatización y la programación de acciones, eventos y alarmas.
- Variedad: conseguir unificar contenidos dispersos y no estructurados, con datos históricos, actuales y/o predictivos para un manejo óptimo de los mismos y para extraer de ellos información de valor.

Pero existe uno más, que es la extracción automática del conocimiento relevante dentro de dichos datos. En esta tesis, se intentan abordar los retos anteriores, ofreciendo técnicas y desarrollando herramientas necesarias para identificar patrones, entregar el conocimiento y la visión adecuada a tiempo, sobre los datos, a los responsables en la toma de decisiones. La innovación mediante la aplicación de sistemas híbridos, mediante la combinación de métodos y técnicas estadísticas e informáticas, para el tratamiento avanzado de datos, con el objetivo de dar un apoyo a la decisión en tiempo real, es de especial relevancia en el campo de la fusión nuclear.

Otro de los grandes conceptos innovadores que se están planteando es el denominado '**Internet de las cosas**', (IoT, siglas del inglés). IoT trata de dar respuesta a la interconectividad entre nodos, bien sean, usuarios, dispositivos, sensores, infraestructuras, etc., no solo para el acceso ágil a los datos sino también para el envío y recopilación de información relevante. El término cubre un amplio rango de conceptos, dispositivos y plataformas. Por ejemplo, la plataforma IoT de Intel[15], es un modelo de referencia con soluciones hardware y software que aportan una base para conectar con fluidez y seguridad diferentes dispositivos, proporcionando no solo conectividad remota en la nube

[14] El tratamiento masivo de datos – Big data. http://rtdibermatica.com/?p=319
[15] Intel-IoT. http://www.intel.com/content/www/us/en/internet-of-things/overview.html

sino más valor añadido a través del análisis. Ninguna empresa puede crear IoT por sí sola. El ecosistema de IoT proporciona hardware, software, herramientas, integración de sistemas e infraestructura de nube y redes que se necesitan para acelerar el desarrollo y la implementación de soluciones que entregan información y datos a la nube. En este sentido, proyectos como "IoT Big Data Challenge", desarrollado en colaboración con Google[16], trata de innovar siguiendo este concepto, analizar flujos de datos continuos en la nube mediante sensores y dispositivos interconectados.

Por tanto, la innovación en el tratamiento masivo de datos no solo está ligada a conceptos de análisis y a la aplicación de métodos y técnicas ingeniosas en el tratamiento de los mismos. Sistemas de virtualización hardware o la convivencia en un solo equipo de diferentes sistemas operativos en sistemas embarcados pueden resultar también de gran ayuda en el tratamiento de los datos, por ejemplo, mejorando el rendimiento de ciertas aplicaciones críticas de control en tiempo real, mientras otras menos críticas, como son multimedia o video, son controladas por otro sistema operativo sobre la misma máquina, o simplemente para mejorar la calidad del servicio, con la posibilidad de separar servicios entre distintas máquinas virtuales y supervisarlas, para que cuando una falla, el sistema pueda reiniciarla.

[16] IoT Big Data Challenge. http://telitgcp4iot.com/

Capítulo 3

El reconocimiento de patrones morfológicos

El análisis visual de los datos es especialmente relevante como una primera evaluación general de los mismos. Esto nos aporta una referencia estructural y espacial de la situación de los datos y de su comportamiento. Las señales adquiridas por los diferentes diagnósticos y digitalizadas por los sistemas de adquisición, son almacenadas para su posterior análisis por el personal científico. La evolución temporal de estas señales representa formas de onda estructurales donde quedan reflejadas los eventos físicos que acontecen y que reflejan un patrón característico en dicha medición. A menudo, buscando estos patrones gráficos característicos, encontramos en que instante de tiempo acontece dicha fenomenología física. El proceso de búsqueda, efectivamente se puede realizar señal a señal, no obstante, esto se hace intratable cuando nuestra base de datos incluye cientos y miles de señales adquiridas en las diferentes descargas de operación. El reconocimiento ágil, rápido y efectivo en la búsqueda de patrones morfológicos en señales de evolución temporal es extremadamente importante. En este tema se documenta y explica la base teórica investigada, no solo la realizada para formas de onda parciales y completas muy similares entre ellas y pertenecientes a grandes bases de datos de fusión, sino también la efectuada y explorada en patrones gráficos semejantes pertenecientes a imágenes y películas de vídeo. Se aportan además estrategias de búsqueda optimizadas que mejoran notablemente las técnicas de reconocimiento de patrones inicialmente implementada.

3.1 Análisis y transformación de datos científicos experimentales

En el análisis de datos de muchas variables es necesario procesar previamente toda la información disponible. No se pueden mezclar ni comparar variables que estén medidas en diferentes unidades y/o en diferentes escalas de medición. La normalización de los datos se encarga de transformar los valores de las variables originales en otros en los que sí se puedan realizar los procesos de comparación o de cualquier otra índole que implique la interrelación de las variables implicadas. En otras ocasiones, ante la ausencia de valores en un determinado rango, es necesario hacer una estimación de los mismos en base a los valores conocidos, bien sea dentro del intervalo de observación original, término conocido como interpolación, o fuera de dicho intervalo, denominado extrapolación. La transformación de datos también incluye la aplicación de otros métodos más sofisticados como pueden ser el análisis de componentes principales mediante el estudio de la variabilidad e importancia de las variables, y otras funciones matemáticas de transformación, como las funciones wavelet y de Fourier, que descomponen la información original en componentes de tiempo y frecuencia, facilitando más información muy relevante para su análisis.

3.1.1 Normalización de datos

La normalización consiste en situar los datos sobre una escala de valores equivalentes que permita la comparación de atributos que toman valores en dominios o rangos diferentes. En efecto, si no hay normalización previa, las comparaciones tienden a quedar sesgadas por la influencia de los atributos con valores más altos o más bajos, hecho que distorsiona el resultado. Bien sea en tareas de clasificación o en tareas de regresión se hace imprescindible la normalización de los valores.

- **Normalización por la diferencia**

También conocido como **escala por intervalo** asigna significados a la diferencia de valores. Trata de ajustar los valores en un intervalo de extremos máximo y mínimo preestablecido, sin tener en cuenta la distribución de los mismos.

  - Entre 0 y 1:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3.\ 1)$$

- Entre máximo y mínimo específico:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} (x_{newMax} - x_{newMin}) + x_{newMin} \quad (3.\ 2)$$

Algunas librerías de software basados en algoritmos de aprendizaje muy populares como ***libsvm***[17] (utilizada en esta investigación), exigen que los datos introducidos vengan ya normalizados en un cierto intervalo, bien sea [0,1] o [-1,1]. De esta forma en los trabajos en los que se han utilizado esta herramienta, como son [González et al., 2010], [González et al., 2012], [González et al., 2012d] y en [Farias et al., 2012], se ha aplicado a los datos esta técnica de normalización.

En los desarrollos y trabajos realizados en la predicción de disrupciones, tales como [Dormido-Canto et al., 2013], [Vega et al., 2013b] y [Vega et al., 2013c], se ha optado por realizar previamente un truncamiento de valores para cada señal a un determinado máximo y mínimo, con todos aquellos datos que superen dichos límites, con el objetivo de eliminar valores muy extremos denominados 'valores atípicos', generados erróneamente por los equipos de medida y/o adquisición. Posteriormente se utilizan esos límites máximos y mínimos para realizar la normalización en el intervalo [0, 1].

- **Normalización basada en la distribución**

También conocido como **escala de proporción** o **escala de medida**, es un intervalo en el que se ha fijado un valor como origen o cero. Si todas las variables que forman la población, tienen las mismas unidades, solamente con centrar dichos valores respecto de la media puede llegar a ser suficiente, con el fin último de agrupar los datos respecto a un valor y eliminar así los desplazamientos que se pudieran encontrar en cada una de las variables, así se consigue que todas las señales tengan la misma referencia en la variable dependiente y no temporal de la señal, la amplitud. Tanto en funciones univariantes como en señales de evolución temporal en las que no es necesario relacionar con otras variables o señales, también se lleva la señal a una línea base tomando como referencia la amplitud 0 voltios [Vega et al., 2008]. Con esto conseguimos que dos señales iguales pero con desplazamiento diferentes puedan ser reconocidas como semejantes. Esto se consigue obteniendo valores dentro del rango elegido que tienen como propiedad que su media sea cero. Esta normalización basada en la media, resulta adecuada para trabajar después con métodos que utilizan distancias, como ocurre en el trabajo anteriormente citado y en otros desarrollos posteriores, como en [Vega et al., 2008c], [Pereira et al., 2010] y [Pereira et al., 2010b].

En cambio, cuando las tareas a desarrollar implican interrelacionar las variables entre sí y éstas han sido muestreadas en diferentes unidades de medida, también necesitaremos escalar dichos valores. Teniendo en cuenta la distribución de los datos, podemos realizar una normalización bien sea respecto a su desviación estándar o respecto a la longitud a la variable centrada.

- Escalado de desviación unidad (estandarización)

[17] A library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

$$x_i' = \frac{x_i - \bar{x}}{\sigma},\ siendo\ \ \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (3.\ 3)$$

Resultando una transformación para los datos donde se cumple ($\bar{x} = 0,\ \sigma = 1$)

- Escalado de longitud unidad

$$x_i' = \frac{x_i - \bar{x}}{L},\ siendo\ \ L = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3.\ 4)$$

Resultando una transformación para los datos donde se cumple ($\bar{x} = 0,\ L = 1$)

La estandarización de datos es muy útil preferiblemente en tareas de regresión a la hora de parametrizar una ecuación para extraer leyes de escala. Con este propósito, se aplicó este método de normalización en los siguientes trabajos, [González et al., 2012] y [Vega et al., 2012], con el objetivo de extraer leyes de escala aplicando métodos de regresión paramétricos y no paramétricos respectivamente en cada trabajo. Existen otras situaciones[18] donde es conveniente utilizar los coeficientes de regresión normalizados o transformados (aplicando cualquiera de estos dos métodos de normalización) en lugar de los correspondientes originales, para que aparezcan gráficamente en la misma escala y poder extraer conclusiones comparables.

- **Escalado decimal**

El escalado decimal permite reducir en un cierto número de potencias de diez el valor de un atributo. Esta transformación resulta especialmente útil al tratar con valores muy elevados. Se utiliza sobre todo en ajustes de regresión donde ciertas variables presentan valores muy alejados respecto de las demás.

$$x'_i = \frac{x_i}{10^j}\ \ siendo\ j, el\ número\ de\ potencia \quad (3.\ 5)$$

Este método también se ha aplicado igualmente en las dos publicaciones últimas anteriormente citadas.

3.1.2 Interpolación de valores desconocidos

Los valores ausentes o desconocidos corresponden a la situación en la que el valor de un atributo para un determinado objeto no se conoce. Las observaciones ausentes pueden causar problemas en los análisis y algunas medidas de series temporales no se pueden calcular si hay valores perdidos en la serie. Existen diferentes métodos para reemplazar los valores desconocidos. Reemplazamiento de la media de puntos adyacentes, de la mediana de los valores circundantes, etc. La interpolación suele ser el método más

[18] Análisis predictivo de datos. Tesis de Master. (Pag. 44-45), [Pereira, 2010]

utilizado para la asignación de valores desconocidos. Se define interpolación como el procedimiento que permite conocer de forma aproximada el valor que toma un dato desconocido a partir de un conjunto de datos observados. La operación de interpolación modifica el número de muestras presentes en una señal de evolución temporal, igualando la frecuencia de muestreo de la señal subyacente. Esta operación es importante en situaciones en las que una misma señal ha sido adquirida con diferentes ratios en la tasa de muestreo. En la interpolación lineal las muestras ficticias tienen un valor que se calcula como promedio de dos muestras originales adyacentes. En la mayoría de los casos el dato necesario no se encuentra explícito en el conjunto de datos sino entre dos valores de ésta, para lo cual es necesario estimarlo de entre los valores que presentan estos datos mediante la interpolación. Otro motivo importante para utilizar la interpolación lineal es la de preparar todo el conjunto inicial de señales con el número de muestras adecuado para poder aplicarle otros métodos. En ocasiones, se requiere que el número de muestras de entrada sea potencia de 2, como se ha requerido en los trabajos [Pereira et al., 2010], [Pereira et al., 2010b], [Vega et al., 2008] y [Vega et al., 2008c]. Para ello aplicamos la interpolación lineal para 2^m puntos entre un tiempo inicial y un tiempo final de forma que $2^{m-1} <= u\text{-}p+1 < 2^m$ siendo u la posición de la última muestra y p la posición de la primera muestra, m es la potencia a calcular y define los límites del intervalo. Básicamente, la expresión anterior explica que la interpolación resultante no puede tener menos muestras que la serie temporal inicial.

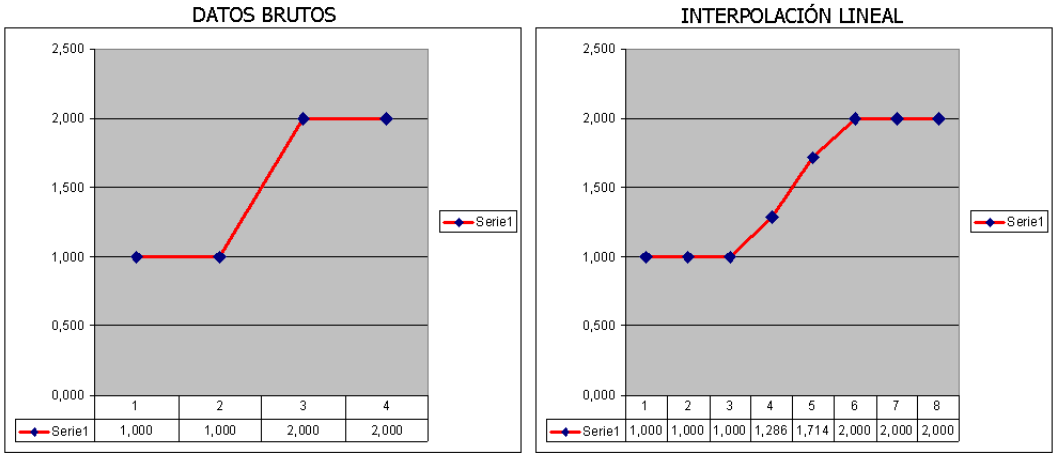


Figura 3. 1. Interpolación lineal

En la Figura 3. 1 anterior, se puede ver un ejemplo de cómo utilizar la expresión del cálculo de potencia de 2 para una colección de datos brutos de 4 elementos iniciales. Se comprueba que aunque las muestras resultantes son 8 y que la unión entre los puntos 3 y 6 no es perfectamente lineal, se sigue manteniendo la forma estructural de la señal original y el error que se comete es asumible y despreciable.

Con respecto a las amplitudes, para el cálculo de la línea recta a interpolar entre dos puntos, podemos decir que si los puntos conocidos vienen dados por las coordenadas (x_0,y_0) y (x_1,y_1), para un valor x en abscisas en el intervalo (x_0,x_1), el valor incógnita de y en ordenadas, a lo largo de la línea recta viene dado por la ecuación,

$$\frac{y - y_0}{y_1 - y_0} = \frac{x - x_0}{x_1 - x_0}$$

$$y = y_0 + (y_1 - y_0)\frac{x - x_0}{x_1 - x_0}$$

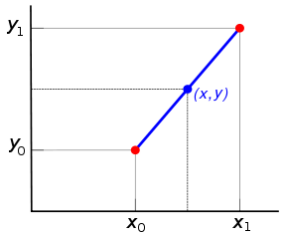


Figura 3. 2. Formulación matemática en la interpolación lineal

Esta fórmula se puede obtener geométricamente según se muestra en la Figura 3. 2, dados los dos puntos rojos, la línea azul representa la interpolación lineal entre dichos puntos. Hay que tener en cuenta que la interpolación lineal se hace por pedazos y no entrega un solo polinomio para todo el conjunto de datos. También se puede realizar la interpolación ajustándolo a un polinomio o a otra función. En la Figura 3. 3 de más abajo se puede observar como la forma de onda original de color verde responde a una forma trigonométrica sinusoidal, el procesado de interpolación mejora significativamente la forma de onda resultante, incluso aumentando la resolución temporal.

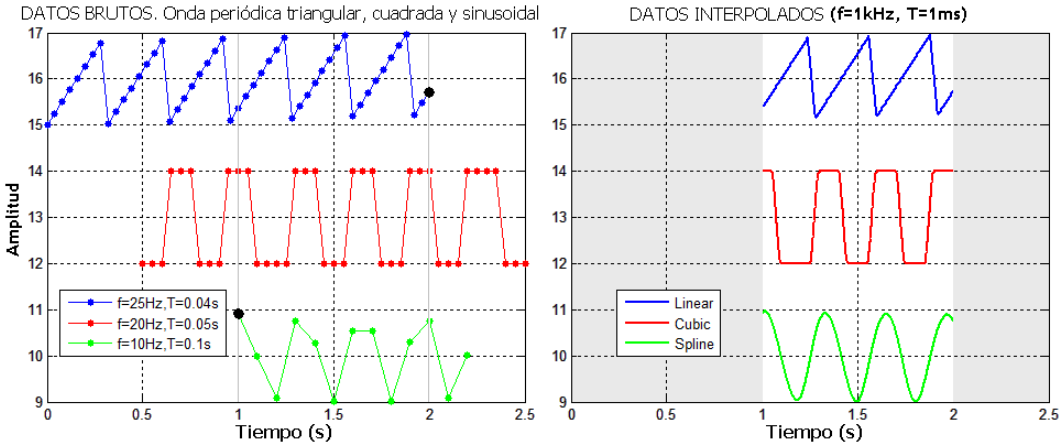


Figura 3. 3. Tipos de interpolación

En la Figura 3. 3 se puede observar diferentes métodos de interpolación sobre diferentes tipos de señales. En este caso el ajuste de interpolación se realiza con la misma frecuencia y periodo de muestreo para las tres señales resultantes. El intervalo de interpolación temporal se realiza desde el máximo valor temporal para el conjunto de todos los mínimos temporales de cada señal original y el extremo superior corresponde al mínimo valor temporal para el conjunto de todos los máximos temporales de cada una de las señales. Con esto se asegura que existen valores no ausentes para cada una de las muestras resultantes del proceso de interpolación. Esto es de gran importancia a la hora de procesar señales de evolución temporal en bases de datos de fusión. En cada descarga experimental se adquieren decenas de señales con diferentes periodos de muestreo e intervalos temporales, es necesario realizar el ajuste explicado anteriormente para asegurarnos de la coincidencia temporal de todas las señales interesadas en el estudio. Este procedimiento de interpolación pero en su variante lineal, se ha utilizado en muchos trabajos de investigación, no solo aplicado a señales utilizadas en el estudio de la transición L/H [González et al., 2012], [González et al., 2010], sino también a las señales utilizadas en el estudio de la predicción de disrupciones [Pereira et al., 2014], [Vega et al., 2014].

El procesado de interpolación en tareas de tiempo real se hace más complicado en aquellos casos en los que el periodo de muestreo utilizado para interpolar resulta menor que el ofrecido por los sistemas de medida. Si los equipos electrónicos con los que se

cuenta, adquieren a escasa resolución temporal y queremos interpolar en tiempo real la señal adquirida, re-muestreando a más elevadas frecuencias, la solución más simple sería la de usar una interpolación con referencia al vecino más próximo, esto es, con referencia a la última muestra adquirida.

3.1.3 Discretización de información

El proceso de **discretización** consiste en la transformación de datos numéricos en categóricos estableciendo un criterio por medio del cual se pueda dividir los valores de un atributo en dos o más conjuntos disjuntos. Teniendo en cuenta que el conjunto de valores sobre el cual se trabajará después de discretizar implica una reducción de los valores por tratar, el número de comparaciones y cálculos que tendrá que realizar el correspondiente método de clasificación será menor, redundando en su beneficio.

Existen muchos métodos o criterios para discretizar flujos de datos numéricos. En esta investigación se ha utilizado la discretización sobre la **codificación delta**[19] en varios trabajos. El término codificación delta, engloba varias técnicas para almacenar datos basadas en la diferencia entre muestras sucesivas o secuenciales y generalmente se utiliza la letra griega delta (δ, Δ) para especificar el cambio de la variable. En nuestro caso, aplicamos la codificación delta a la diferencia entre muestras secuenciales de la variable dependiente de la señal representada en el eje de ordenadas. Al aplicar la interpolación lineal a los datos de una señal, conseguimos que el periodo de muestreo (Δx) de la variable independiente de la señal, esto es, la variable temporal representada en abscisas, sea constante. La expresión (δ/Δx) representa la pendiente de la recta entre dos puntos o muestras consecutivas [Vega et al., 2008c]. Al ser Δx constante, el valor delta δ, se convierte en el único dato que necesitamos conocer y almacenar. Tanto en el trabajo de investigación anterior, como en [Farias et al., 2006] y en [Dormido-Canto et al., 2008] se categorizaba el valor delta en diferentes rangos, asignándole una categoría a cada intervalo para representar finalmente ésta, con una letra alfabética (5 valores, *d,c,e,a,b*, que dependían de la pendiente de la recta). Trabajos posteriores [Pereira et al., 2010], [Pereira et al., 2010b], simplificaron la asignación de intervalos a solamente dos, pudiendo ser dicha pendiente de subida o de bajada, solamente valores positivos o negativos, codificando el carácter '*a*' si el delta es negativo y el carácter '*b*' si el delta es positivo. El objetivo de esta asignación, fue el obtener una división en intervalos más simple a partir de un atributo numérico continuo, de manera que a cada intervalo se le pueda asociar una etiqueta o categoría binaria. Esta discretización es una potente herramienta dentro del campo de la transformación de valores.

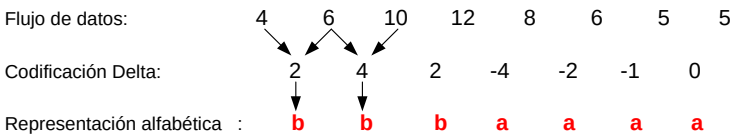


Figura 3. 4. Discretización de información

[19] Delta encoding. (Pag. 486), [Smith, 1997]

La representación alfabética (Figura 3. 4), nos permitirá facilitar la búsqueda de caracteres por medio de consultas en bases de datos relacionales y si esos caracteres los limitamos a únicamente dos letras, el reconocimiento de los mismos aumenta, pudiendo incluso considerar las transformaciones delta muy cercanas a cero (variaciones de pendiente visualmente imperceptibles), como valor indiferente '?' de búsqueda, aumentando así el reconocimiento y mayor recuperación de cadenas de caracteres más largas e igualmente similares.

En los trabajos de búsqueda de patrones dentro de imágenes realizados en [Vega et al., 2009], [Vega et al., 2008b] se utilizó igualmente la discretización de información para poder simplificar la búsqueda de sub-patrones dentro de imágenes pertenecientes a películas de vídeo.

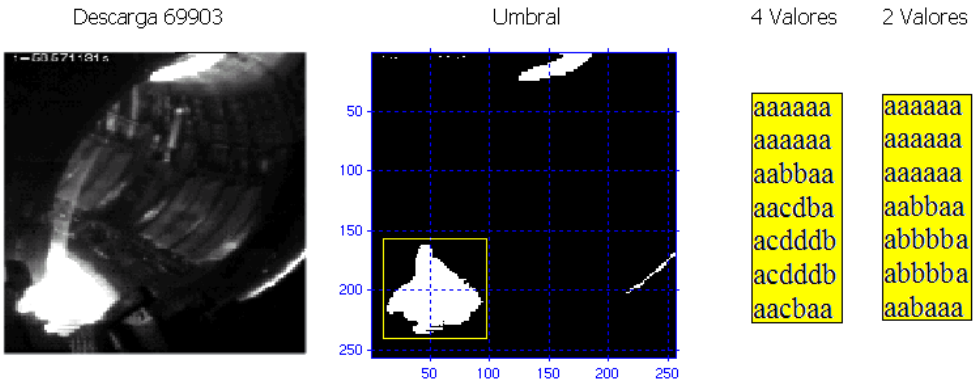


Figura 3. 5. Ejemplo de transformación de valores numéricos en una imagen del JET

En la imagen superior de la Figura 3. 5, se puede observar el plasma confinado para la descarga 69903 perteneciente a la instantánea número 75 de una película de video del JET. El proceso de umbralización del patrón señalado se puede observar a la derecha de la imagen, con 4 letras alfabéticas y con solamente 2 letras. Como se verá posteriormente, diferentes rangos de umbralización dan lugar a diferentes resultados tanto en la calidad de las imágenes recuperadas como en la cantidad de las mismas.

3.1.4 Transformadas matemáticas

La transformada **wavelet** facilita el análisis de una señal mediante su descomposición en componentes de tiempo-frecuencia de forma casi simultánea, mientras que la transformada de **Fourier** solamente permite la representación de las señales en componentes frecuenciales. Las dos transformaciones son capaces de revelar aspectos importantes en los datos tales como tendencias, puntos de ruptura, discontinuidades en las derivas, auto similaridad, etc. Por medio de la transformada wavelet de una señal obtenemos información valiosa que no se dispone en los datos originales brutos de esa señal, además, puede comprimir o eliminar ruido sin degradación apreciable. La

transformada **Haar**[20], englobada dentro del análisis wavelet discreto, es una de las variantes más simples y robustas de entre todas las funciones wavelet y es un método muy importante y útil en la reducción de información. A partir de los datos brutos de una señal discreta, se calculan los coeficientes de aproximación, realizando la media de cada 2 muestras consecutivas, de tal forma que en cada nivel de transformación obtenemos tantas muestras como la mitad de los puntos originales antes de la transformación. Esto quiere decir que antes de aplicar dicha transformada debemos asegurarnos que el número total de muestras es potencia de 2, además podemos realizar tantas transformaciones como queramos siempre y cuando nos queden muestras residuales a transformar para el número de características que estemos calculando. La transformada Haar descompone una señal discreta f en dos sub-señales que tienen tamaño $n/2$, una sub-señal contiene la **tendencia** o **coeficientes de aproximación** y la otra las **fluctuaciones** o **coeficientes de detalle**. La función Haar se realiza a varios niveles y el primer nivel corresponde a la primera transformación H_1, ésta se define de la siguiente manera,

$$f \overset{H_1}{\rightarrow} (a^1|d^1) \tag{3. 6}$$

Siendo, $f = (f_1, f_2, \dots, f_n)$ las muestras originales de la señal, n un número par, $a^1 = (a_1, a_2, \dots, a_{n/2})$ los coeficientes de aproximación del nivel 1, $d^1 = (d_1, d_2, \dots, d_{n/2})$ los coeficientes de detalle del nivel 1, a_m la media de las muestras f_{2m-1} y f_{2m} multiplicado por $\sqrt{2}$, d_m el valor de $(f_{2m-1} - f_{2m})/2$ multiplicado igualmente por $\sqrt{2}$, en cada posición $m = (1,2,\dots,n/2)$.

Niveles sucesivos de la transformada Haar (Figura 3. 6), se pueden aplicar descendentemente sobre los coeficientes de detalle del nivel anterior, reduciendo en cada transformación a la mitad dichos coeficientes.

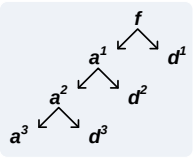


Figura 3. 6. Niveles de transformación wavelet-Haar

En la Figura 3. 7 se puede observar un ejemplo de cómo funciona la transformada wavelet-Haar aplicando las fórmulas anteriores.

[20] The Haar transform. (Pag. 9), [Walker, 1999]

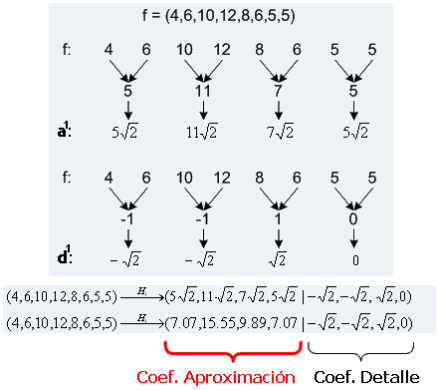


Figura 3. 7. Coeficientes de aproximación y de detalle wavelet-Haar

El factor multiplicativo $\sqrt{2}$ es muy importante no solo para poder conservar la energía de la señal original en los datos transformados, sino también para compactar esa energía en los coeficientes de aproximación. Por energía de una señal se entiende la suma de los cuadrados de sus valores.

$$\varepsilon_f = \varepsilon_{(a^1|d^1)} \tag{3. 7}$$

$$\sum_{i=1}^{n} f_i^2 = \sum_{i=1}^{n/2} {a_i^1}^2 + \sum_{i=1}^{n/2} {d_i^1}^2 = 440 + 6 = 446 \tag{3. 8}$$

Con esto se consigue también que a los datos transformados se le pueda hacer la función inversa obteniendo nuevamente la señal original,

$$(a^1|d^1) \xrightarrow{{H_1}^{-1}} f \tag{3. 9}$$

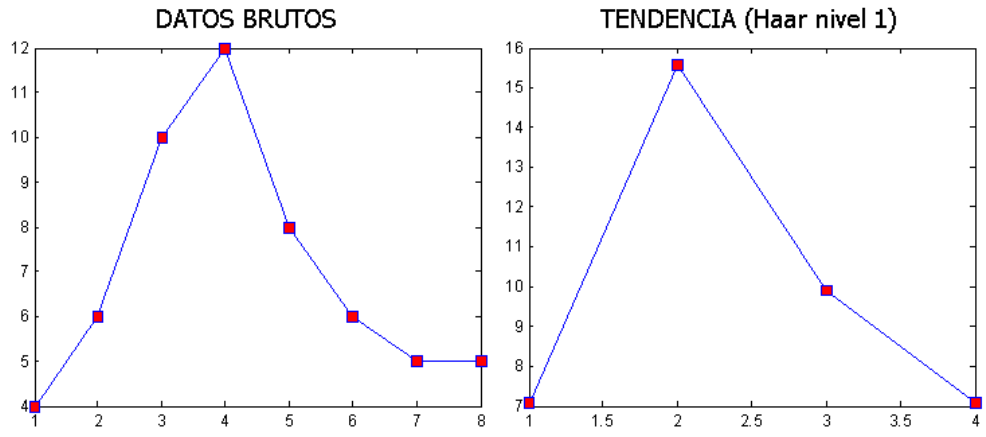


Figura 3. 8. Tendencia de los datos (coeficientes de aproximación)

Que la función Haar posea inversa no es lo más relevante. Como se ha comprobado, la mayor concentración de la energía recae sobre la tendencia de la señal transformada, esto quiere decir que solamente con los coeficientes de aproximación se puede conservar y representar la estructura morfológica de una señal sin perder apenas información visual y

con el añadido de reducir los datos a la mitad. En la Figura 3. 8 se puede apreciar como a partir de los 8 datos muestreados y utilizando solamente los datos de tendencia de la primera transformación Haar, podemos representar la misma forma estructural de la señal con únicamente esos 4 coeficientes de aproximación.

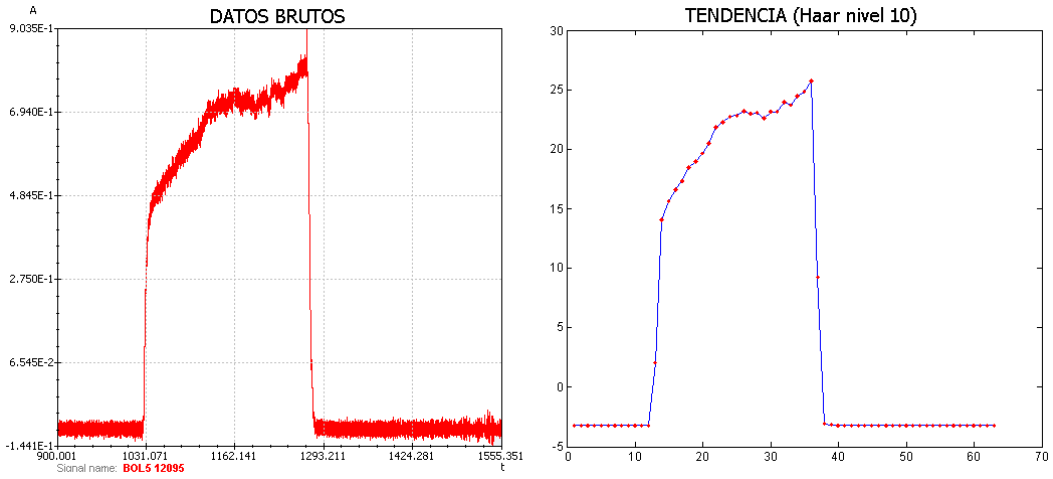


Figura 3. 9. Tendencia de la señal. Haar nivel 10

En la Figura 3. 9 se puede apreciar cómo podemos reconstruir la forma estructural de la señal original (65536 muestras, señal de Bolometría perteneciente a la descarga 12095 del TJ-II) con tan solo 64 coeficientes Haar y utilizando solamente los coeficientes de aproximación. Para ello ha sido necesario disminuir desde la potencia 2^{16} hasta llegar a 2^{6}. El nivel de transformación a adoptar dependerá de la resolución que necesitemos mantener. Aplicando un nivel 10 de transformación obtenemos 1 coeficiente Haar por cada 10 muestras de la señal original. De esta manera hemos reducido notablemente los datos de la señal original manteniendo una resolución aceptable. Si necesitáramos buscar detalles de la señal más precisos que estuvieran dentro de una resolución mínima, tendríamos que obtener más coeficientes Haar (128, 256, etc), disminuyendo así el intervalo temporal entre coeficientes.

Se puede apreciar también como la transformada Haar suaviza notablemente el ruido presente en la señal de referencia. No se puede dar el caso de que dos señales diferentes transformadas sean 100% similares, ya que para datos de entrada diferentes su transformada será diferente. El concepto final de similitud recae sobre la medida de la distancia para esos datos transformados, esto mismo se explicará más en detalle posteriormente. En los ejemplos anteriores, existe pérdida de información, ya que no aplicamos los coeficientes de detalle, no obstante no es ningún inconveniente, ya que una de las utilidades más potentes de los wavelets consiste en reducir los datos a ciertos niveles para que pueda realizarse una primera selección de señales parecidas en el menor tiempo posible. En los trabajos [Pereira et al., 2010], [Pereira et al., 2010b], y [Vega et al., 2008], se utilizan los coeficientes de aproximación de la transformada wavelet-Haar para extraer información de las señales y reducir notablemente sus datos con el objetivo de reconocer, buscar y recuperar formas de onda estructurales en tiempos de computación mínimos.

La descomposición wavelet es una potente herramienta en la transformación de datos para procesar señales de evolución temporal. Durante el transcurso de esta investigación también se ha tratado el procesamiento de imágenes y otros trabajos relacionados igualmente con la búsqueda de información y patrones dentro de ellas.

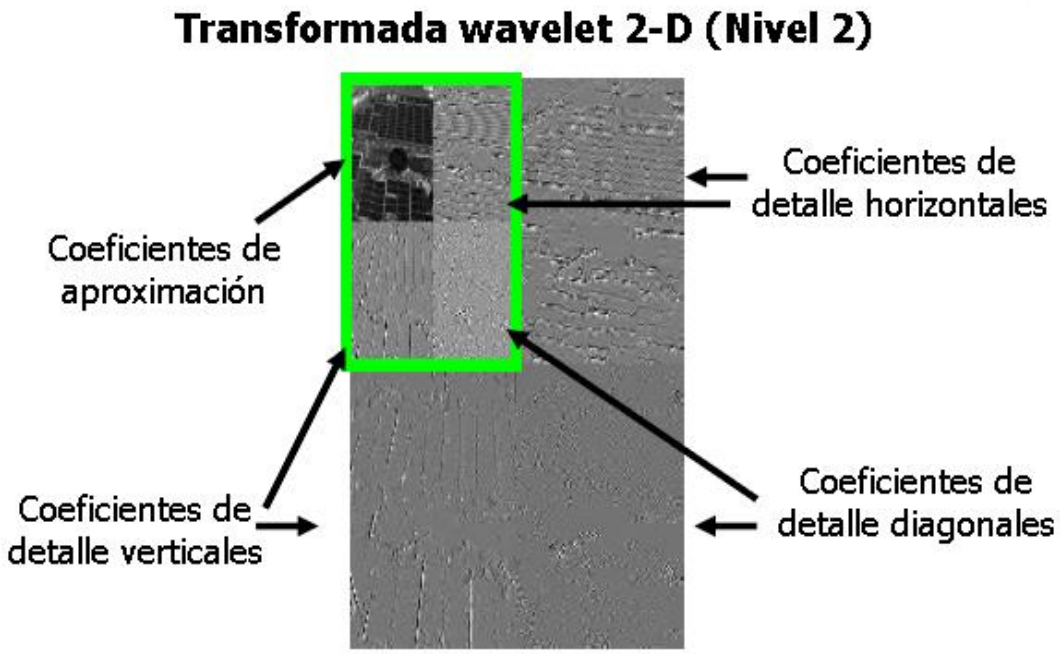


Figura 3. 10. Transformada wavelet 2D en imágenes

La extensión de las funciones wavelets a dos dimensiones permite hacer un análisis multi-resolución de la variabilidad de la imagen en las distintas direcciones: horizontal, vertical y diagonal. Al descomponer una imagen $n \times n$ y sub-muestreando, obtendremos 4 imágenes o matrices de información de $n/2 \times n/2$, una sería la aproximación y las demás tendrán los detalles en las direcciones mencionadas. Dicho de otra forma podemos decir que la transformada wavelet es una transformación de la imagen que la divide en dos tipos de imágenes de menor tamaño, la tendencia y las fluctuaciones. La tendencia viene a ser una copia de la imagen original a menor resolución y las fluctuaciones almacenan información referida a los cambios locales en la imagen original. La tendencia y las fluctuaciones más significativas permiten una compresión de la imagen a cambio de descartar información irrelevante y de la eliminación de ruido. El estudio de la tendencia y las fluctuaciones permite, entre otras cosas, la comparación con patrones para detectar formas en una imagen. La descomposición y reducción de información basada en técnicas wavelet han sido aplicadas a imágenes de JET [Vega et al., 2009], [Vega et al., 2008b] y a imágenes del diagnóstico de esparcimiento Thomson del TJ-II [Vega et al., 2005], [Vega et al., 2010].

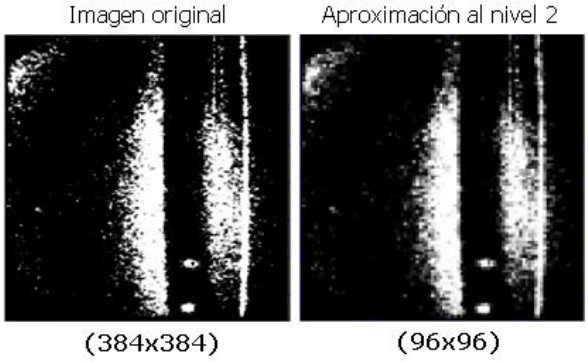


Figura 3. 11. Transformada wavelet-Haar 2D sobre una imagen del diagnóstico Thomson

En la Figura 3. 11 se puede observar una imagen del diagnóstico Scattering Thomson y su transformada Wavelet a un nivel 2, conservando la misma representación visual pero con solo 96x96 pixeles de información.

Otra función matemática muy utilizada para la transformación de valores es la transformada de Fourier, TF. La TF es una técnica matemática para transformar la representación temporal de una señal a una representación en el dominio de la frecuencia y viceversa. El análisis de Fourier lleva a cabo una descomposición de la señal en sus componentes sinusoidales para diferentes frecuencias Figura 3. 12.

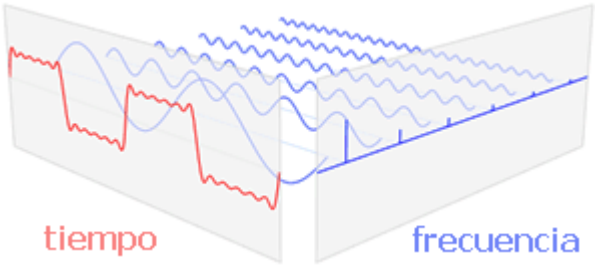


Figura 3. 12. Descomposición frecuencial de una señal de evolución temporal

La transformada de Fourier discreta TFD, es el nombre dado a la función TF cuando se aplica a una señal digital o discreta en vez de a una señal analógica o continua. La transformada de Fourier rápida TFR, es una versión más rápida de la TFD que puede ser aplicada cuando el número de muestras de la señal es potencia de dos. Un cálculo de TFR toma aproximadamente *n*log2(n)* operaciones, mientras que TFD toma aproximadamente n^2 operaciones, así es que la TFR es significativamente más rápida. La salida de la función TFR incluye frecuencias negativas que existen puramente como propiedades matemáticas de la TF. La primera mitad de la salida de TFR contiene frecuencias en orden ascendente empezando desde la componente continua 0 *Hz*. La segunda mitad del arreglo contiene frecuencias negativas. En los trabajos [Vega et al., 2014], [Dormido-Canto et al., 2013] y en [Pereira et al., 2014] se ha utilizado la TFR eliminando la componente continua y las frecuencias negativas de los valores transformados. En sendos trabajos se utilizaron 7 señales del JET para poder caracterizar el estado del plasma como disruptivo o no disruptivo. Estas señales fueron procesadas utilizando ventanas temporales de 32 *ms* con una frecuencia de muestreo de 1 *kHz*. Para cada ventana de 32 muestras se utilizaron 2 características, la media de sus amplitudes y la desviación estándar del espectro de Fourier de las muestras, previa eliminación de la componente continua de los valores transformados. Por tanto, un total de 14 características de partida fueron utilizadas para poder predecir comportamientos disruptivos en las señales del JET.

3.1.5 Distancia entre vectores y similaridad

La aplicación práctica de los algoritmos de aprendizaje obliga a considerar medidas o cuantificaciones, habitualmente numéricas, de cuánto son de semejantes o similares dos objetos o vectores de datos.

- **Distancia de Hamming**

Previa umbralización de los datos en valores binarios, la distancia de Hamming consiste en la cuenta o el número de coincidencias entre coordenadas. La función OR-exclusiva nos informará de la coincidencia o diferencia de cada posición en la cadena numérica (comparación de dos números binarios), dando por resultado 0 cuando los valores en las entradas son coincidentes y 1 cuando son diferentes.

$$d_{Hamming}(u,v) = \#\{i = 1,\dots,n:\ u_i \oplus v_i = 0\} \tag{3.10}$$

Si queremos medir dos señales utilizando la distancia de Hamming, esto se interpreta de la siguiente manera, cuando las pendientes de las muestras o los tramos a comparar son las dos de subida o las dos de bajada quiere decir que la señal en ese tramo comparado es similar, en cambio cuando una de ellas es de subida y la otra es de bajada quiere decir que son diferentes estructuralmente hablando para ese tramo comparado, por tanto lo que estamos realizando es una comparación de rampas o desniveles, la suma de todas las coincidencias corresponderá a la distancia entre esas dos señales. Si auto comparamos una señal consigo misma, lógicamente nos resultará similitud 100% de todos sus valores.

- **Distancia euclídea**

Corresponde al valor de la raíz cuadrada de la suma de las diferencias cuadráticas entre los valores de cada coordenada.

$$d_{euclídea}(u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \tag{3.11}$$

Como esta distancia crece cuando el número de coordenadas crece, a menudo se calcula una distancia media dividiéndola por el número total de puntos.

$$d_{euclídea\ media}(u,v) = \sqrt{\frac{\sum_{i=1}^{n}(u_i - v_i)^2}{n}} \tag{3.12}$$

- **Similaridad del coseno**

En este caso la similitud de dos vectores de datos es el valor absoluto del coseno del ángulo que forman esos dos vectores (producto escalar entre vectores).

$$d_{coseno}(u,v) = \frac{|u\cdot v|}{\|u\|\|v\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\left(\sqrt{\sum_{i=1}^{n} u_i{}^2}\right)\left(\sqrt{\sum_{i=1}^{n} v_i{}^2}\right)},\ 0 \le d_{coseno}(u,v) \le 1 \tag{3.13}$$

La virtud de utilizar el valor absoluto de esta distancia es que el método no depende de las ganancias de amplificación de los valores ni de los signos de polaridad de los mismos.

La distancia del producto escalar entre vectores de datos ha sido utilizado en esta investigación en la búsqueda y similitud entre patrones para formas de onda completas en los trabajos [Vega et al., 2008], [Vega et al., 2008c]. La distancia euclídea media también ha sido utilizada en [Pereira et al., 2010] y en [Pereira et al., 2010b] como medida de semejanza en la recuperación de sub-patrones que son más similares a uno de referencia y que pertenecen y están incluidos dentro de formas de onda más largas. Así mismo en [Vega et al., 2008b] se utilizó la distancia euclídea, como medida de similaridad entre imágenes y sub-patrones dentro de imágenes. El estudio[21] que se hizo sobre la similaridad de formas de onda completas mediante la distancia de Hamming no aportó resultados más relevantes (incluso inferiores) que los obtenidos hasta ese momento con las métricas anteriores.

[21] Detección de formas de onda completas. (Pag. 15-16), [Pereira, 2009]

3.2 El proceso de búsqueda y descubrimiento de patrones ocultos.

Métodos de búsqueda y reconocimiento de patrones morfológicos en señales de evolución temporal han sido aplicados a bases de datos masivas en el campo de la fusión nuclear. Dicho reconocimiento se ha tratado de abordar en dos fases, la primera de ellas atiende a lo que se denomina forma de onda completa y consiste en encontrar las señales completas más parecidas a una de referencia. La segunda aproximación se refiere a formas estructurales contiguas dentro de la señal (patrones), consistente igualmente en indicar y en encontrar donde y en que señales se repiten esas similaridades para un patrón de referencia seleccionado por el usuario. Para ello se han abordado previamente diferentes técnicas de transformación de datos como son, la normalización, interpolación, transformadas wavelet, discretización, etc., que han hecho posible la preparación de los datos para que los métodos de búsqueda a aplicar fueran eficaces tanto en calidad de resultados como en el tiempo empleado para conseguirlo.

Posteriormente se trasladó la técnica de búsqueda usada en señales hacia la búsqueda gráfica de patrones dentro de imágenes.

3.2.1 Formas de onda similares en señales de evolución temporal

Mostrando una visión inicial de abajo hacia arriba para explicar las siguientes técnicas que se muestran a continuación, se puede decir que la idea subyacente para la búsqueda de patrones dentro de señales es el de poder representar estas señales con unas características muy reducidas, las cuales van a ser almacenadas como una cadena de caracteres en una base de datos relacional, para que recaiga así toda la potencia de búsqueda en el motor de dicha base de datos. Por tanto, un paso muy importante para la consecución de los objetivos planteados es el de preparar y acondicionar las señales convenientemente mediante esta representación. Entremos en detalle, ya de una forma ordenada desde el principio, para cada técnica utilizada

3.2.1.1 Detección de formas de onda completas

En esta sección se van a detallar los procesos para analizar formas enteras de señal o completas y la recuperación de las mismas.

- **Primera aproximación, distancia del coseno.**

Se trata de buscar las señales completas más parecidas o similares a una de referencia y expresar como de diferentes son todas las recuperadas con respecto a esa señal de referencia. La primera aproximación que se ha utilizado para cuantificar la similitud de señales completas es el valor absoluto del producto escalar normalizado, utilizado en [Vega et al., 2008] y aplicado a señales del TJ-II. En esa ocasión el sistema de clasificación se basaba en la combinación de técnicas supervisadas de clustering (agrupamiento basado en una estructura arborescente) y en el producto escalar de vectores normalizados como medida de similaridad. Para no tener que mantener un sistema de agrupamiento paralelo, posteriormente se substituyó el sistema de clustering por un motor de base de datos relacional, aplicando directamente la función de similaridad a todos los registros a recuperar en dicha base de datos [Pereira et al., 2010]. En el caso que nos ocupa los vectores de referencia a comparar son los coeficientes wavelets-Haar obtenidos para cada señal que se empareja de todas las señales indexadas en la base de datos. Finalmente se ordenarán de mayor similitud (1 es la mayor similitud) a menor similitud (0 es la menos similar). La virtud de utilizar la distancia del producto escalar normalizada es que el método no depende de ganancias de amplificación de la señal ni de los signos de la polaridad de las mismas.

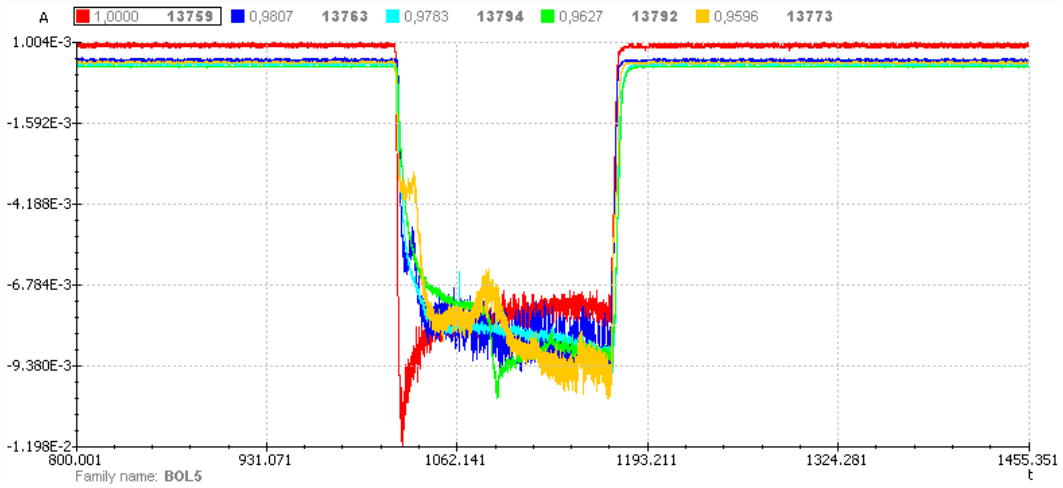


Figura 3. 13. Comparación de señales mediante la distancia del producto escalar

- **Segunda aproximación, distancia de Hamming.**

Primeramente recuperamos todas las señales que tengamos almacenadas en la base de datos. Cada señal está representada en nuestra base de datos por unos coeficientes Haar de una longitud determinada 64, 128, etc. y unos coeficientes delta, que son la diferencia entre muestras consecutivas de los coeficientes de la señal. Dependiendo del signo de los

coeficientes delta, asignamos una representación binaria, al signo negativo le asignamos 0 y al positivo le asignamos el número 1. Hacemos este mismo procedimiento con la señal de referencia. Procedemos seguidamente a realizar la función algebraica OR-Exclusiva XOR, entre la señal de referencia y cada una de las señales recuperadas. Está función nos informará de la coincidencia o diferencia de cada posición en la cadena numérica (comparación de dos números binarios), dando por resultado 0 cuando los valores en las entradas son coincidentes y 1 cuando son diferentes. En las señales esto se interpreta de la siguiente manera, cuando las pendientes a comparar son las dos de subida o las dos de bajada quiere decir que la señal en ese tramo comparado es similar, en cambio cuando una de ellas es de subida y la otra es de bajada quiere decir que son diferentes estructuralmente hablando para ese tramo comparado, por tanto lo que estamos realizando es una comparación de rampas o desniveles.

La función de similitud de Hamming entre dos señales viene definido por el número de coincidencias encontradas en la comparación de las muestras, los valores deltas, de las dos señales para el mismo índice de comparación. Ordenando finalmente todas las señales por el número de coincidencias encontradas, desde las más similares a las menos, cuantificada por dicha distancia de Hamming, obtendremos el resultado esperado.

BOL5 12095	00100110010111111111111111010111011100000100100101010100110100	
BOL5 12095	00100110010111111111111111010111011100000100100101010100110100	
XOR:	00000000000000000000000000000000000000000000000000000000000000000	**Similitud: 63**
BOL5 12095	00100110010111111111111111010111011100000100100101010100110100	
BOL5 12096	01100100111101000001101010101010101010101010101010100110100010011	
XOR:	01000010101010111110010100000100010010100011100000110000010011	**Similitud: 38**
BOL5 12095	00100110010111111111111111010111011100000100100101010100110100	
BOL5 12097	01011010011100010101010100100110010110010001110101000011101000 1	
XOR:	01111100001011101010101010001000101110011000111111101001110010 1	**Similitud: 29**

Figura 3. 14. Distancia de Hamming

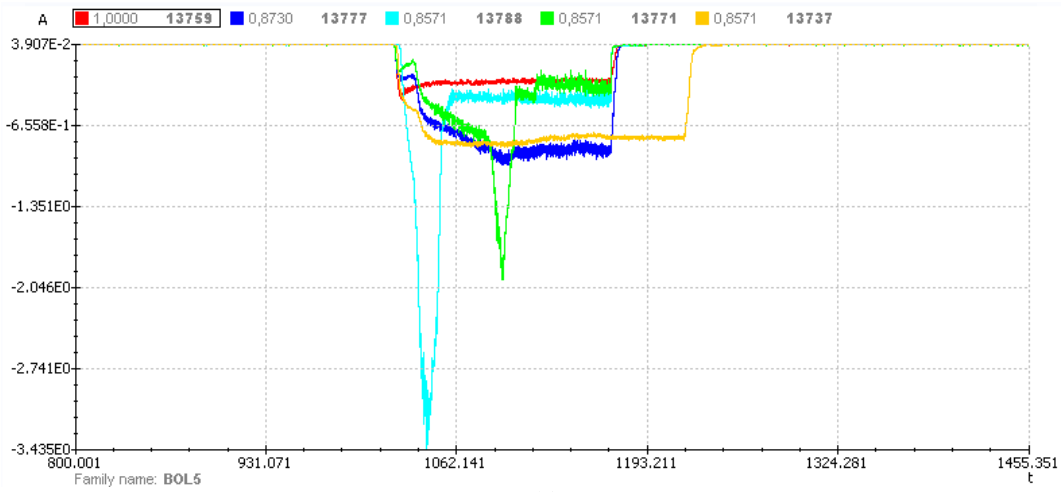


Figura 3. 15. Similaridad entre señales del TJ-II mediante la distancia de Hamming

3.2.1.2 Detección de patrones dentro de señales

Para la detección de patrones, esto es un subconjunto de muestras consecutivas de una señal, las consultas de similaridad se complican, ya que se hace intratable recorrer todo el espacio secuencial segmento a segmento para cada una de las señales en busca de un posible patrón coincidente. Para ello y previa discretización de valores reales en caracteres alfabéticos o primitivas[22], haremos uso de la potencia que nos brinda el motor de una base de datos relacional, trasladando el problema del reconocimiento de patrones a un problema de emparejamiento de caracteres alfabéticos gestionado por el motor de las bases de datos relacionales.

- **Primera aproximación, primitivas de longitud constante y múltiples pendientes.**

En [Dormido-Canto et al., 2006] se utiliza una longitud de señal constante, o sea un segmento con el mismo número de muestras para cada primitiva, resultando un número coincidente de primitivas totales para cada una de las señales que componen la base de datos relacional. La umbralización de cada segmento, atendiendo a la pendiente de la recta que representa, identifica a esa primitiva. En dicho trabajo se ha utilizado el método de primitivas de longitud constante (PLC) mediante 5 primitivas diferentes y representadas por las letras [a c d e z].

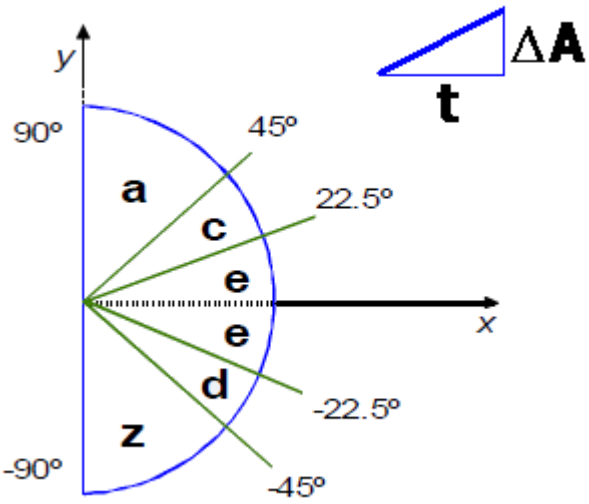


Figura 3. 16. Primitivas de longitud constante y múltiples pendientes

Buscar un patrón de una señal significa, consultar por una sub-cadena de longitud variable y combinación de esos cinco tipos de caracteres. Posteriormente, de entre todas las recuperaciones realizadas en la base de datos por la consulta de ese sub-conjunto de caracteres, se comparan los valores reales de dichas cadenas con los originales de referencia, para obtener una medida de similaridad que, ordenadas posteriormente, resulta de la obtención de una lista de sub-patrones colocados desde los más similares a los menos con respecto al sub-patrón de referencia objeto de la búsqueda.

[22] Pattern primitives. Pág. 243. [Marqués de Sa, 2001]

- **Segunda aproximación, primitivas de longitud adaptable.**

Las primitivas de longitud adaptable (PLA) es otro método para descomponer una señal por medio de aproximaciones lineales en trozos. En el método anterior (PLC), la señal original se descompone en segmentos iguales, sin embargo ahora, la longitud de cada segmento es diferente, los segmentos no tienen un número fijo de muestras. Todos los segmentos son codificados y agrupados en una cadena de primitivas, cada segmento con su propia etiqueta, pero con la diferencia que el número final de primitivas en cada señal no es constante. Por ejemplo, a una señal plana linealmente sin pendientes, le corresponderá solamente una primitiva, mientras que una señal muy ruidosa con muchas pendientes pronunciadas, le corresponderá un número muy elevado de primitivas, no superior al número de muestras. Con esta peculiaridad, es deseable adaptar el número de segmentos a un valor óptimo. En esta investigación se ha minimizado este número introduciendo un factor, la desviación estándar, como una medida de la dispersión de las amplitudes de la señal. Un error máximo fijo denominado E_{max}, se define para cada uno de los segmentos lineales. Este valor es función de la desviación de los datos de cada señal. El método PLA se implementa aplicando el siguiente algoritmo, mientras existan muestras a transformar para una señal:

1. Incrementar el número de segmento *k* (inicialmente *k*=1).
2. Incrementar el número de muestras *l* (inicialmente en cada nuevo segmento, *l*=3).
3. Generar una línea de regresión con *l* muestras.
4. Si el error del ajuste del segmento (error cuadrático medio, MSE, siglas del inglés) es menor que el máximo establecido ($MSE < MSE_{max}$), ir al punto 2, sino guardar el segmento e ir al punto 1.

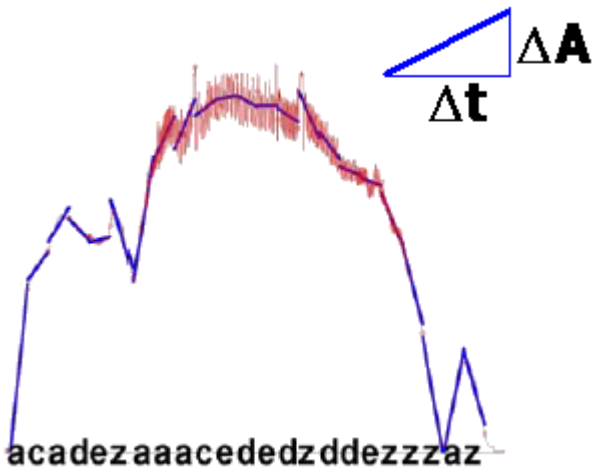


Figura 3. 17. Primitivas de longitud adaptable

En cada segmento creado, se almacena el rango horizontal y vertical (Δt, ΔA), necesarios posteriormente no solo para conocer la pendiente del segmento y poder umbralizarlo en una primitiva, siguiendo el mismo procedimiento que en PLC y poder almacenarlo en la base de datos relacional, sino también para obtener una similaridad numérica de comparación entre el patrón de referencia y cada uno de los patrones recuperados en una consulta a la base de datos.

La función de similitud aplicada consistirá en la media del producto escalar normalizado para cada segmento del patrón de referencia con respecto a cada segmento del patrón recuperado de la base de datos.

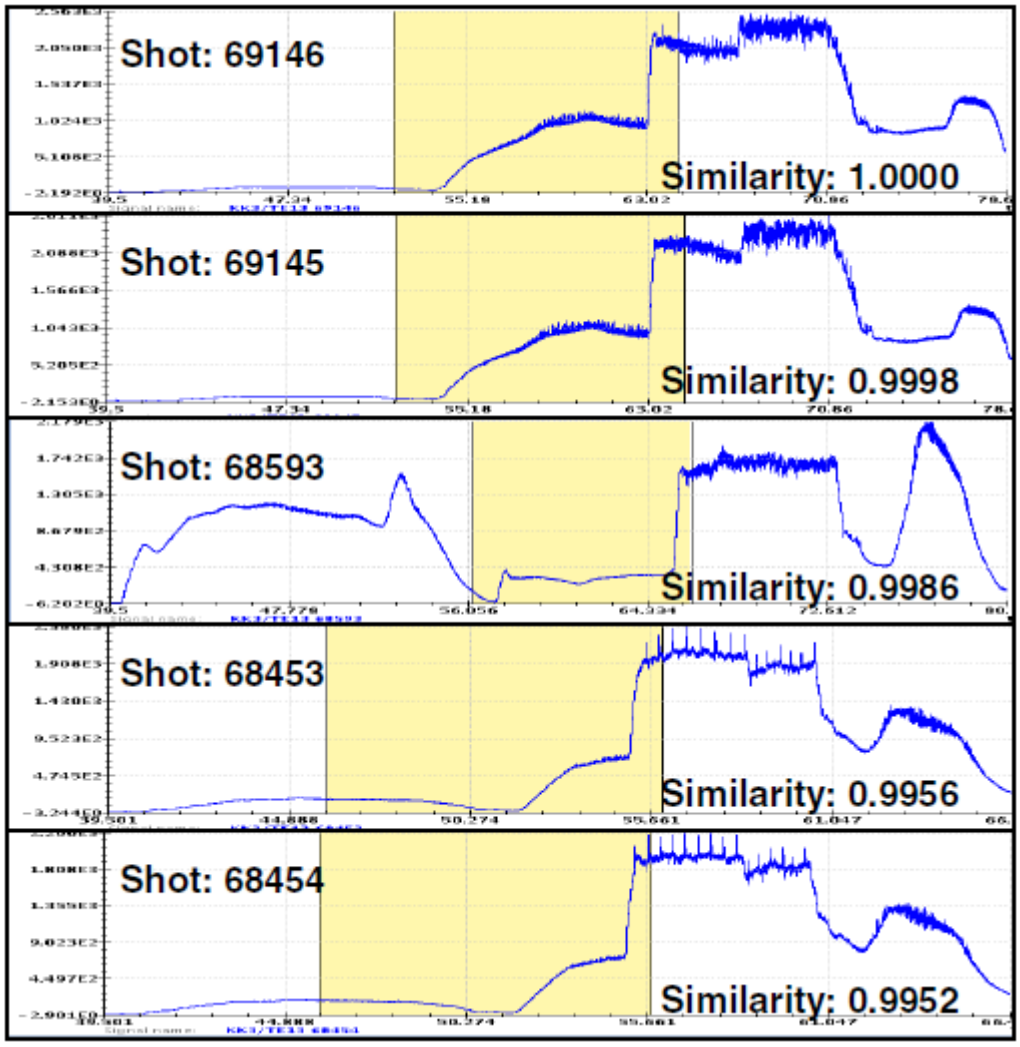


Figura 3. 18. Primitivas de longitud adaptable en señales del JET

- **Tercera aproximación, primitivas de longitud constante y polarización de pendientes.**

En esta ocasión en lugar de tener múltiples valores de primitivas, la cadena de caracteres resultante tiene una representación binaria de dos únicos valores (primitiva 'a' y primitiva 'b'), estas corresponden a las pendientes negativas o positivas de la recta resultante entre dos coeficientes delta consecutivos de la señal. Cada señal es procesada con diferentes técnicas antes de ser codificada en dicha cadena de caracteres. Los coeficientes Haar son la representación reducida de las muestras interpoladas de la señal original. La interpolación de las muestras originales se realiza para poder unificar la frecuencia de muestreo de la señal de partida y para poder acondicionar el número de muestras de entrada que se necesitan para las transformaciones Haar en base a potencias de 2.

En la Figura 3. 1 se esquematiza todas las etapas por las que atraviesa una señal antes de ser almacenada en la base de datos relacional.

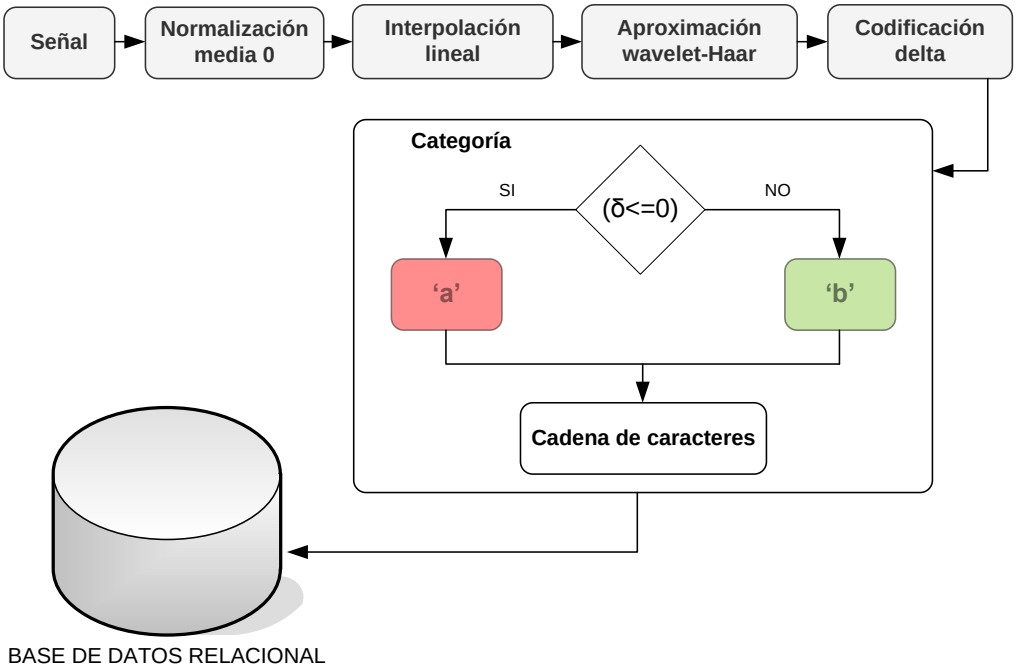


Figura 3. 19. Primitivas de longitud constante y polarización de pendientes

La elección de solamente dos primitivas es la diferencia esencial con respecto a las dos aproximaciones anteriores. Con solo dos primitivas, la probabilidad de encontrar el patrón más largo, esto es, la señal completa, en toda la base de datos es $\Omega_{(2\ primitivas)} = 1/2^{64}$ para 64 coeficientes delta, una probabilidad mucho mayor que la representada en la primera aproximación y comparándolo con el mismo número de segmentos, $\Omega_{(5\ primitivas)} = 1/5^{64}$. Con 2 primitivas se incrementa notablemente la posibilidad de encontrar sub-patrones muy largos y similares ocultos en grandes bases de datos.

3.2.2 Patrones gráficos semejantes en imágenes y películas de vídeo

Paralelamente a la investigación llevada a cabo en la búsqueda y recuperación de patrones en señales de evolución temporal, se ha venido investigando tales tareas pero aplicado a imágenes completas y a patrones gráficos dentro de imágenes.

3.2.2.1 Reconocimiento y clasificación de imágenes completas

En cada descarga de operación del TJ-II y por medio del diagnóstico Scattering-Thomson se generan los perfiles de densidad y temperatura correspondientes al plasma obtenido. El espectro en dos dimensiones de estos perfiles determina las imágenes generadas desde dicho diagnóstico. Las imágenes capturadas pueden ser de 5 tipos diferentes y dependiendo de la imagen obtenida se determina la clase de análisis a

realizar. La finalidad perseguida en el trabajo [Vega et al., 2005] fue la clasificación automática de las imágenes, desde los datos brutos y sin intervención humana para continuar en la determinación del análisis posterior a efectuar.

Cada imagen del diagnóstico Scattering Thomson del TJ-II tiene 221760 atributos (pixeles) constituidos a partir de una matriz de dimensiones 385 x 576. Ante tanta cantidad de información es necesario e imprescindible reducir el tamaño de los datos sin perder la información morfológica de las mismas. Métodos de reducción de dimensionalidad basados en la transformada wavelet en dos dimensiones se utilizaron en este trabajo para reducir la información de partida.

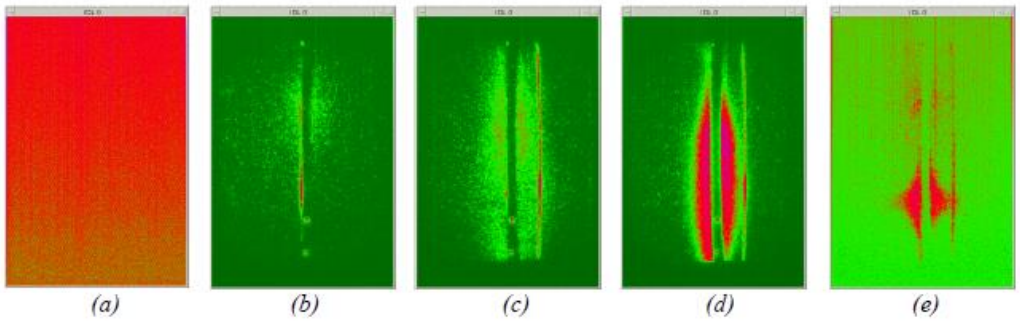


Figura 3. 20. Imágenes del diagnóstico de esparcimiento Thomson del TJ-II

El mejor nivel de reducción para las imágenes del espacimiento Thomson ha sido la transformada wavelet-Haar al nivel 4 con los coeficientes de detalle vertical, pasando de 221760 atributos a solamente 900 atributos. Para dotar un procedimiento automático en la clasificación de las imágenes del diagnóstico Thomson se desarrollaron aplicaciones software exclusivamente para ello, tanto para la sincronización del diagnóstico con la operación pulsada del TJ-II (explicado en el punto 5.1.1) como para el propio reconocimiento en sí de los patrones de cada imagen (utilizando un clasificador SVM, explicado en el apartado 4.2.1). Posteriormente se desarrolló una mejora del sistema de clasificación [Makili et al., 2010] debido a una mejora en la óptica del diagnóstico, manteniendo a la vez la misma arquitectura de sincronización de los procesos. Los mejores porcentajes de aciertos fueron del 92,7% y del 98% en sendos trabajos para diferentes campañas de operación del TJ-II. Un tercer trabajo [Vega et al., 2010] fue publicado igualmente para clasificar imágenes del diagnóstico Thomson del TJ-II con excelentes resultados (tasas de acierto del 97%), basado esta vez en una clasificación conformal (ver apartado 4.2.3) mediante el algoritmo del vecino más próximo. Finalmente en [González et al., 2012b] se aplicó un método para seleccionar las regiones características útiles y más importantes en cada tipo de imagen del diagnóstico Thomson, reduciendo notablemente la cantidad de datos a emplear en la clasificación y el tiempo computacional empleado en ello. Se aplicaron igualmente medidas conformales de clasificación pero esta vez basadas en SVM, consiguiendo tasas de acierto del 95.3% para un conjunto total de 634 imágenes.

Siguiendo con el enfoque basado en el reconocimiento de patrones morfológicos mediante la transformación de los datos en primitivas y gestionado por un motor de base de datos relacional, un sistema de recuperación de imágenes de la cámara visible de alta velocidad del JET fue desarrollado en [Vega et al., 2008b]. En esta ocasión las video-películas contienen millares de imágenes con una resolución de 274x300 pixeles. El proceso a seguir es similar al utilizado en señales de evolución temporal. Las imágenes

necesitan una preparación basada primeramente, en una umbralización de los datos para eliminar ruido y quedarse con los pixeles más significativos, la aplicación de la transformada wavelet-Haar en dos dimensiones, para reducir drásticamente los datos a procesar, quedándonos con los coeficientes de aproximación a un cierto nivel de descomposición (32x32 y 16x16). El último paso consiste en la discretización de los valores en primitivas o letras alfabéticas, atendiendo a los niveles de categorización que hayamos realizado. Estas primitivas se almacenan en la base de datos relacional de forma que la primera fila de la matriz de primitivas corresponderá a la primera columna o primera variable de nuestra base de datos, la segunda fila será la segunda variable de datos y así sucesivamente. Buscar por una imagen se convierte en la búsqueda de caracteres alfabéticos en la base de datos relacional utilizando consultas que hagan uso de todas y cada una de las columnas en conjunción de la base de datos. Finalmente, de entre todas las recuperaciones, aplicamos una medida de similaridad basada en la distancia euclídea de los coeficientes Haar almacenados. La ordenación posterior verificará la semejanza con la imagen de referencia.

[5x6]

aabbaa
aacdba
acdddb
acdddb
aacbaa

n Frame	Column 1	Column 2	Column 3	Column 4	Column 5
1	aabbaa	aacdba	acdddb	acdddb	aacbaa

Figura 3. 21. Indexación de una imagen en la base de datos

De la misma forma que ocurrió en las señales, demasiados tipos de primitivas puede ocasionar recuperaciones escasas, fundamentalmente debido a la rigidez de las mismas, teniendo que existir mucha información en la base de datos para que la probabilidad de encontrar patrones similares aumente. En dicho trabajo se probaron indexaciones con 4 y con 2 primitivas (imagen binaria cuyos pixeles son valores 0 y 1), obteniéndose resultados en tiempo y forma diferentes.

3.2.2.2 Patrones gráficos dentro de imágenes

Aprovechando el sistema de indexación descrito en el apartado anterior, la selección de un patrón gráfico dentro de una imagen supone construir una consulta a la base de datos mucho más compleja pero no por ello menos efectiva.

aabbaa
aacdba
acdddb
acdddb
aacbaa

n Frame	Column 1	Column 2	Column 3	Column 4	Column 5
1	aabbaa	aacdba	acdddb	acdddb	aacbaa
2	aacbaa	aacbaa	aacdba	aacbaa	acdddb
3	aabbaa	aacdba	acdddb	aacbaa	aacbaa
4	aacdba	aacbaa	aacbaa	aaaaaa	aaaaaa
5	aacbaa	aacbaa	aacdba	acdddb	acdddb

Figura 3. 22. Búsqueda de un patrón en una imagen

La búsqueda del patrón que se refleja en la Figura 3. 22 se realiza mediante la concatenación de tres consultas a la base de datos que englobaría en la primera consulta la columna 1, la columna 2 y la columna 3; la siguiente consulta englobaría las columnas 2 ,3 y 4; finalmente la última consulta correspondería a las columnas 3, 4 y 5. El método descrito es invariante a traslaciones y ha sido publicado en [Vega et al., 2009] e

implementado igualmente para imágenes pertenecientes a video-películas del JET. El autor ha contribuido además con una aplicación gráfica desarrollada en Matlab que facilita la visualización de las películas así como la selección de patrones dentro de las imágenes.

Veamos un ejemplo real de indexación, búsqueda y recuperación de un patrón perteneciente a una imagen del diagnóstico de esparcimiento Thomson del TJ-II, Figura 3. 23. Originalmente la imagen bruta de partida tiene unas dimensiones de 384x384 píxeles de información. Le aplicamos una transformación wavelet-Haar al nivel 2, reduciendo su información a una matriz de 96x96 píxeles. Posteriormente umbralizamos la imagen, eliminando los puntos de menor intensidad a un cierto nivel (por ejemplo, al 60%), destacando aquellos de mayor intensidad. Finalmente discretizamos las intensidades restantes en 2 primitivas atendiendo al valor máximo de intensidad para todo el conjunto de imágenes (intensidades cero hasta un cierto umbral, primitiva A y el resto de intensidades máximas, primitiva B).

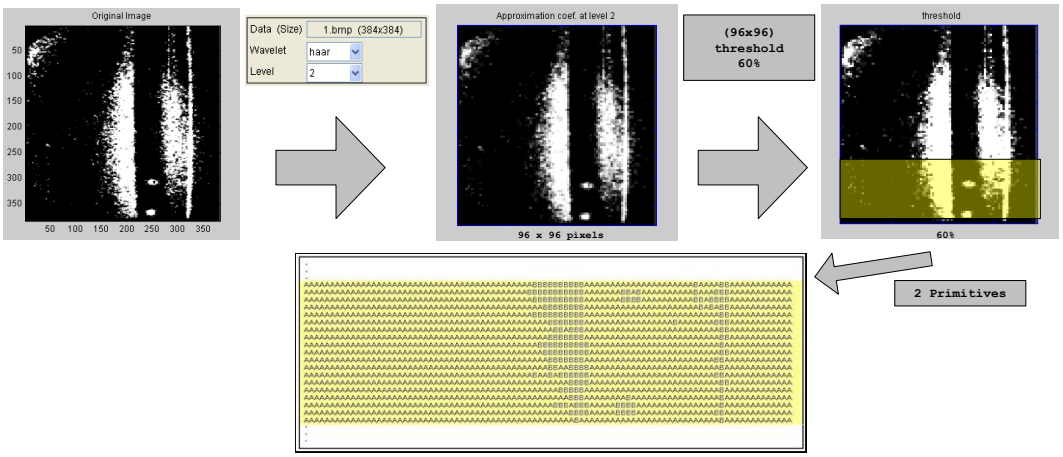


Figura 3. 23. Indexación, búsqueda y recuperación en imágenes

Posteriormente procedemos a introducir dichos valores en la base de datos, de forma que la primera fila corresponde a la primera columna, la segunda fila a la segunda columna y así sucesivamente hasta la fila 96, ver Figura 3. 24.

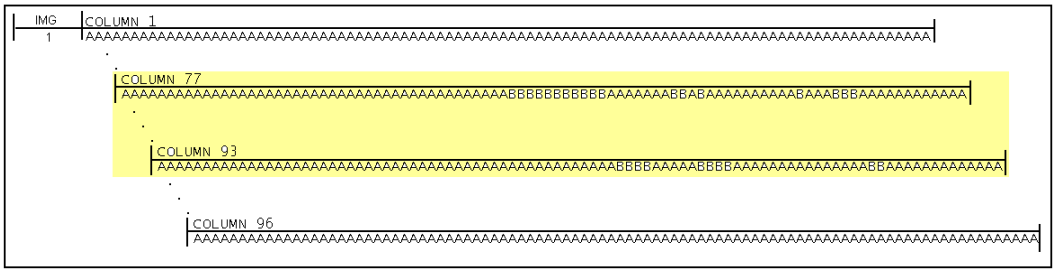


Figura 3. 24. Información almacenada en la base de datos

Añadimos a la base de datos una segunda imagen (4.bmp) realizando el mismo procedimiento. Sobre la primera imagen (1.bmp) realizamos una búsqueda muy localizada para un patrón característico, según se indica en la imagen correspondiente, Figura 3. 25. Se procede a convertir las coordenadas reales del patrón sobre la imagen original a las correspondientes para el nivel de transformación Haar que se ha decidido,

en el ejemplo que nos ocupa, el nivel 2 y seguidamente se construye sintácticamente la consulta a la base de datos.

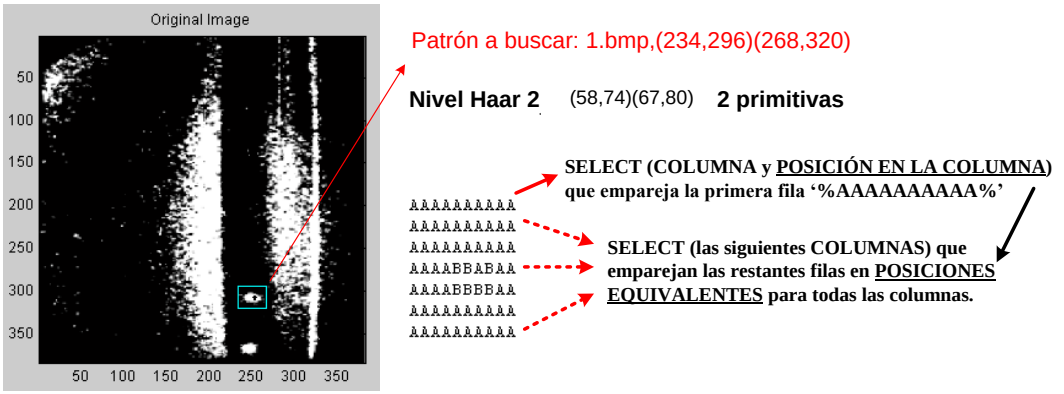


Figura 3. 25. Elaboración de una consulta

La ejecución de la consulta sobre esta base de datos (2 imágenes indexadas) devuelve tres patrones, el primero coincide con el patrón de referencia de la primera imagen, los otros dos residen en la segunda imagen y están localizados en posiciones diferentes, ver Figura 3. 26. De entre todos los emparejamientos recuperados se calcula finalmente la distancia euclídea con respecto al patrón de referencia teniendo en cuenta el valor de las intensidades en cada punto coincidente del patrón.

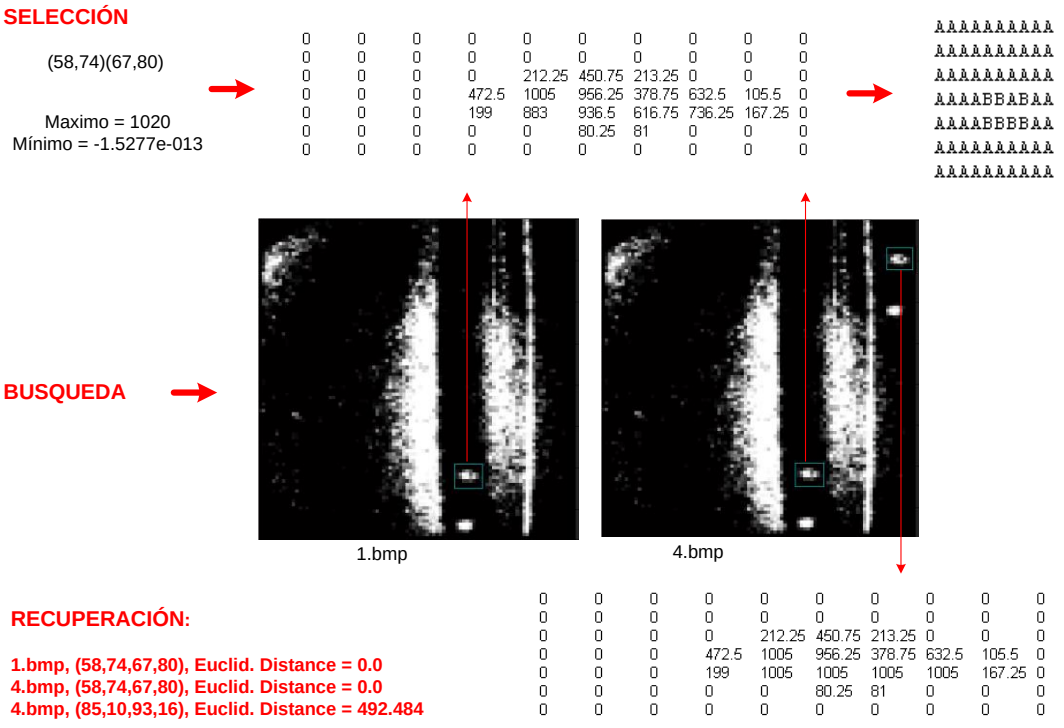


Figura 3. 26. Recuperación de un patrón

Los resultados de la aplicación de este método sobre imágenes reales del JET se pueden consultar en el apartado 6.2 de la presente memoria. Este método de indexación de la base de datos junto con el sistema de construcción de consultas aquí detallado se mejoró posteriormente mediante búsquedas optimizadas, tanto en la elaboración de la

consulta como en el diseño e implementación del sistema. Se hace apreciable que la elección del nivel de profundidad elegido en la descomposición Haar condiciona notablemente el tamaño de la base de datos, además debe existir un compromiso, ya que reducir demasiado el tamaño de la base de datos, supone que hemos elegido un nivel de descomposición muy elevado pero esto va en detrimento de la calidad en las recuperaciones finales. El sistema de construcción de consultas tiene una dificultad añadida, consultas de búsqueda de patrones que sean largos en altura y estrechos en anchura son realmente ineficientes, debido al excesivo número de consultas a realizar sobre la base de datos, tantas como el número de filas en origen del patrón. Además si en la primera SELECT, las primitivas están integradas solamente por umbrales residuales basadas en la primitiva A, que son las más numerosas, esto tendría una dificultad añadida, se aumentaría notablemente el tiempo empleado en la obtención de los resultados.

3.2.3 Consultas optimizadas de similaridad en bases de datos masivas

Con respecto a las señales de evolución temporal, como se comprobó en apartados anteriores, la elección de solamente 2 primitivas aumenta notablemente la recuperación de patrones semejantes, pequeñas cadenas de caracteres y compuestas por dos primitivas ‘a’ y ‘b’, son fácilmente localizables incluso aunque no existan muchas señales indexadas en la base de datos, no obstante, cadenas más largas, no son tan fáciles de identificar, ya que un simple desajuste de un carácter en la comparación puede comprometer la similaridad entre los dos patrones comparados. Si solo hiciéramos búsquedas exactas, patrones compuestos por una larga cadena de caracteres tendrán una baja probabilidad de ser encontrados. La probabilidad de que el patrón más largo posible sea encontrado para una longitud de 64 caracteres binarios sigue siendo muy baja incluso con solamente 2 caracteres en la codificación de las primitivas. Esto quiere decir que si hacemos una búsqueda exacta de los 64 caracteres, tendríamos que tener 1.8447e+19 de señales en nuestra base de datos para que posiblemente llegáramos a emparejar por lo menos un único patrón de esa longitud máxima. Por ello se hace necesario flexibilizar (sobre todo en subconjuntos largos de caracteres) todo lo que se pueda la consulta, para poder realizar suficientes emparejamientos y manteniendo una muy alta relación entre ellos al mismo tiempo.

En cuanto a la búsqueda de regiones similares en imágenes se hace intratable indexar tanta cantidad de información aún a sabiendas de las técnicas de reducción de datos explicadas en apartados anteriores. Métodos más optimizados de indexación que los vistos hasta el momento y una mejor distribución tanto de la carga de datos como de los procesos de búsqueda son explicados igualmente en el presente apartado.

3.2.3.1 Estrategias de búsqueda en señales de evolución temporal

Los pequeños saltos que se producen en los coeficientes delta y por tanto muy cercanos al valor 0, sean positivos o negativos, apenas suponen una distinción estructural

entre ellos, incluso puede suponerse que se trata solamente de ruido, ver Figura 3. 27. Si identificamos que coeficientes delta están muy próximos a cero, podemos tomar estos coeficientes y realizar consultas más complejas en las cuales dichos coeficientes puedan tomar indistintamente el valor negativo o positivo en la búsqueda a realizar. Con esto se flexibiliza mucho la consulta, aumentando el número de patrones recuperados, amplificando los resultados, desde la base de datos relacional.

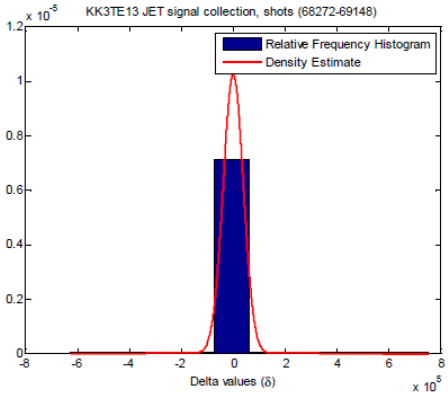


Figura 3. 27. Distribución de los valores delta

Atendiendo a la dificultad en la recuperación de formas de onda similares que responden a patrones más largos, se propuso en [Pereira et al., 2010b] lo que se denominan **consultas flexibles de similaridad**, ver Figura 3. 28. Estas consultas de similaridad las podemos dividir en dos tipos, **búsquedas exactas** o **búsquedas aproximadas**.

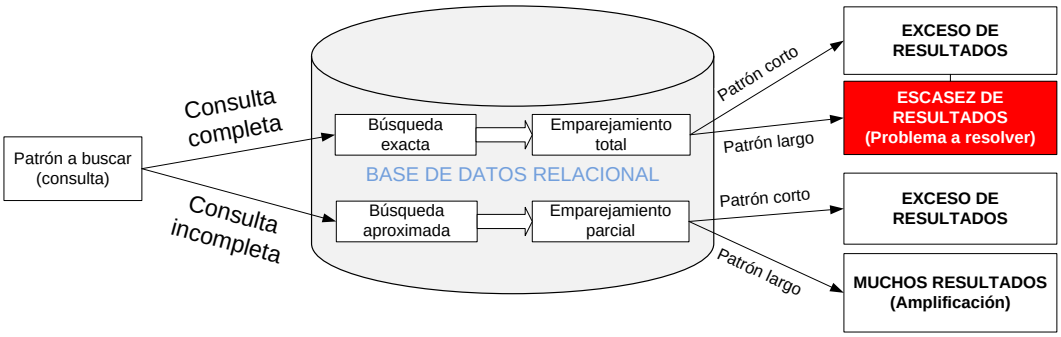


Figura 3. 28. Tipos de consultas a la base de datos

Las búsquedas exactas consisten en localizar un subconjunto de muestras consecutivas dentro de una señal que es coincidente en todos y cada uno de sus caracteres y para la misma posición con la consulta origen de referencia. Las búsquedas aproximadas localizan un subconjunto de muestras consecutivas que no tienen por qué coincidir con el carácter comparado de una misma posición de la consulta origen de referencia. El nivel de flexibilidad requerido en la consulta dependerá de la proporción de casos que caerán en la categoría central para cada tipo de señal. Mediante un análisis estadístico en cada señal se determina el rango de deltas donde las primitivas pueden tomar cualquier valor indistinto.

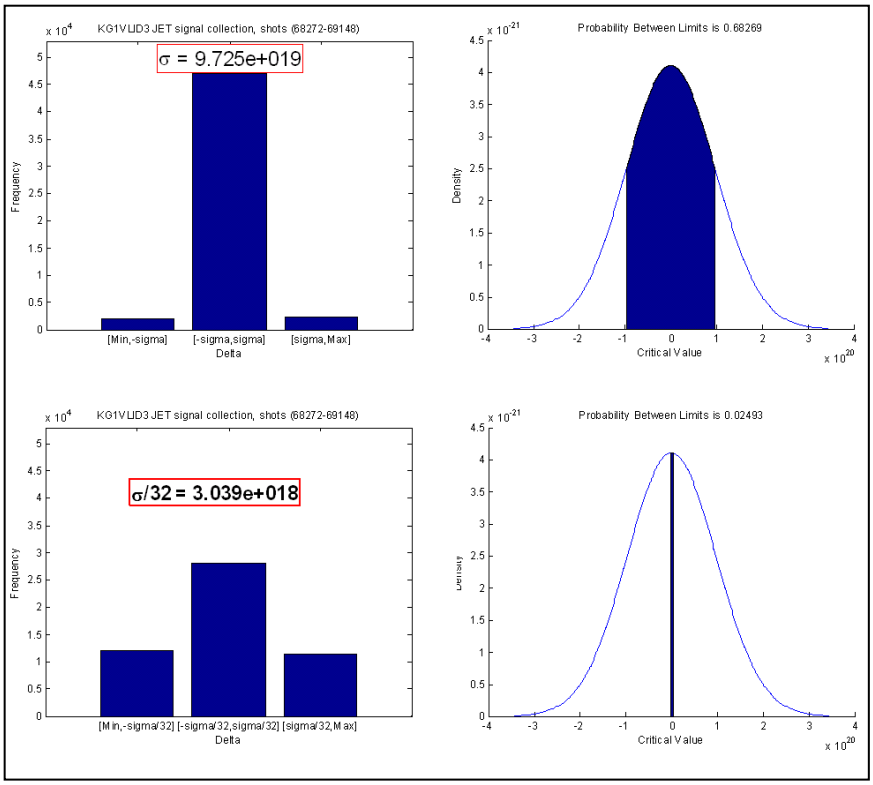


Figura 3. 29. Distribución de los deltas y densidad de probabilidad señal LID3 del JET

Para definir y acotar un límite positivo y negativo en el cual los valores delta puedan tomar el carácter indiferente (bien sea 'a' o 'b' indistintamente) recurrimos a la dispersión general (desviación típica) de todos los coeficientes delta almacenados en la base de datos para ese tipo de señal, ver Figura 3. 29. La desviación típica, es una medida de dispersión usada en estadística que nos dice cuánto tienden a alejarse los valores puntuales del promedio en una distribución. De hecho, específicamente, la desviación es, el promedio de la distancia de cada punto respecto de la media. Este valor en un conjunto de datos representa una medida de cuánto se desvían los datos de su media, y por tanto nos es muy útil para en función de un valor de esta sigma poder flexibilizar las consultas según necesitemos. Por ejemplo, un intervalo [-σ,σ] no es una buena elección, ya que serían unos límites muy amplios en la que casi todos los valores serían indiferentes 'a' o 'b'. Es necesario reducir el intervalo y acercarlo más al valor cero para concretar las primitivas donde se hacen más imperceptibles esas diferencias, aumentando así la proporción de las pendientes más pronunciadas, esto es, fuertemente positivas o negativas (las barras de los extremos en la imagen inferior). Después de haber analizado empíricamente muchas señales, la elección [-σ/32,σ/32] es muy conveniente y se ajusta bastante bien con resultados muy satisfactorios. No obstante, en la implementación mediante una aplicación Java y presentada en el trabajo [Pereira et al., 2010b] se ofrece la posibilidad de que sea el usuario el que flexibilice la consulta. Es entonces el usuario el que determina la cantidad y la calidad en las recuperaciones de similitud. Con esto se flexibiliza mucho la consulta, aumentando el número de patrones recuperados y amplificando los resultados de sub-patrones más largos desde la base de datos relacional. Finalmente, para identificar que sub-secuencia de las recuperadas es más similar, mediante un término más cualitativo con respecto al patrón de referencia, utilizamos la distancia euclidiana media como

función de similaridad para los coeficientes Haar que coinciden con la posición de los caracteres de ese patrón.

En la Figura 3. 30 se refleja todo el proceso de búsqueda de un patrón y el cálculo de sus similitudes. Es de resaltar que si no se hubiera flexibilizado la consulta a 1/8 de sigma no se hubiera encontrado ningún patrón semejante en otras señales, salvo el suyo propio de referencia. Se muestra el patrón original para poderlo comparar con el patrón flexible empleado en la consulta de búsqueda.

Patrón a buscar (perteneciente a la señal BOL5 12095): bbbbbbbbbababbbabb

Codificación Delta del patrón anterior
(Los coeficientes entre limitDown y limitUp definen la flexibilidad):

sigma de la familia BOL5: 2.3469485522401246
factor de sigma (definido por el usuario): 1/8

limitDown = -(factor de sigma) * (sigma): -0.29336856903001557
limitUp = (factor de sigma) * (sigma): 0.29336856903001557

1.0935364 0.53551865 0.7214756 0.77724266 1.365324 0.4253521 0.474308 0.08353424 0.44216537 -0.24660492 0.06099701 -0.449152 0.5157604 0.043748856 0.7741833 -0.25364304 0.7762165 0.40686035

Patrón flexible obtenido: bbbbbbb_b__ab_b_bb

Consulta a realizar:

select * from BOL5 where PRIMITIVES like '%bbbbbbb_b__ab_b_bb%'

RECUPERACION DE LA SELECT:

Patrón completo: bbbbbbbbbababbbabb
BOL5 12095 aabaabbaabababbbbbbbbbbbbababbbabbbaaaaabaabaabababababaabbabaa
Patrón flexible: bbbbbbb_b__ab_b_bb

Patrón completo: bbbbbbbbbababbbabb
BOL5 12094 abababaabbbbbbbbbbbbbbbbbaabbbbbbbaaaaabbaabaabababaababaabbab
Patrón flexible: bbbbbbb_b__ab_b_bb

Patrón completo: bbbbbbbbbababbbabb
BOL5 12076 abaaabababbbbbabbbbbbbbabbbabbbbbbbaaaaabbbabaabababaababababab
Patrón flexible: bbbbbbb_b__ab_b_bb

CALCULO DE LA SIMILITUD:

Coeficientes Haar del patrón:

BOL5 12095: 17.33995 18.433487 18.969006 19.690481 20.467724 21.833048 22.2584 22.732708 22.816242 23.258408 23.011803 23.0728 22.623648 23.139408 23.183157 23.95734 23.703697 24.479914
BOL5 12094: 17.916977 18.804163 19.576035 20.29543 20.995758 22.15168 22.84143 23.26261 23.58244 23.803488 24.25478 23.970913 23.66587 23.984425 23.99398 24.403572 24.742928 25.043581
BOL5 12076: 19.084679 19.60894 19.622728 20.277103 20.67527 21.185513 21.397366 21.682236 21.609581 22.161766 22.240763 22.387775 22.269394 23.1203 23.682467 24.670261 25.247355 26.398571

Función de similitud euclidiana:

BOL5 12095: f = 1.0000
BOL5 12094: f = 0.8406
BOL5 12076: f = 0.7794

Figura 3. 30. Proceso completo de búsqueda de un patrón

En el proceso descrito en la Figura 3. 31 se puede observar como una búsqueda flexible entre los límites [-σ/32,σ/32], permite encontrar muchos patrones no localizados en la misma posición temporal de la señal y en los que solamente en muchas señales cambia una única primitiva respecto al patrón de referencia, siendo igualmente similares.

Figura 3. 31. Búsqueda flexible de un patrón en la señal LID3 del JET

En la Figura 3. 32 se puede observar el patrón a buscar, perteneciente a la señal LID3 descarga 68598 del JET y en color azul una descarga diferente de la misma señal con un patrón equivalente y localizado en una región temporal muy diferente.

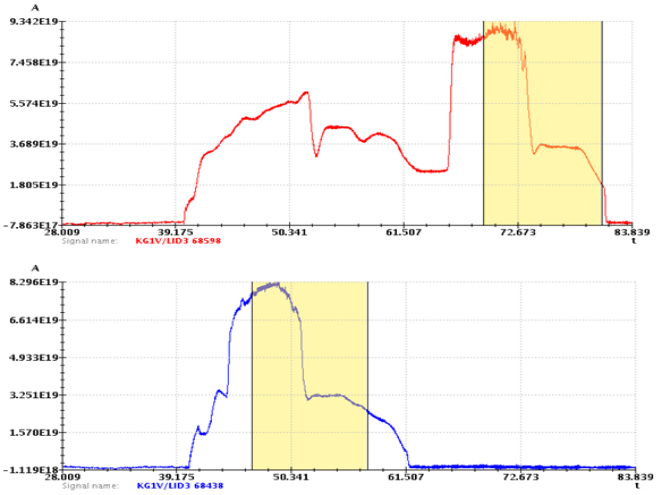


Figura 3. 32. Recuperación de un patrón similar en la señal LID3 del JET

3.2.3.2 Búsquedas optimizadas en imágenes y películas de vídeo

Con el objetivo de reducir el espacio de almacenamiento de información en la base de datos y el tiempo de búsqueda de patrones en imágenes, se propuso en [Vico, 2010] un nuevo método de indexación de primitivas, un planteamiento novedoso en la elaboración de las consultas de búsqueda y una nueva arquitectura distribuida basada en entornos de computación escalable tanto en el número no solo de equipos sino también en el número de procesos. En dicho trabajo se colaboró en la implantación, configuración e instalación final del sistema.

El método de indexación propuesto consiste en almacenar solamente las primitivas cuyas intensidades difieren del valor residual de fondo (primitiva A), incluyendo además como registros, la información del número de fila y las posiciones inferior y superior en los cuales aparecen los umbrales importantes para cada fila, Figura 3. 33.

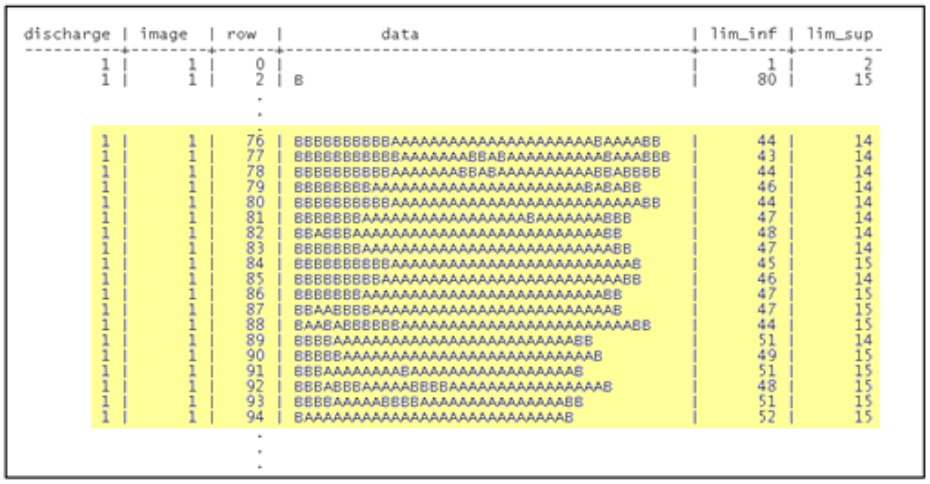


Figura 3. 33. Método de indexación mejorado

De esta forma, la búsqueda de cualquier patrón, independientemente de su dimensión, se hace mucho más ágil, al no contar con el valor residual de fondo tanto en la primera SELECT como en las sucesivas filas de las que se compone el patrón en origen, Figura 3. 34. La elaboración sintáctica de las consultas se hace algo más compleja debido a la gestión de los índices, no solo con las filas sino con la posición de la columna para cada una de las cadenas de primitivas.

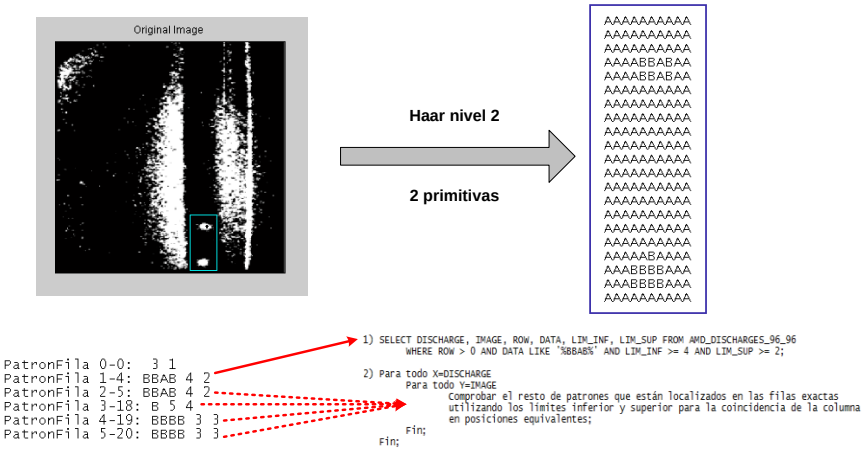


Figura 3. 34. Consultas optimizadas a la base de datos relacional

Finalmente se diseñó una arquitectura distribuida y escalable no solo para el almacenamiento de la información sino también para los procesos de búsqueda, paralelizando dicha búsqueda entre varios ordenadores, ver Figura 3. 35. El sistema está integrado por un conjunto de rutinas, programas distribuidores o '*dispatcher*' y aplicaciones servidoras, conectadas a bases de datos distribuidas PostGreSQL. Los programas distribuidores planifican y balancean el almacenamiento de los datos y se encargan también de elaborar las consultas a las bases de datos a petición de los diferentes clientes remotos. Las aplicaciones servidoras ejecutan los mandatos y consultas de los programas distribuidores, recuperando la información y entregándola a sus peticionarios. Todos estos procesos pueden ejecutarse en un mismo ordenador o cada uno de ellos en varios ordenadores distribuidos en un entorno de computación totalmente remoto.

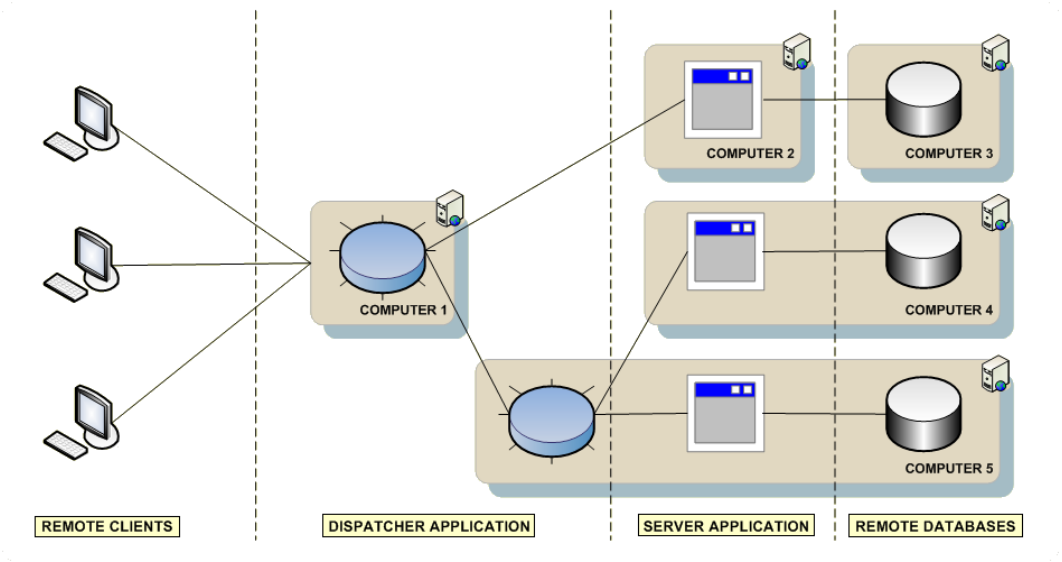


Figura 3. 35. Arquitectura distribuida y escalable para la recuperación de imágenes

Las dos pruebas realizadas y documentadas en [Vico, 2010] mediante la búsqueda de un patrón, recuperando 12 imágenes que contienen patrones similares, con diferentes configuraciones de distribución, ejecutadas solamente en un único ordenador (Intel Xeon E5450 @ 3.00 GHZ, 8 procesadores, 4 núcleos, 32 GB de RAM) y con una cantidad de información cercana tanto en tamaño como en número de películas a las pruebas realizadas y explicadas en el apartado 6.2, alcanzaban recuperaciones cercanas al segundo. Es de resaltar que la paralelización de los procesos y la distribución de las diferentes bases de datos relacionales necesita muchos recursos hardware y la disponibilidad de ordenadores para su uso, no obstante, la gestión inteligente en el balanceo y el almacenamiento de los datos permite reducir drásticamente el tiempo empleado en la búsqueda y presentación de la información a los diferentes clientes remotos.

Capítulo 4

Técnicas de aprendizaje para selección de características

Las técnicas de aprendizaje automatizadas, constituyen un tema vigente en las investigaciones actuales, sobre todo por el amplio espectro en que pueden ser aplicadas. La aplicación de diferentes métodos de selección de características incluye necesariamente la utilización de diferentes técnicas de aprendizaje automático, con el objetivo de conseguir y determinar los factores más relevantes en la evaluación que se haga de todos los datos disponibles. Bien sea utilizando algoritmos de clasificación o mediante ajustes de regresión, las predicciones obtenidas por estos algoritmos de aprendizaje son útiles para ponderar la importancia de los factores evaluados. Se incluye en el presente capítulo una explicación teórica de los diferentes métodos utilizados en la selección de características, los algoritmos de clasificación y regresión utilizados y aquellos otros que aportan más información a la predicción resultante con valores añadidos de confianza y credibilidad y cuyas salidas son conocidas como predicciones conformales.

4.1 Reducción de dimensionalidad y selección de características

La selección de características, también denominados atributos, componentes, variables, columnas, coordenadas o dimensiones, es un término usado habitualmente en minería de datos para describir las herramientas y las técnicas disponibles para reducir las entradas de los datos a un tamaño apropiado para su procesamiento y análisis. Se debe seleccionar o descartar activamente los atributos en función de su utilidad para el análisis. La capacidad de aplicar la selección de características es esencial para un análisis eficiente, ya que los conjuntos de datos suelen contener mucha más información de la necesaria para la generación del modelo, ocasionando degradar la calidad de los patrones a detectar. Tanto las variables ruidosas, como las redundantes o correlacionadas y las variables irrelevantes, dificultan la detección de patrones significativos a partir de los datos. Todo proceso de selección de atributos tiene un punto de partida, que puede ser el conjunto completo de atributos, el conjunto vacío o cualquier estado intermedio. Tras evaluar el primer subconjunto, se examinarán otros subconjuntos según una dirección de búsqueda, hacia adelante, hacia atrás, aleatoria o cualquier variación o mezcla de las anteriores. El proceso terminará cuando se recorra todo el espacio o cuando se cumpla una condición de parada, según la estrategia de búsqueda seguida. Existen otros métodos de selección de atributos que se basan más en la transformación de los valores de entrada que en técnicas optimizadas de búsqueda, aportando información de cuanto de relevante es cada variable en su conjunto, pudiendo descartar aquellas que sean irrelevantes o que estén por debajo de un cierto umbral de relevancia.

4.1.1 Dependencia entre atributos y correlación

Una de las principales premisas a tener en cuenta en la selección de características es que las variables independientes no posean ningún tipo de dependencia lineal entre ellas y no existan variables repetidas. Cuando una variable independiente posee alta correlación con otra u otras o puede ser explicada como una combinación lineal de alguna de ellas, se dice que el conjunto de datos presenta el fenómeno denominado **multi-colinealidad** [García et al., 2006]. La correlación entre variables dificulta[23] tanto los procesos de clasificación, variables innecesarias que aumentan los tiempos de cómputo, como las tareas de regresión, problemas de precisión en la estimación de los coeficientes.

[23] Colinealidad entre variables independientes. (Pag. 20-27), [Pereira, 2010]

- **Matriz de correlaciones**

Mediante el coeficiente de **correlación de Pearson** (4.1) se puede medir la relación lineal entre dos variables cuantitativas. El valor resultante se encuentra en el intervalo [-1,1] y es independiente de la escala de medida de las variables. Cuando $r = 0$, no existe ninguna correlación entre las variables y las variables independientes son ortogonales. Si r toma cualquier valor de los extremos del intervalo, existe una correlación lineal perfecta entre ellas, si el índice es negativo, la pendiente de la recta es negativa y positiva en caso contrario.

$$r = \frac{S_{xy}}{\sigma_x \sigma_y},\ S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \quad (4.1)$$

S_{xy} es la **covarianza** de las dos variables y σ_x σ_y las desviaciones típicas de cada distribución independientemente. Una forma muy práctica de determinar el grado de colinealidad entre todas las variables involucradas es la construcción de una **matriz de correlación**. Las variables se colocan en filas y en columnas y sus intercepciones deben presentar el coeficiente de regresión lineal de Pearson. Inicialmente se puede construir también una matriz de correlación con la covarianza en sus intercepciones, cada valor de la matriz indica el grado de variación conjunta de dos variables aleatorias, no obstante esta medida no suele ser de gran utilidad cuando las variables tienen diferente escala de medida. Asimismo, es de gran utilidad la construcción de una tercera matriz con los **diagramas de dispersión** de los datos para comprobar visualmente la lejanía o cercanía de dichos datos sobre la tendencia lineal que llegaran a mostrar.

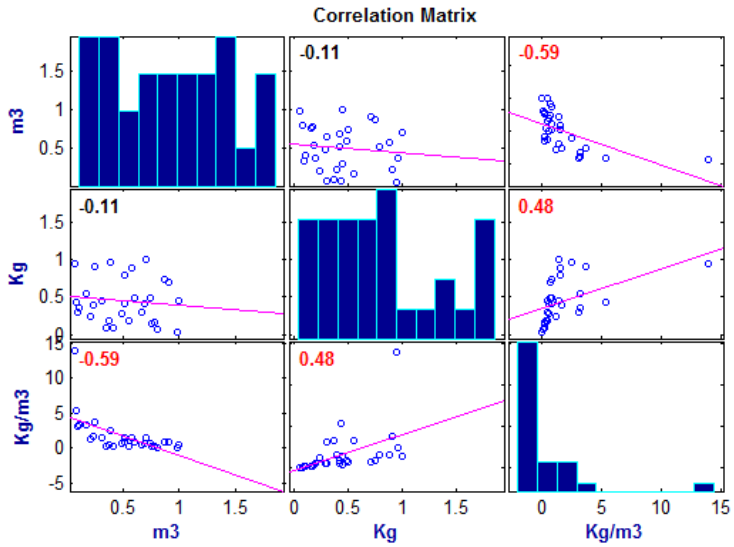


Figura 4. 1. Matriz de correlación y diagramas de dispersión

En la figura anterior se puede observar como ante dos variables aleatorias independientes y casi en ausencia de relación (*volumen* y *masa*, r = -0.11) y una tercera variable sintética, función de las otras dos *densidad* = *masa*/*volumen*, la matriz de correlación evidencia la colinealidad de la densidad con las otras dos variables (r = -0.59

y r = 0.48). En tareas de clasificación podemos prescindir de la tercera variable, ya que sería una variable redundante y el incluirla supondría aumentar la carga de datos y por consiguiente aumentar igualmente los tiempos de computación innecesariamente. En tareas de regresión, la conclusión más evidente es que la tercera variable está funcionando como variable dependiente de las dos primeras, si existiese una cuarta variable y la densidad la incluyéramos como variable independiente en nuestro análisis de regresión, esto ocasionaría imprecisión en los coeficientes de la ecuación paramétrica resultante tanto en los valores propios como en la polaridad de los mismos. El inconveniente del estudio de la colinealidad es que estos análisis solamente encuentran relaciones en los datos que sean lineales entre ellos dos a dos y no de otra índole como, tendencias y relaciones curvilíneas, polinomiales, exponenciales, etc., y que también son perjudiciales a la hora de generar modelos de clasificación y regresión. En esta investigación se han realizado análisis de correlación entre variables que están implicadas en la transición L/H en plasmas del JET con el objetivo de generar modelos paramétricos de regresión [González et al., 2012]. Por último, es de señalar que el coeficiente de correlación de Pearson coincide con la distancia del valor angular del coseno para datos normalizados respecto de la media, esto es, $mean(x)=mean(y)=0$, el coeficiente de correlación se convierte así, en el coseno entre ambos vectores centrados, cumpliéndose que:

- Si r = 1, el ángulo es *0°*, ambos vectores son colineales (paralelos).
- Si r = 0, el ángulo es de *90°*, ambos vectores son ortogonales.
- Si r = -1, el ángulo es de *180°*, ambos vectores son colineales de dirección opuesta.

- **Análisis de componentes principales**

El análisis de componentes principales (ACP) es una técnica proveniente del análisis exploratorio de datos cuyo objetivo es la síntesis de información, esto es, la reducción de la dimensión o el número de variables implicadas. Ante una tabla de datos con muchas variables (Figura 4. 2), el objetivo será reducirlas a un menor número de variables transformadas perdiendo la menor cantidad de información posible.

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{l1} & X_{l2} & \dots & X_{ln} \end{bmatrix} \rightarrow \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \vdots & \vdots & & \vdots \\ C_{l1} & C_{l2} & \dots & C_{ln} \end{bmatrix}$$

100% de la información 80% 16% 0.02%

Figura 4. 2. Transformación mediante ACP

Esta aproximación se basa en el hecho de que cualquier conjunto de variables pueden ser transformadas a otro conjunto de variables ortogonales y por tanto independientes entre sí, sin ninguna relación. Las nuevas variables ortogonales son conocidas como **componentes principales**. Estos nuevos componentes principales o factores, son calculados como una combinación lineal de las variables originales normalizadas y además serán linealmente independientes entre sí. Técnicamente, el ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados y construye una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño del

conjunto de datos es capturada en el primer eje, llamado el Primer Componente Principal, la segunda varianza más grande es el segundo eje, y así sucesivamente. La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. La transformación que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformación lineal necesaria para reducir la dimensionalidad de datos. Además las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales. Una de las ventajas del ACP para reducir la dimensionalidad de un grupo de datos, es que retiene aquellas características del conjunto de datos que contribuyen más a su varianza.

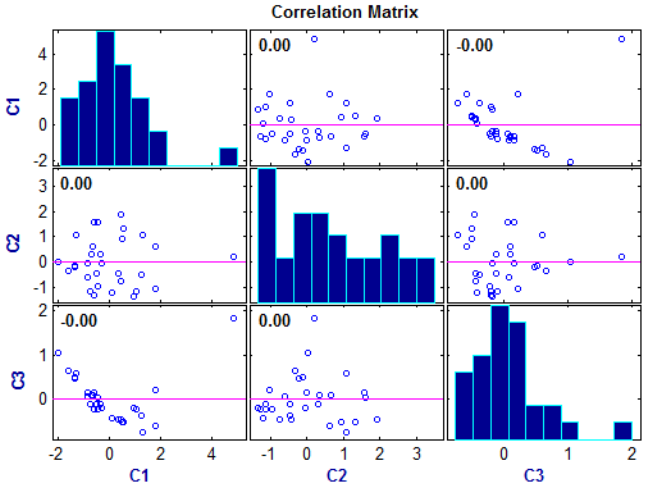


Figura 4. 3. Diagrama de dispersión de los componentes principales

La diagonal de la matriz de correlación de los componentes principales resultantes se denomina *eigenvalues* o **auto-valores** y los valores que no están en la diagonal de la matriz son ceros debido a que los componentes principales son ortogonales En la Figura 4. 3 se puede observar los coeficientes principales de las variables *volumen*, *masa* y *densidad*. Los auto-valores de estas variables son respectivamente [1.81, 0.89, 0.29] y el porcentaje de variabilidad total se calcula para cada auto-valor *i* de la forma:

$$\frac{100\ \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{4. 2}$$

Dichos porcentajes corresponden a [60.41, 29.77, 9.81], donde se puede comprobar como las dos primeras variables concentran más del 90% de la variabilidad total, pudiéndose despreciar la tercera componente correspondiente a la *densidad*. No obstante, si se decide conservar la tercera componente, nos aseguraríamos que estas son totalmente independientes y por tanto se puede trabajar en términos de mínimos cuadrados en análisis de regresión.

4.1.2 Búsqueda completa y exhaustiva

Para una base de datos con *n* atributos, existen 2^n subconjuntos candidatos. Una búsqueda exhaustiva en este espacio es totalmente ineficiente, incluso para bases de datos pequeñas, siendo necesario el uso de diferentes estrategias para atajar este problema. Esta búsqueda garantiza la localización del resultado más óptimo conforme a un criterio dado. Si para seleccionar los subconjuntos óptimos se ha de examinar todos los conjuntos posibles del espacio, coincidirá con la exhaustiva. Sin embargo, según la medida de evaluación utilizada o algún criterio a priori sobre las variables, puede no ser necesario examinar todos los subconjuntos posibles. Podemos decir que una búsqueda exhaustiva siempre es completa, pero a la inversa no se cumple en todos los casos[24]. Una implementación de búsqueda completa pero no exhaustiva fue realizada en [Vega et al., 2014]. En este trabajo, parte del mismo consistió en encontrar el mejor subconjunto de características de un total de catorce, utilizadas en la predicción de disrupciones. Todas las posibles combinaciones con un número de características entre 2 y 7 fueron tenidas en cuenta y el resto fueron descartadas a priori. El análisis combinatorio completo resultante supuso la evaluación de 9893 predictores, mucho menor de los 16384 necesarios si se hubiera realizado un análisis exhaustivo de las 14 características totales.

$$\sum_{n=2}^{7} C_{14,n} = 9893, \text{ where } C_{m,n} = \frac{m!}{n!(m-n)!} \quad (4.\ 3)$$

La búsqueda completa y exhaustiva es sencilla de implementar y, siempre encuentra la solución **mejor** de entre todas las posibles. Sin embargo, su coste de ejecución es proporcional al número de soluciones candidatas, el cual es exponencialmente proporcional al tamaño del problema. Por el contrario, se usa habitualmente cuando el número de soluciones candidatas no es elevado, o bien como método base cuando se desea comparar el desempeño de otros algoritmos meta-heurísticos, éstos son algoritmos semejantes con mejores tiempos de ejecución y buenas soluciones, usualmente las **óptimas**, converjan o no con las mejores o las **exactas**. Los mejores resultados alcanzados en [Vega et al., 2014] han servido como base de comparación de los resultados conseguidos en [Pereira et al., 2014] aplicando una técnica alternativa de selección de características fundamentada en los algoritmos evolutivos.

4.1.3 Selección hacia adelante y eliminación hacia atrás

Cuando la búsqueda de características no englobe ni la completa ni la exhaustiva es necesario introducir un criterio de finalización en el proceso y una estrategia de búsqueda.

[24] Búsqueda completa. (Pag. 40), [Ruiz., 2006]

Dicho proceso se convierte entonces en un proceso de optimización. Una métrica de evaluación que pondere el resultado de la predicción sobre las características elegidas en cada ejecución secuencial del proceso, dictaminará si la función criterio escogida ha llegado a la finalización del proceso de optimización. Por tanto diferentes métodos de selección de atributos con diferentes métricas de evaluación en las salidas de los algoritmos de búsqueda, obtendrá distintos subconjuntos de características finales y tiempos de cómputo muy desiguales. La solución óptima depende entonces de la configuración empleada. El hecho de no explorar todo el espacio de búsqueda implica que no se garantiza llegar a una solución inmejorable, pero sí que se pueden alcanzar resultados muy óptimos en tiempos de cómputo inferiores.

El método de búsqueda secuencial mediante la **selección hacia adelante**, parte de un conjunto de características vacío y en cada iteración añade al conjunto, la mejor característica de entre todas las evaluadas. El criterio de finalización puede ser o bien un número fijo de características alcanzadas o que no existan más características a añadir. El método de búsqueda secuencial mediante la **eliminación hacia atrás**, parte de un conjunto formado por todas las características disponibles y en cada iteración se elimina del conjunto la peor característica de entre las evaluadas. El criterio de parada es igualmente un número exacto de atributos o que no existan más características a eliminar. Efectivamente, tanto el método de selección hacia adelante como el método de eliminación hacia atrás tienen sus inconvenientes. Por ejemplo, no se pueden corregir adiciones o eliminaciones anteriores. Este problema puede ocasionar la elección de sub-conjuntos finales poco óptimos, se puede dar el caso que aplicando el método de eliminación hacia atrás, el mejor conjunto de tres variables independientes sean $\{x_2, x_3, x_4\}$, quedando en la siguiente iteración solamente $\{x_2, x_3\}$, no coincidiendo con el más óptimo conocido a priori $\{x_2, x_5\}$, debido a que x_5 fue eliminado en una etapa anterior.

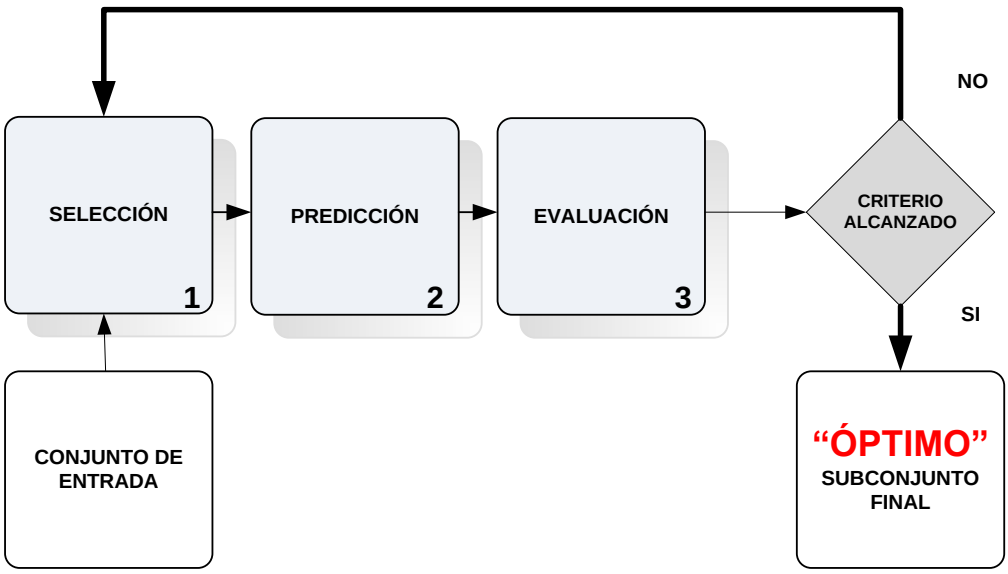


Figura 4. 4. Proceso de búsqueda de características

Diferentes métodos de evaluación en los resultados de las predicciones, bien sea mediante técnicas de clasificación o de regresión, determinará igualmente, distintas salidas en los subconjuntos finales. Por tanto, la salida del óptimo subconjunto final de

características dependerá no solo del método de selección de características elegido sino también del algoritmo de predicción a utilizar, así como del tipo de métrica usada en la evaluación que se haga a las salidas generadas por dichos predictores.

En [González et al., 2010] se seleccionaron ocho señales muy relevantes, para un conjunto total de 27 características, que caracterizan la transición de confinamiento desde regímenes bajos a regímenes altos de confinamiento (transición L/H), en los plasmas del Tokamak JET, utilizando una selección de atributos basada en el método de eliminación hacia atrás, para un total de 749 descargas utilizadas como conjunto de entrenamiento. El menor de los valores absolutos de todos los coeficientes que integran la ecuación del hiperplano de separación, obtenida por un clasificador SVM lineal, fue el criterio de evaluación seguido para eliminar la característica correspondiente en cada iteración. En dicho trabajo se colaboró con la implementación de un algoritmo basado en la extracción de los coeficientes no normalizados de la ecuación de separación, a partir del modelo generado en cada ejecución del proceso de entrenamiento mediante SVM (ver apartado 4.2.1). Finalmente se aplicó el clasificador más óptimo, formado por las ocho características ganadoras más relevantes, a un conjunto de 150 nuevas descargas y se obtuvieron resultados cercanos al 91% en tasas de acierto. Es de resaltar que el clasificador más óptimo finalmente seleccionado en dicho trabajo no es el mejor (con 24 características se alcanzan tasas de acierto del 94%), no obstante es insignificante el incremento de la tasa de acierto que se obtiene con la adicción de más características, provocando esto más complejidad en el clasificador, con el inconveniente añadido de mayores tiempos de cómputo en la obtención de los resultados para nuevas observaciones. Señalar también que, en este trabajo se utilizó una versión paralelizada del algoritmo SVM debido a la gran cantidad de información a utilizar y de los elevados tiempos de ejecución en el procesamiento de los mismos [Ramírez et al., 2010]. El mismo procedimiento de selección de características hacia atrás fue utilizado posteriormente en [Farias et al., 2012] para seleccionar las características más óptimas que intervienen en la transición L/H, pero en esta ocasión utilizando señales del tokamak DIII-D.

4.1.4 Búsqueda aleatoria y algoritmos genéticos

Las estrategias anteriores son deterministas, es decir, todas las veces que las ejecutamos devuelven los mismos resultados. La búsqueda aleatoria es no determinista, es decir, no se espera que en diferentes ejecuciones se proporcione exactamente el mismo resultado. Este tipo de estrategia, al contrario que las anteriores, busca a través del espacio formado por los atributos de manera aleatoria, es decir, no existe un estado siguiente o anterior que se pueda determinar según alguna regla. En este tipo de búsquedas se pretende usar el carácter aleatorio para no caer en mínimos locales e incluso moverse temporalmente a otros estados con peores soluciones. Esta búsqueda se puede detener en cualquier momento, por lo que es difícil determinar su coste. Normalmente, el criterio de parada es un número de iteraciones, y aunque a priori pueda dar la impresión de que no se puede tener garantías de que el subconjunto elegido sea el óptimo debido a la aleatoriedad del proceso, sí se puede demostrar que puede llegar a converger al mejor subconjunto con las mejores tasas de acierto. Existen evidencias empíricas de que utilizando algoritmos aleatorios se encuentran soluciones de un nivel aceptable, en un

tiempo competitivo aceptable, comparándolos con el resto de algoritmos de optimización combinatoria.

Los Algoritmos Genéticos (AG) son métodos adaptativos basados en alteraciones y recombinaciones aleatorias que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos y una analogía directa con el comportamiento natural. A lo largo de las generaciones, las poblaciones evolucionan en la naturaleza de acorde con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin (1859). En la naturaleza, los individuos de una población compiten entre sí en la búsqueda de recursos tales como comida, agua y refugio. Incluso los miembros de una misma especie compiten a menudo en la búsqueda de un compañero. Aquellos individuos que tienen más éxito en sobrevivir y en atraer compañeros tienen mayor probabilidad de generar un gran número de descendientes. Por el contrario individuos poco dotados o enfermos producirán un menor número de descendientes. Esto significa que los genes de los individuos mejor adaptados se propagarán en sucesivas generaciones hacia un número de individuos creciente. La combinación de buenas características provenientes de diferentes ancestros, puede a veces producir descendientes "súper-individuos", cuya adaptación y supervivencia es mucho mayor que la de cualquiera de sus ancestros. De esta manera, las especies y las sucesivas generaciones evolucionan logrando unas características cada vez mejor adaptadas al entorno en el que viven.

Los principios básicos de los AG aplicados en ingeniería informática fueron establecidos en [Holland, 1975], y se encuentran bien descritos en varios textos y aplicados a muchos y diferentes proyectos [Whitley et al., 1988], [Haupt et al., 2007].

El **Algoritmo Genético Simple** (AGS), también denominado **Canónico**[25], se representa en la siguiente figura.

```
BEGIN /* Algoritmo Genético Simple */
  Generar una población inicial.
  WHILE NOT Terminado DO
  BEGIN /* Producir nueva generación */
    FOR Tamaño población  DO
    BEGIN /* Ciclo Reproductivo */
      Seleccionar dos individuos de la anterior generación, para el cruce
           (probabilidad de selección proporcional a la función de
           evaluación del individuo).
      Cruzar con cierta probabilidad los dos individuos obteniendo dos
           descendientes.
      Mutar los dos descendientes con cierta probabilidad.
      Computar la función de evaluación de los dos descendientes
           mutados.
      Insertar los dos descendientes mutados en la nueva generación.
    END
    IF la población ha convergido THEN
      Terminado := TRUE
  END
END
```

Figura 4. 5. Algoritmo genético simple

Como se verá a continuación, se necesita una codificación o representación del problema, que resulte adecuada al mismo. Además se requiere una **función de ajuste** o adaptación al problema, también denominada **función de evaluación**, la cual asigna un número real a cada posible solución codificada. Durante la ejecución del algoritmo, los

[25] Métodos Matemáticos en Ciencias de la Computación. AG, (Tema 2. Pág. 3)

padres deben ser seleccionados para la reproducción, a continuación dichos padres seleccionados se cruzarán generando dos hijos, sobre cada uno de los cuales actuará un operador de mutación. El resultado de la combinación de las anteriores funciones será un conjunto de **individuos** o posibles soluciones al problema, los cuales en la evolución del AG formarán parte de la siguiente **población**. Se supone que los individuos, pueden representarse como un conjunto de parámetros que denominaremos **genes**, los cuales agrupados forman una ristra de valores a menudo referida como **cromosoma**. Si bien el alfabeto utilizado para representar los individuos no debe necesariamente estar constituido por el (0,1) buena parte de la teoría en la que se fundamentan los AG utiliza dicho alfabeto.

En términos biológicos, el conjunto de parámetros representando un cromosoma particular se denomina **fenotipo**. El fenotipo contiene la información requerida para construir un organismo, el cual se refiere como **genotipo**. Los mismos términos se utilizan en el campo de los AG. La adaptación al problema de un individuo depende de la evaluación del genotipo. Esta última puede inferirse a partir del fenotipo, es decir, puede ser computada a partir del cromosoma, usando la función de evaluación.

La función de adaptación debe ser diseñada para cada problema de manera específica. Dado un cromosoma particular, la función de adaptación le asigna un número real, que se supone refleja el nivel de adaptación al problema del individuo representado por el cromosoma. Durante la fase reproductiva se seleccionan los individuos de la población para cruzarse y producir descendientes, que constituirán, una vez mutados, la siguiente **generación** de individuos. La selección de padres se efectúa al azar usando un procedimiento que favorezca a los individuos mejor adaptados, ya que a cada individuo se le asigna una probabilidad de ser seleccionado que es proporcional a su función de adaptación. Este procedimiento se dice que está basado en la **ruleta sesgada** o también conocido como **rueda de la ruleta**. Según dicho esquema, los individuos bien adaptados se escogerán probablemente varias veces por generación, mientras que los pobremente adaptados al problema, no se escogerán más que de vez en cuando. Una vez seleccionados dos padres, sus cromosomas se combinan, utilizando habitualmente los operadores de cruce y mutación. Las formas básicas de dichos operadores se describen a continuación. El operador de **cruce**, coge dos padres seleccionados y corta sus ristras de cromosomas en una posición escogida al azar, para producir dos sub-ristras iniciales y dos sub-ristras finales. Después se intercambian las sub-ristras finales, produciéndose dos nuevos cromosomas completos. Ambos descendientes heredan genes de cada uno de los padres. Este operador se conoce como operador de cruce basado en un punto. Habitualmente el operador de cruce no se aplica a todos los pares de individuos que han sido seleccionados para emparejarse, sino que se aplica de manera aleatoria, normalmente con una probabilidad comprendida entre 0.5 y 1.0. En el caso en que el operador de cruce no se aplique, la descendencia se obtiene simplemente duplicando los padres. El operador de **mutación** se aplica a cada hijo de manera individual, y consiste en la alteración aleatoria, normalmente con probabilidad pequeña, de cada gen componente del cromosoma. Si bien, puede en principio pensarse que el operador de cruce es más importante que el operador de mutación, ya que proporciona una exploración rápida del espacio de búsqueda, este último asegura que ningún punto del espacio de búsqueda tenga probabilidad cero de ser examinado, y es de capital importancia para asegurar la convergencia de los AG. Existen muchos criterios y variantes prácticas para llegar a la convergencia. Si el AG ha sido correctamente implementado, la población evolucionará a lo largo de las generaciones sucesivas de tal manera que la adaptación media extendida a todos los individuos de la población, así como la adaptación del mejor individuo se irán incrementando hacia el óptimo global. El concepto de convergencia está relacionado con

la progresión hacia la uniformidad. A medida que el número de generaciones aumenta, es más probable que la adaptación media se aproxime a la del mejor individuo.

4.1.4.1 Configuraciones y restricciones impuestas al AGS

Para el estudio de los AG y su aplicación práctica, hay que tener en cuenta una serie de parámetros. Estos parámetros son configurables en cualquier AGS, además es necesario imponer unas restricciones. Generalmente, el criterio de parada puede ser la uniformidad de convergencia explicada anteriormente, pero existen otros muchos. Los pasos básicos de un AG se pueden repetir hasta que se dé una condición de terminación fijando un número máximo de iteraciones o detenerlo cuando no se produzcan más cambios en la población. Una variante del AGS es el que se presenta a continuación.

```
BEGIN                    /* Algoritmo Genético Adaptado */
  Obtener                /* Población al azar */
  REPEAT
    BEGIN
      Predicción         /* De la toda la población */
      Evaluación         /* Función objetivo */
      Selección          /* De los mejores individuos */
      Cruzamiento        /* Entre pares */
      Mutación           /* A nivel de gen */
    END
  UNTIL (Criterio alcanzado)
END
```

Figura 4. 6. Algoritmo genético adaptado

En el AGS los operadores genéticos de reproducción, cruzamiento y mutación, son aplicados antes que la función de adaptación. El **Algoritmo Genético Adaptado** (AGA) se acerca más al planteamiento expuesto en el organigrama de la Figura 4. 4. y tiene una configuración algo diferente del AGS. Básicamente, la selección o reproducción de los individuos no se hace dos a dos, sino que se aplican los operadores genéticos básicos a todos los individuos previamente evaluados [Kumar et al., 2014]. Esta configuración fue utilizada en el trabajo [Pereira et al., 2014] para seleccionar las mejores características utilizadas para la predicción de disrupciones en plasmas del JET.

- **Tamaño de la población**

Este parámetro nos indica el número de cromosomas o individuos que tenemos en nuestra población para una generación determinada. En caso de que esta medida sea insuficiente, el AG tiene pocas posibilidades de realizar reproducciones, con lo que se realizaría una búsqueda de soluciones escasa y poco óptima. Por otro lado si la población es excesiva, el algoritmo genético será muy lento. De hecho estudios revelan que hay un límite a partir del cual es ineficiente elevar el tamaño de la población puesto que no se consigue una mayor velocidad en la resolución del problema. En el trabajo [Alander,

1992], basándose en evidencia empírica, se sugiere un tamaño de población λ, comprendida entre l='número de genes' y $2l$ sea suficiente para atacar con éxito los AG. En cuanto a la población inicial, habitualmente se escoge generando ristras al azar, pudiendo contener cada gen uno de los posibles valores del alfabeto con probabilidad uniforme. En [Pereira et al., 2014] se utilizó una longitud para cada individuo $l=14$ genes y un tamaño de población $\lambda=2l=28$ individuos.

- **Probabilidad de cruce**

El **cruzamiento**, consiste en el intercambio de material genético entre dos cromosomas. El objetivo del cruce es conseguir que el descendiente herede y mejore las características de sus padres. La probabilidad de cruzamiento indica la frecuencia con la que se producen cruces entre los cromosomas padre, es decir, que haya probabilidad de reproducción entre ellos. En caso de que no exista probabilidad de reproducción, los hijos serán copias exactas de los padres. En caso de haberla, los hijos tendrán **alelos** (subconjunto secuencial de genes coincidentes en la misma posición para cada individuo) de los cromosomas de los padres. El intercambio de alelos puede variar y existen diferentes estrategias para materializar dicho intercambio. Como se ha visto en párrafos anteriores, el operador de intercambio basado en un punto es el más común. En la tesis doctoral [Jong, 1975], se investigó el comportamiento del operador de cruce basado en múltiples puntos, concluyendo que el cruce basado en dos puntos, podía representar una mejora mientras que añadir más puntos de cruce no beneficiaba el comportamiento del algoritmo. Otra opción de cruzamiento es seleccionar una máscara u operador de cruce uniforme. En caso de que el bit correspondiente a la máscara esté a 1, se copia el gen de un progenitor y en caso de que esté a 0 se copia el gen del otro progenitor. No obstante, la elección de la máscara también supone un compromiso. Por otra parte, la idea de que el cruce debería de ser más probable en algunas posiciones ha sido descrita por varios autores. Nuevamente, la heurística llevada a cabo en este tema se ciñe al problema a tratar y debería estar justificada dicha decisión. En [Pereira et al., 2014] se ha optado por el operador de intercambio basado en un punto mediante una posición dentro del cromosoma cambiante aleatoria, no fija, y con una probabilidad de cruzamiento $P_{cruzamiento} = 0.55$ entre dos individuos.

- **Probabilidad de mutación**

Tras el cruce, tiene lugar la mutación. Si nos referimos en términos de evolución, la mutación se manifiesta de forma extraordinaria, nada común. Las mutaciones suelen en promedio ser beneficiosas pues contribuyen a la diversidad genética de la especie. Además previenen a las soluciones de la población de verse limitadas por un óptimo local. Por lo tanto la mutación consiste en modificar ciertos genes de forma aleatoria atendiendo a la probabilidad de mutación establecida con anterioridad. La mutación depende de la codificación y de la reproducción. Si se abusa de la mutación, podemos caer en el uso del AG como una simple búsqueda aleatoria. La probabilidad de mutación debe ser baja. Ésta nos indica la frecuencia con la que los genes de un cromosoma son mutados. Si no hay mutación, los descendientes son los mismos que había tras la reproducción. En caso de que haya mutaciones, parte del cromosoma descendiente es modificado y si la probabilidad de mutación es del 100%, la totalidad del cromosoma se cambia. En este caso, no se cambian simplemente unos bits del cromosoma sino que se cambian todos, lo que significa que se produce una inversión en el cromosoma y no una

mutación por lo que la población degenera muy rápidamente. La búsqueda del valor óptimo para la probabilidad de mutación, es una cuestión que ha sido motivo de varios trabajos. Así, en [Jong, 1975] se recomienda la utilización de una probabilidad de mutación del gen de l^{-1}, siendo l la longitud del cromosoma, como se ha visto anteriormente. En otros trabajos [Schaffer et al., 1989], se utilizan resultados experimentales para estimar la tasa óptima, proporcional a $1/\lambda^{0.9318}l^{0.4535}$, donde λ denota el número de individuos en la población. Si bien en la mayoría de las implementaciones de AG se asume que tanto la probabilidad de cruce como la de mutación permanecen constantes, en otras ocasiones se han obtenido mejores resultados experimentales modificando la probabilidad de mutación a medida que aumenta el número de iteraciones. La tasa de mutación para un gen que se ha utilizado en [Pereira et al., 2014] fue la siguiente.

$$P_{mutación} = 1/mean(\lambda + l) = 2/(\lambda + l) = 2/(3l) \approx 0.05 \quad (4.4)$$

Esta probabilidad de mutación, empíricamente con mejores resultados en el caso que nos ocupa, es mayor que la sugerida por [Schaffer et al., 1989] y menor que la sugerida por [Jong, 1975] y cercana al valor medio de las dos.

$$\frac{1}{\lambda^{0.9318}l^{0.4535}} < \frac{2}{3l} < l^{-1}, \quad para\ l = 14, \lambda = 2l \quad (4.5)$$

- **Método de selección**

Como ya hemos visto anteriormente, es necesario hacer una selección con los individuos más capacitados para que éstos sean los que se reproduzcan con más probabilidad de acuerdo con la teoría de Darwin, en la cual los más capacitados son los que deben sobrevivir y crear una nueva descendencia más facultada y mejor adaptada al medio. Por lo tanto, una vez evaluado cada cromosoma y obtenida su puntuación, se tiene que crear la nueva población teniendo en cuenta que los buenos rasgos de los mejores se transmitan a ésta. Esta selección se puede realizar de varias formas. La selección por rueda de ruleta es la más utilizada. Cada cromosoma tendrá una parte de esa ruleta mayor o menor en función de la puntuación que tenga cada uno. Se hace girar la ruleta y se selecciona el cromosoma en el que se para la ruleta. Obviamente el cromosoma con mayor puntuación saldrá con mayor probabilidad. En caso de que las probabilidades difieran mucho, este método de selección dará problemas puesto que si un cromosoma tiene un 90% de posibilidades de ser seleccionado, el resto apenas saldrá, lo que reduciría la diversidad genética. La selección **elitista**, puede mejorar el funcionamiento de los AG al evitar que se pierda la mejor solución para cada generación. La selección por **escalada** consiste en hacer más discriminadora la función de aptitud, ya que al incrementarse la aptitud media de la población, la fuerza de la presión selectiva también aumenta. Este método puede ser útil para seleccionar más tarde, cuando todos los individuos tengan una aptitud relativamente alta y sólo les distingan pequeñas diferencias en la aptitud. En [Baker, 1987], se introduce un método denominado **muestreo universal estocástico**, el cual utiliza giros de la ruleta mínimos. Este método de reproducción fue el utilizado en [Pereira et al., 2014]. Los individuos son seleccionados a partir del marcador de la ruleta, proporcional entre 0 y 1 a los valores máximo y mínimo del valor de la función de ajuste para cada individuo en cada generación. Se selecciona en cada muestreo todos los

individuos que superan el valor de la ruleta. Se repite el procedimiento hasta que se alcanzan el número de individuos que integran la población entera. Con este método se consigue que para cada generación, se dupliquen o repitan varias veces los mejores individuos y sean descartados los individuos que estén por debajo del valor de la ruleta para cada muestreo.

4.1.4.2 La importancia de la función de evaluación

La función de evaluación, también conocida como **función de ajuste**, **función de aptitud**, **función de adaptación** o **función objetivo**, es fundamental para hacer converger los AG en el menor tiempo posible. Si los parámetros de configuración anteriormente descritos son importantes, la función de evaluación del AG se hace todavía aún más fundamental. Se puede decir que la rápida o lenta adaptación y convergencia para encontrar las mejores soluciones del problema también depende de la elección que se haga de esta función. Esta función debe ser capaz de "castigar" a las malas soluciones y de "premiar" a las buenas, de forma que sean estas últimas las que se propaguen con mayor rapidez. Esta función es la clave para que el AG sea útil y conduzca a una solución tempranamente exitosa. Los conceptos de buenos y malos candidatos son evaluados por ecuaciones o métricas. Estás métricas asignarán una puntuación a cada individuo para poder cuantificarlas y poder compararlas con el objetivo final de discernir y poder seleccionar los mejores candidatos en función de esa puntuación. La regla general para construir una buena función objetivo es que ésta debe reflejar el valor del individuo de una manera lo más real posible. Las salidas generadas por los clasificadores o por otros predictores utilizados en regresión, se convierten en las entradas de la función de ajuste. La evaluación que haga dicha función de esas entradas, generará una única salida que se convertirá en la puntuación obtenida para un individuo concreto. Tradicionalmente, las tasas de acierto globales directamente obtenidas por los clasificadores, se utilizaban como funciones de ajuste para los AG [Rattá et al., 2012], como se explicará posteriormente en el apartado 4.4.1. Diferentes ecuaciones que incluyen otras salidas, ponderando y teniendo más en cuenta las tasas de error, también obtenidas por los algoritmos de aprendizaje, permitirán optimizar mucho mejor y en menor tiempo los resultados de los AG. En [Pereira et al., 2014] se exponen y evalúan cinco métricas diferentes a modo de función de ajuste, utilizando para cada caso y para su comparación, la misma población inicial aleatoria y el mismo algoritmo de clasificación. Los resultados y las conclusiones finales se explican y se exponen detalladamente en el capítulo 7.

4.2 Algoritmos de clasificación

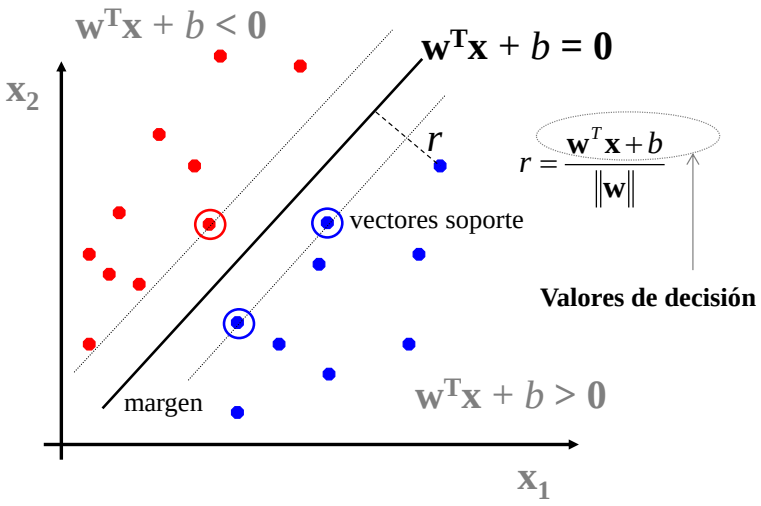
Las máquinas de vectores soporte (SVM, siglas procedentes del inglés) son herramientas basadas en algoritmos de aprendizaje supervisado y deducen por medio de una función generada a partir de unos datos de entrenamiento catalogados previamente por un tutor. Están muy desarrolladas y extendidas en el campo de la inteligencia artificial y utilizada en tareas tanto de clasificación como de regresión. En el trabajo [Vega et al., 2005], se plasmó la realización de un clasificador automático SVM mediante funciones de base radial (RBF, siglas procedentes del inglés) capaz de identificar con elevada precisión, las imágenes resultantes del diagnóstico de esparcimiento Thomson del TJ-II. Las funciones *kernel* [An et al., 2007] son funciones matemáticas que permiten convertir lo que sería un problema de clasificación no lineal en el espacio dimensional original, a un sencillo problema de clasificación lineal en un espacio de dimensión mayor. En la publicación [González et al., 2010], se utilizó un clasificador SVM, mediante un *kernel* de separación lineal, como método de selección de las mejores características para la predicción del instante de la transición L/H en plasmas del JET. En dicho trabajo se colaboró con un algoritmo que extrae los coeficientes de la ecuación del hiperplano lineal a partir del modelo generado en la etapa de entrenamiento del clasificador (que está basado en el software ***libsvm***) y paralelizado en un entorno de supercomputación Linux de altas prestaciones [Ramírez et al., 2010]. Para cada modelo generado, la característica o atributo con el menor de los coeficientes en la ecuación del hiperplano obtenido es descartado, en un proceso repetitivo de selección de características, utilizando la técnica de eliminación hacia atrás y comenzando con un total de 27 señales a modo de características. Este método fue aplicado a una base de datos integrada por 749 descargas del JET y probado en otra base de datos diferente constituida por 150 descargas adicionales.

4.2.1 Máquinas de vectores soporte

Dado un conjunto de puntos en el que cada uno pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo que predice si un punto nuevo pertenece a una categoría o a la otra. Los puntos de entrada son vistos como un vector n-dimensional. La separación en categorías se realiza mediante la construcción de un hiperplano n-dimensional que separa de forma óptima los datos. La distancia de un punto p a un plano Π, es la menor de las distancias desde ese punto de referencia a los infinitos puntos del plano. Esta distancia corresponde a la perpendicular trazada desde el punto al plano.

$$d(p,\pi) = \frac{w_1 x_1 + w_2 x_2 + b}{\sqrt{w_1^2 + w_2^2}} para\ 2\ dimensiones, \tag{4. 6}$$

$$en\ general \quad r = \frac{w^T x + b}{\|w\|}\ para\ n\ dimensiones \tag{4. 7}$$



Función de decisión = signo (**Valores de decisión**)

Figura 4. 7. Función de decisión SVM

En el ejemplo de la Figura 4. 7, idealizado para 2 dimensiones, la representación de los datos a clasificar se realiza en el plano x_1x_2. El algoritmo SVM trata de encontrar un hiperplano, en el ejemplo que nos ocupa es una línea, que une a las variables predictoras y constituye el límite que define si un elemento de entrada pertenece a una categoría o a la otra. Existe un número infinito de posibles hiperplanos o líneas que realicen la clasificación pero, ¿cuál es la mejor y cómo la definimos?. La mejor solución es aquella que permita un **margen**[26] máximo entre los elementos de las dos categorías. Se denominan **vectores de soporte** a los puntos que conforman las dos líneas paralelas al hiperplano de ese margen, siendo la distancia entre ellas la mayor posible. En el mismo ejemplo de la figura se puede comprobar que la distancia de cualquier punto azul al hiperplano es siempre positiva, esto quiere decir que los **valores de decisión** son también positivos, mientras que la distancia de cualquiera de los puntos rojos al mismo hiperplano es siempre negativa, al igual que sus valores de decisión. Por tanto, la **función de decisión** a tener en cuenta en tareas de clasificación consistirá en fijarse en el signo que toman los valores de decisión para cada punto. El modelo basado en SVM produce un hiperplano que separa completamente los datos del problema estudiado en dos categorías mediante la función de decisión. Un hiperplano lineal constituido por *m* características tiene la siguiente expresión:

$$w_1x_1 + w_2x_2 + \cdots + w_mx_m + b = 0 \qquad (4.\ 8)$$

Donde w_j es el peso de cada característica x_j, siendo b, el valor del término independiente. La idea que subyace para utilizar dicha ecuación como criterio para la selección de características por eliminación, es la de ir eliminando el peso más pequeño de entre todos los que forman el vector de coeficientes *w* mediante un proceso repetitivo y empezando con las *m* características disponibles para cada una de las ecuaciones calculadas en cada iteración. Esto es, el valor más pequeño de los coeficientes, correspondería a la característica más irrelevante y sabiendo que un valor muy próximo a cero correspondería a una característica totalmente irrelevante. La solución a dicho problema[27] toma siempre la forma:

[26] The role of the margin. Página 192. [Scholkopf et al., 2002]
[27] Optimal separating hyperplane. Página 423. [Cherkassky y Mulier, 2007]

$$w_j = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (4.\ 9)$$

Los y_i corresponden a las etiquetas de las respectivas clases (+1,-1, en una clasificación binaria) para cada observación. Los α_i son las soluciones al problema de optimización cuadrática y resuelto por SVM[28], muchos de los cuales son cero. Los vectores asociados a los multiplicadores no nulos, corresponden con los vectores soporte. Por tanto los demás no son necesarios a efectos de la clasificación, quedando la ecuación de los coeficientes de la forma,

$$w_j = \sum_{i=1}^{VS} \alpha_i y_i x_i \quad (4.\ 10)$$

La expresión matemática final de la función de decisión quedaría:

$$D(x) = signo(w^T x + b) = signo\left(\sum_{i=1}^{VS} \alpha_i y_i (x_i x) + b\right) \quad (4.\ 11)$$

Existen paquetes software que permiten el cálculo y la resolución eficiente de las fórmulas anteriormente comentadas. En el trabajo presente, se ha utilizado *libsvm* en un entorno de supercomputación con algoritmos paralelizados para la obtención de los modelos de entrenamiento SVM. La salida del modelo generado por *libsvm* corresponde a un fichero de texto (Figura 4. 8) que incluye toda la información de los vectores soporte para las diferentes características de entrada [Tong y Svetnik, 2002].

```
svm_type c_svc
kernel_type linear
nr_class 2
total_sv 3
rho 1.99948
label 1 -1
nr_sv 2 1
SV
 0.2497230769230769 1:0.3 2:0.1
 0.250166153846154  1:0.3 2:0.03
-0.499889230769231  1:0.1 2:0
```

Figura 4. 8. Ejemplo de modelo SVM

Para la extracción de los diferentes coeficientes de cada una de las características junto con el término independiente se implementó un algoritmo para procesar el modelo de salida generado por el software *libsvm*. La función de decisión es cero justo en el hiperplano y por tanto:

$$0 = \sum_{i=1}^{VS} \alpha_i y_i [(x_i x) + 1] \quad (4.\ 12)$$

$$0 = \sum_{i=1}^{VS} \alpha_i y_i \left[\left(x_{i,1} x_1, \ldots, x_{i,m} x_m\right) + 1\right] \quad (4.\ 13)$$

$$0 = \sum_{i=1}^{VS} (\alpha_i y_i x_{i,1} x_1) + \cdots + \sum_{i=1}^{VS} (\alpha_i y_i x_{i,m} x_m) + \sum_{i=1}^{VS} \alpha_i y_i \quad (4.\ 14)$$

[28] Optimal separating hyperplane. Página 422. [Cherkassky y Mulier, 2007]

$$0 = \left(\sum_{i=1}^{VS} \alpha_i y_i x_{i,1}\right) x_1 + \cdots + \left(\sum_{i=1}^{VS} \alpha_i y_i x_{i,m}\right) x_m + \sum_{i=1}^{VS} \alpha_i y_i \quad (4.\ 15)$$

Donde,

$$w_j = \sum_{i=1}^{VS} \alpha_i y_i x_{i,j} \; para \; j = 1 \ldots m \qquad b = \sum_{i=1}^{VS} \alpha_i y_i \quad (4.\ 16)$$

La implementación de estas ecuaciones en un algoritmo mediante código C llamado *gethyperplane.c* se encuentra adjunto en el anexo B.1 de la presente tesis. Un requisito de *libsvm* es el trabajar con datos escalados para poder compararlos.

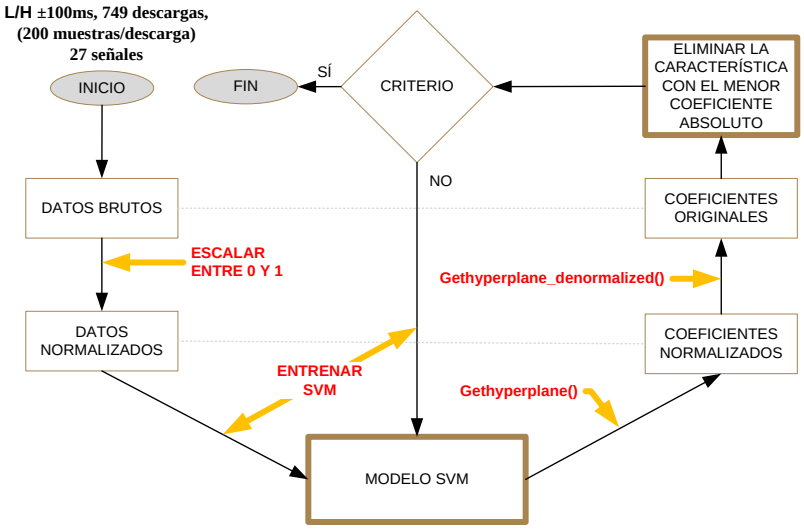


Figura 4. 9. Diagrama de flujo para la eliminación de características

Escalar los datos entre 0 y 1 significa realizar una transformación de la forma:

$$x_i \rightarrow x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (4.\ 17)$$

De esta forma la ecuación del hiperplano quedaría:

$$v^T x' + b' = 0 \quad (4.\ 18)$$

$$\sum_{j=1}^{m} v_j x_j' + b' = 0 \quad (4.\ 19)$$

Donde v_j es el peso de cada característica x'_j ya normalizada, siendo b', el valor del término independiente normalizado. Para obtener otra vez los coeficientes originales procederíamos de la siguiente manera.

$$v_1 \frac{x_1 - x_{1_{min}}}{x_{1_{max}} - x_{1_{min}}} + v_2 \frac{x_2 - x_{2_{min}}}{x_{2_{max}} - x_{2_{min}}} + \cdots + v_m \frac{x_m - x_{m_{min}}}{x_{m_{max}} - x_{m_{min}}} + b' = 0 \quad (4.\ 20)$$

Si tenemos en cuenta que,

$$w_j = v_j \frac{1}{x_{j_{max}} - x_{j_{min}}} \quad \text{y} \quad b = b' - \sum_{j=1}^{m} w_j x_{j_{min}} \tag{4. 21}$$

Entonces despejando, obtenemos los coeficientes originales:

$$w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b = 0 \tag{4. 22}$$

Otro algoritmo denominado *gethyperplane_denormalized.c* (anexo B.2) se encarga de obtener los coeficientes originales a partir de los generados por el algoritmo del Anexo B.1., *gethyperplane.c*. Todo el proceso de normalización y de selección llevado a cabo se explica mediante un diagrama de flujo en la Figura 4. 9.

Los resultados finales del trabajo se muestran en la figura siguiente. Se pueden apreciar las 27 señales iniciales y el conjunto final de las señales más relevantes con los coeficientes resultantes más elevados. Se adjunta también un gráfico que muestra la evolución del número de vectores soporte para cada modelo generado en cada una de las iteraciones. La complejidad del modelo, reflejado por el elevado número de vectores soporte, aumenta al disminuir el número de características.

TABLE I. Initial set of signals for the L/H transition.

F1. Beta normalized with respect to the diamagnetic energy
F2. R coordinate inner lower strike point
F3. R coordinate outer lower strike point
F4. Electron temperature at PSI=0.95 surface
F5. Toroidal magnetic field at PSI=0.95 surface
F6. Outer interferometry channel
F7. Upper triangularity
F8. Z coordinate lower X point
F9. Toroidal magnetic field
F10. Radial inner gap
F11. D_α view
F12. Top outer gap
F13. Line integrated density
F14. Plasma inductance
F15. Time derivative of diamagnetic energy
F16. Z coordinate inner lower strike point
F17. Z coordinate outer lower strike point
F18. Electron temperature gradient at PSI=0.95 surface
F19. Magnetohydrodynamic energy
F20. R coordinate lower X point
F21. Total heating power
F22. Safety factor at PSI=0.95 surface
F23. Plasma current
F24. Radial outer gap
F25. Lower triangularity
F26. Radiated power
F27. Elongation boundary

TABLE II. Coefficients of the hyperplane of the final set of signals.

Feature	Original data	Normalized data
F5. Toroidal magnetic field at PSI=0.95 surface	$-6.925\,446 \times 10^{1}$	−93.652 816
F9. Toroidal magnetic field	$5.457\,961 \times 10^{1}$	95.530 676
F11. D_α view	-2.7359×10^{-16}	−23.191 389
F13. Line integrated density	$4.423\,290 \times 10^{-2}$	10.037 411
F14. Plasma inductance	−8.888 935	−12.933 389
F19. Magnetohydrodynamic energy	$6.333\,208 \times 10^{-6}$	27.013 612
F23. Plasma current	$2.850\,196 \times 10^{-6}$	8.884 307
F26. Radiated power	$-2.366\,97 \times 10^{-6}$	−109.091 167
Bias	$1.163\,064 \times 10^{1}$	−3.073 230

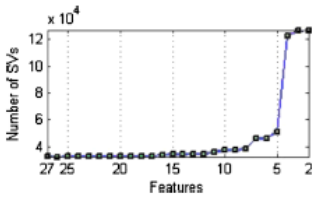


Figura 4. 10. Conjunto de señales iniciales y coeficientes del hiperplano generados

Paralelamente a este trabajo y con la finalidad de obtener, no solo los valores de predicción SVM, sino también las distancias normalizadas y reales al hiperplano de separación, se modificaron las rutinas paralelizadas SVM, tanto las del algoritmo "*train*", como las del algoritmo "*predict*", para poder incluir dicha información (Figura 4. 11).

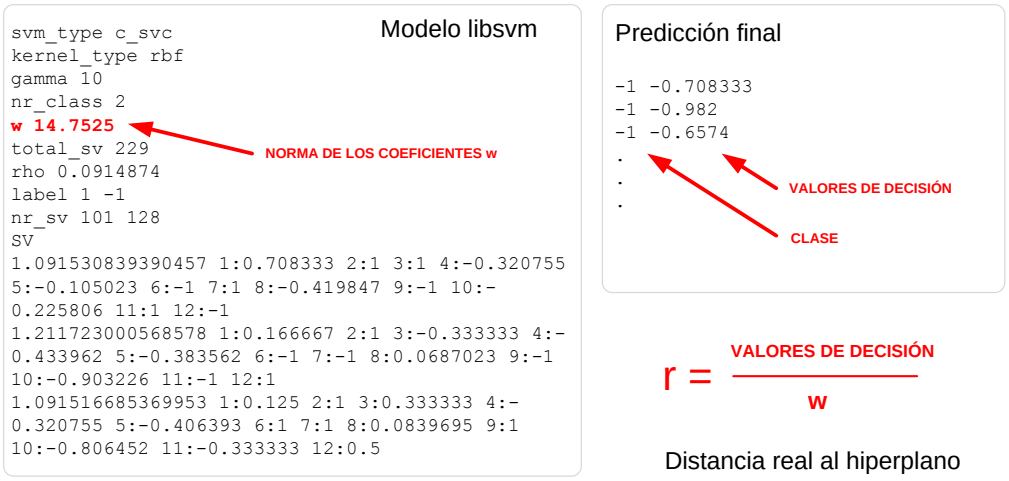


Figura 4. 11. Adicción de información a los ficheros de salida de libsvm

De esta forma, el modelo resultante incluye una línea adicional con el valor de la norma de *w*. Para clasificar nuevas instancias, el algoritmo "*predict*" añade una columna adicional en la salida, aportando información sobre los valores de decisión o distancias normalizadas al hiperplano. Con esta información y la norma de *w*, se puede obtener los valores de las distancias reales al hiperplano de separación.

4.2.2 Predictores probabilísticos

Por lo general, en la mayoría de los algoritmos de aprendizaje, las predicciones no incluyen una estimación de cómo de exactas o verídicas son éstas. Sus resultados suelen ser categóricos, sin incluir ninguna información más, respecto a su exactitud. Es evidente que las predicciones correspondientes a diferentes vectores de características pueden tener diferentes niveles de probabilidad o confianza. Un clasificador probabilístico expresa la probabilidad de que un objeto a clasificar, representado por un vector de características concreto, pertenezca a una clase determinada. De forma matemática y usando el **teorema de Bayes** [Theodoridis et al., 2003] se puede expresar de la siguiente manera:

$$P(c_i|\vec{x}) = \frac{P(c_i)P(\vec{x}|c_i)}{P(\vec{x})} = \frac{P(c_i)P(\vec{x}|c_i)}{\sum_{i=1}^{n.clases} P(c_i)P(\vec{x}|c_i)} = \frac{Priori \cdot Verosimilitud}{Evidencia} \quad (4.\ 23)$$

Esta probabilidad tiene que darse bajo la estricta condición de **aleatoriedad estadística**. Todos los vectores de características deben ser independientes e idénticamente distribuidos (iid). Esto quiere decir que dichos vectores deben ocurrir con la misma distribución de probabilidad y tienen que ser mutuamente independientes, esto es, las probabilidades de que sucedan no pueden estar relacionadas. No obstante, la necesidad de un conocimiento a priori, es una dificultad. Si no se tiene este conocimiento, las probabilidades a priori han de ser estimadas, al igual que su predicción. Algo parecido le ocurre a la **regresión logística** [Hosmer et al., 2000], que es un ajuste paramétrico para estimar también probabilidad condicionada, de la forma:

$$P(\vec{x}) = \frac{1}{1+\exp(-qf)} \quad (4.\ 24)$$

En la ecuación anterior, f es la salida del clasificador y q es un parámetro que debe ser calculado empíricamente, con lo cual, diferentes valores de q condicionarán, diferentes probabilidades para el vector de características. Tanto el teorema de Bayes como la regresión logística tienen el inconveniente de que necesitan calcular empíricamente ciertos coeficientes para poder obtener estimaciones de probabilidad. Los **predictores Venn** [Vovk et al., 2005], en cambio, obtienen estimaciones de probabilidad de una forma más directa. No tienen que realizar cálculos intermedios para encontrar los valores de otros coeficientes secundarios y el único requisito que deben cumplir es que sus vectores de características cumplan con la norma iid.

Básicamente, el procedimiento de aprendizaje mediante predictores Venn se puede ver como el proceso de encontrar la hipótesis más probable de entre todas las posibles. Ante una nueva observación a clasificar, ésta se realiza como función de la predicción de múltiples hipótesis, ponderadas por las probabilidades de que esa determinada observación pertenezca a cada una de las clases posibles.

Sean $(x_i,y_i)_{,\ i=1,\ldots,\ n-1}$, las observaciones de entrenamiento conocidas, donde x_i son los vectores de características, $y_i \in \{Y_1, Y_2, \ldots, Y_C\}$ las etiquetas clasificadas a las que pertenecen esos vectores. La observación (x_n,y_n) es el objeto a clasificar, con etiqueta desconocida para $y_n \in \{Y_1, Y_2, \ldots, Y_C\}$. La operativa mediante predictores Venn, asigna cada una de las posibles clasificaciones $Y_{j,\ j=1,\ldots,C}$ al vector x_n y divide todos los ejemplos disponibles $\{(x_1,y_1)_{,\ \ldots,}\ (x_{n-1},y_{n-1}), (x_n,Y_j)_,\}$ en un número de categorías basadas en una taxonomía $\tau_{i,\ i=1,\ldots,T}$. Una taxonomía es una función A que clasifica en T categorías la relación entre una muestra (x_k,y_k) perteneciente a dicho conjunto y el resto de observaciones.

$$\tau_i = A\left(\{(x_{1,}\, y_1), \ldots, (x_{n,}\, y_n)\}, (x_{k,}\, y_k)\right) \quad (4.\ 25)$$

Se han utilizado muchas taxonomías con los predictores Venn, como pueden ser las redes neuronales [Papadopoulos, 2013], regresión logística [Nouretdinov et al., 2012], etc. En esta tesis se ha optado por utilizar la taxonomía del centroide más próximo, CMP [Dashevskiy et al., 2008]. La expresión matemática de dicha taxonomía viene dada por:

$$\tau_i = A\left(\{(x_{1,}\, y_1), \ldots, (x_{n,}\, y_n)\}, (x_{k,}\, y_k)\right) = Y_j \quad (4.\ 26)$$

$$j = arg\ min_{j=1,\ldots,c} \|x_i - C_j\| \quad (4.\ 27)$$

C_j son los centroides de las c clases y $\|.\|$ es una distancia métrica que corresponde a la distancia Euclídea.

El procedimiento a seguir consiste en escoger una taxonomía con T categorías, para poder clasificar un vector x_n, con etiqueta desconocida y_n. Se supone primero que $y_n = Y_1$, seguidamente que, $y_n = Y_2$, y así sucesivamente hasta $y_n = Y_c$. La distribución de probabilidad empírica de las clasificaciones, en la categoría τ, que contiene (x_n,Y_1) es:

$$p^{Y_1}(Y_k) = \frac{|\{(x^*,y^*)\in\tau : y^* = Y_k\}|}{|\tau|}, k = 1, \ldots, c \quad (4.\ 28)$$

Procediendo para *c* componentes, obtenemos un vector fila de la forma:

$$\left(p^{Y_1}(Y_1), \dots, p^{Y_1}(Y_c)\right) \tag{4. 29}$$

Igualmente, la distribución de probabilidad que contiene (x_n, Y_2) es:

$$p^{Y_2}(Y_k) = \frac{|\{(x^*, y^*) \in \tau : y^* = Y_k\}|}{|\tau|}, k = 1, \dots, c \tag{4. 30}$$

Sucesivamente, se obtiene otro vector fila con *c* componentes:

$$\left(p^{Y_2}(Y_1), \dots, p^{Y_2}(Y_c)\right) \tag{4. 31}$$

Después de asignar todas las posibles clasificaciones a x_n, se genera un conjunto de distribuciones de probabilidad P_c que responde a una matriz cuadrada de dimensión *c*.

$$P_c = \begin{pmatrix} p^{Y_1}(Y_1) & p^{Y_1}(Y_2) & p^{Y_1}(Y_c) \\ p^{Y_2}(Y_1) & p^{Y_2}(Y_2) & p^{Y_2}(Y_c) \\ p^{Y_c}(Y_1) & p^{Y_c}(Y_2) & p^{Y_c}(Y_c) \end{pmatrix} \tag{4. 32}$$

El último paso del proceso predictivo es asignar una etiqueta a la muestra x_n que deseamos clasificar, siendo $y_n = Y_{cmejor}$, donde

$$c_{mejor} = arg\ max_{k=1,\dots,c}\ \overline{p(k)} \tag{4. 33}$$

Donde $\overline{p(k)}$ es la media de las probabilidades obtenidas por la etiqueta Y_k entre todas las distribuciones de probabilidad. Dicho de otra forma, la predicción será la etiqueta de la columna con el valor máximo de todos los valores medios para cada columna. El intervalo de probabilidad para esta predicción es $[L(Y_k), U(Y_k)]$, que corresponde con la probabilidad mínima y máxima de la columna donde reside el valor medio máximo.

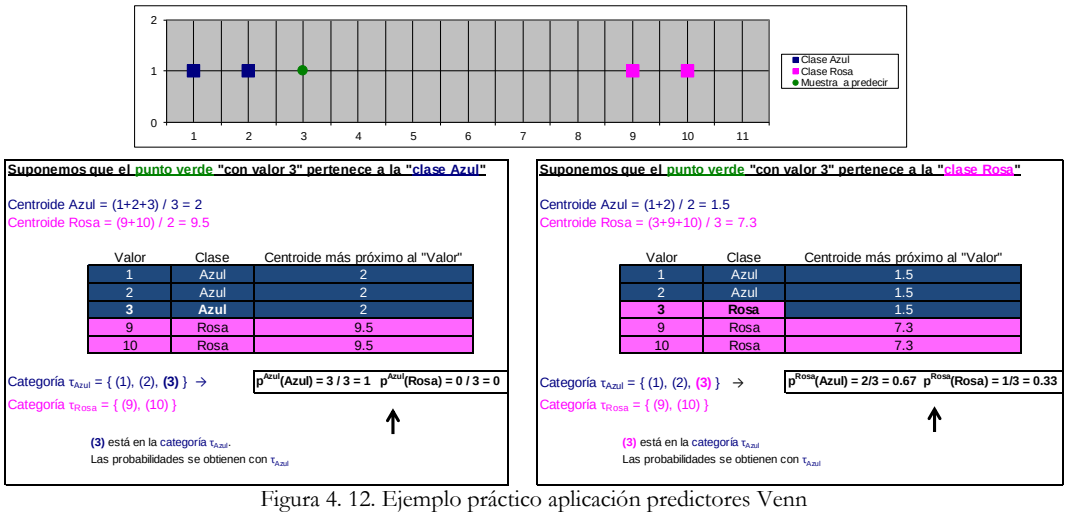


Figura 4. 12. Ejemplo práctico aplicación predictores Venn

Para el ejemplo de la figura de arriba, la matriz cuadrada de dimensión $c = 2$ que recoge todas las distribuciones de probabilidad es:

$$P_2 = \begin{pmatrix} 1 & 0 \\ 0.67 & 0.33 \end{pmatrix} \tag{4. 34}$$

Las medias de las probabilidades obtenidas por cada clase son:

$$\left.\begin{array}{l} \overline{p(Azul)} = \frac{1+0.67}{2} = 0.84 \\ \overline{p(Rosa)} = \frac{0.33}{2} = 0.16 \end{array}\right\} \rightarrow y_n = \{Azul\}, \text{ con una probabilidad entre, } [0.67, 1] \tag{4. 35}$$

Finalmente podemos decir que el punto de color verde a clasificar, corresponde a la clase Azul, con una probabilidad de valor 0.84 y una barra de error que abarca el intervalo [0.67, 1].

En los trabajos [Pereira et al., 2014] y [Vega et al., 2014] se aplicaron los predictores Venn utilizando la taxonomía CMP. En sendos trabajos, existen dos clases bien definidas para predecir las disrupciones, $Y = \{Y_{Disruptiva}, Y_{No\text{-}disruptiva}\}$, que coinciden con el número de categorías a utilizar en CMP. La elección de esta taxonomía permite condensar el espacio muestral de entrada y reducir significativamente toda la información disruptiva y no disruptiva. Las predicciones son realizadas secuencialmente, empezando solamente con una observación de tipo disruptivo y otra de tipo no disruptivo, el clasificador tienen que ir aprendiendo sin apenas información relevante desde sus inicios. El predictor Venn es capaz de aprender directamente por transducción [Vapnik, 2000], en vez de tener que continuamente estar generando modelos de aprendizaje intermedios de forma inductiva, muy costosos computacionalmente, ver Figura 4. 13. Este paradigma de predicción por transducción es comúnmente utilizado por los predictores conformales [Gammerman et al., 1998], que trabajan directamente desde los datos de aprendizaje y por tanto, muy útil en tareas de predicción en tiempo real.

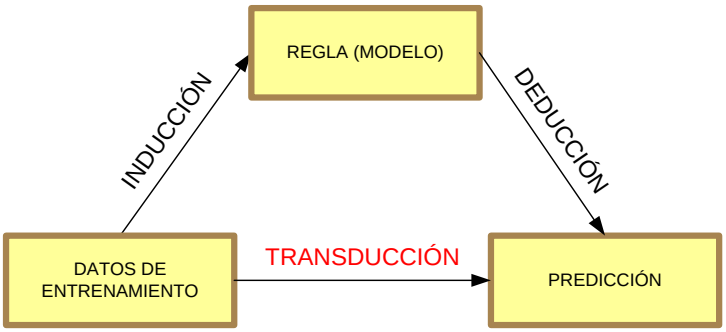


Figura 4. 13. Tipos de aprendizaje

Efectivamente, los clasificadores Venn pertenecen a la familia de los predictores conformales, cuya otra característica significativa es la de proporcionar información cualitativa de cómo de relevante es la información aportada en la predicción, bien sea mediante la estimación de una barra de error como hemos visto anteriormente, o bien

mediante intervalos de confianza y credibilidad [Saunders et al., 1999], conceptos éstos que serán explicados en el siguiente apartado.

4.2.3 Reconocimiento de imágenes mediante el algoritmo conformal del vecino más próximo.

Las técnicas predictivas conformales, complementan la predicción de métodos que se engloban dentro del aprendizaje automático (SVM, Vecino más Próximo, etc.), mediante la aportación de medidas cuantitativas de confianza y credibilidad, ofreciendo niveles comparativos de precisión y fiabilidad [Gammerman et al., 2007]. La propiedad más importante de estos conceptos es su validez automática. Establecen límites para no sobreestimar predicciones con elevada confianza ni descartar las más inciertas, bajo la hipótesis de aleatoriedad estadística, donde las muestras y sus clases se generan mutuamente independientes, de forma que, las probabilidades de que aparezcan dichas muestras no estén relacionadas unas con otras, pero todas ellas deben ajustarse a la misma distribución de probabilidad.

Predicciones multiclase con un alto nivel de fiabilidad y significancia se aplicaron también en [Vega et al., 2010] en modo diferido al reconocimiento de patrones en imágenes, pertenecientes al diagnóstico Thomson del estellerator TJ-II y utilizando el algoritmo del vecino más próximo de forma inductiva y conformal. Igualmente, la implementación que se hizo del algoritmo conformal del vecino más próximo fue obtenida del trabajo [Shafer y Vovk, 2008] e implementado también en Matlab.

Se define previamente una **medida de no-conformidad** como una función que asigna a cada secuencia de datos del conjunto de entrenamiento, una secuencia de números ($\alpha_1,\ldots, \alpha_n$) llamados puntuaciones o **resultados de no-conformidad**. Se define también el término **p-valor** como la proporción de los α's que son iguales o más grandes que el último α calculado.

$$p = \frac{\#\{i=1,\ldots,l+1 \mid \alpha_i \geq \alpha_{l+1}\}}{l+1} \tag{4. 36}$$

El algoritmo conformal, a partir del conjunto de entrenamiento de datos, de la nueva muestra (x_l+1), y de cada nivel de confianza, decide si incluir la clase (y_l+1) en el conjunto de predicción.

$$(x_1, y_1, \ldots, x_l, y_l, x_{l+1}, \ldots) \tag{4. 37}$$

La metodología a seguir por el algoritmo es la siguiente. Provisionalmente asignamos cada muestra x_l+1 a una clase diferente '*y*' de entre las posibles. Para cada $i=1,\ldots,l$, calculamos los resultados de no conformidad α_i. Calculamos el p-valor para esa clase 'y'. Incluimos '*y*' en el conjunto de predicción si y solo si el p-valor es mayor que ε, siendo ésta el **nivel de significancia** o error en la predicción o probabilidad de que la predicción sea engañosa y equivocada. Esto implica que para cada l la probabilidad de que la clase eventual y_l+1 pertenezca al conjunto de predicción anterior es al menos 1- ε. Se asigna el valor de la **credibilidad** al mayor p-valor calculado. Se asigna el valor de la **confianza** como 1 menos el segundo mayor p-valor.

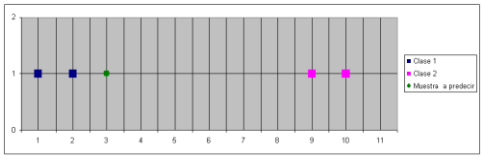
En el aprendizaje anterior, podemos actuar de dos formas diferentes. Predicción desde solamente viejos ejemplos, esto es, se observan las muestras anteriores, un conjunto de entrenamiento fijo, y predecimos su clase a la que pertenece. O bien, predicción usando las características de nuevos ejemplos, donde cada predicción y su muestra son incluidas en el conjunto de entrenamiento. Así el conjunto de entrenamiento aumenta con cada

nueva muestra. La primera opción realmente no tiene interés a no ser que el conjunto inicial de entrenamiento sea realmente grande. Si escogemos la segunda opción, la calidad de nuestras predicciones mejorará a medida que nosotros acumulemos más muestras, mejorando el proceso de aprendizaje con la experiencia.

Lo que realmente diferencia un algoritmo conformal de otro es la elección de la medida de no conformidad a utilizar. En el trabajo [Vega et al., 2010] y con el objetivo de clasificar las imágenes generadas por el diagnóstico de esparcimiento Thomson del TJ-II mediante el algoritmo conformal del vecino más próximo, la elección de la medida de no conformidad fue la siguiente:

$$\alpha_i = \frac{\min\{|x_j - x_i| : 1 \le j \le n \ \ \& \ \ j \ne i \ \ \& \ \ y_j = y_i\}}{\min\{|x_j - x_i| : 1 \le j \le n \ \ \& \ \ j \ne i \ \ \& \ \ y_j \ne y_i\}} = \frac{distancia\ al\ vecino\ más\ prox.\ con\ la\ misma\ clase}{distancia\ al\ vecino\ más\ prox.\ con\ diferente\ clase} \qquad (4.\ 38)$$

Siguiendo con el mismo ejemplo ilustrativo que el expuesto en la Figura 4. 12, pero en esta ocasión, y para abreviar de forma binaria, la clase Azul le haremos corresponder con la clase numérica 1 y la clase Rosa se corresponderá con la clase tipo 2. En el ejemplo de la Figura 4. 14 podemos decir que, con un nivel máximo del 80% de confianza, predecimos que la etiqueta y5 que le corresponde a x5, es igual a 1, con credibilidad 40%.

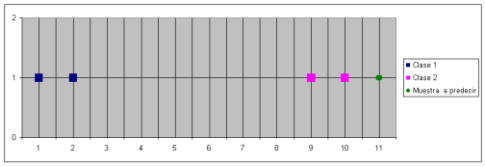


	valor	clase y	alpha y=1	alpha y=2
x1	1	1	0,125	0,500
x2	2	1	0,143	1,000
x3	9	2	0,167	0,143
x4	10	2	0,143	0,125
x5	3	1	0,167	
x5	3	2		6,000
p1			0,400	
p2				0,200

y_5 =1
confianza = 80%
credibilidad = 40%

Figura 4. 14. Ejemplo del vecino más próximo

En el siguiente ejemplo, con un 80% de confianza, y una credibilidad del 100% predecimos que y5 = 2.

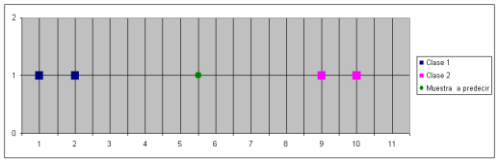


	valor	clase y	alpha y=1	alpha y=2
x1	1	1	0,125	0,125
x2	2	1	0,143	0,143
x3	9	2	0,500	0,143
x4	10	2	1,000	0,125
x5	11	1	9,000	
x5	11	2		0,111
p1			0,200	
p2				1,000

y_5 =2
confianza = 80%
credibilidad = 100%

Figura 4. 15. Clasificación de otro punto con el vecino más próximo

Existen otras ocasiones en las que no podemos decir nada acerca de la clasificación, por ejemplo:



	valor	clase y	alpha y=1	alpha y=2
x1	1	1	0,125	0,222
x2	2	1	0,142	0,285
x3	9	2	0,285	0,142
x4	10	2	0,222	0,125
x5	5.5	1	1,000	
x5	5.5	2		1,000
p1			0,200	
p2				0,200

y_5 = ¿?
confianza = ¿?
credibilidad = ¿?

Figura 4. 16. Punto no clasificable

Con el mismo nivel del 80% de confianza, no podemos decir que y5 = 1 y al mismo tiempo que y5 = 2. No podemos incluir en una región de predicción válida dicha predicción. En estas situaciones lo más lógico sería informar de la constatación de una salida con valor 'indeterminado', donde el clasificador es incapaz de discernir una clase concluyente al respecto.

En una predicción conformal, alta confianza nos indica que todas las alternativas a la clase predicha son improbables, por tanto la confianza se puede expresar como la probabilidad de que la predicción sea verídica. Baja credibilidad significa que el elemento a clasificar no es representativo del conjunto de entrenamiento, por tanto, la credibilidad expresa como de representativo es el objeto de prueba, el elemento a clasificar, respecto del conjunto de entrenamiento. Los p-valores definen la proporción de elementos que son iguales o más raros que el elemento que se quiere predecir y los α's reflejan la rareza, extrañeza o la no conformidad de los ejemplos respecto a una clase determinada. Un valor alto significa que es muy raro o extraño que una muestra pertenezca a la clase supuesta.

Con toda esta información se aplica el algoritmo del vecino más próximo conformal a un total de 165 imágenes (576 x 385 pixels). Se comienza inicialmente con una sola imagen de cada clase (fondo: 17, parásita: 18, ECH: 56, NBI: 32 y corte: 42). Para cada imagen que se predice, se añaden las características que definen dicha imagen al conjunto de entrenamiento anterior, de forma que este conjunto va creciendo con cada nueva iteración que se produce. Los resultados finales obtenidos fueron los siguientes:

Nivel wavelet-haar	**Resultados**
3 (72x48)	Aciertos: 155 (96.87%) Fallos: 3 (1.87%) Sin clasificar: 2 (1.25%)
4 (36x24)	Aciertos: 155 (96.87%) Fallos: 3 (1.87%) Sin clasificar: 2 (1.25%)
5 (18x12)	Aciertos: 151 (94.37%) Fallos: 6 (3.75%) Sin clasificar: 3 (1.87%)

Tabla 4. 1. Tasas de acierto con el algoritmo conformal del vecino más próximo

Como vectores de características de entrada se escogió el valor resultante de la suma de los coeficientes de aproximación del nivel correspondiente (se probaron niveles 3, 4 y 5) con los coeficientes de detalle verticales del mismo nivel. Se pudo observar también como el sistema puede aprender con cada nueva imagen añadida al conjunto de entrenamiento, la confianza aumenta a medida que el sistema aprende.

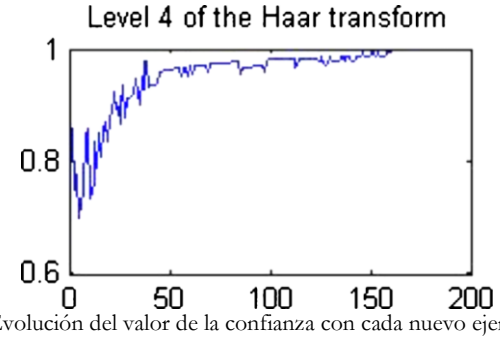


Figura 4. 17. Evolución del valor de la confianza con cada nuevo ejemplo clasificado

4.3 Técnicas de regresión

El análisis de regresión es una técnica estadística para estudiar la relación entre variables. Tanto en el caso de dos variables, conocido como regresión simple, como en el caso de más de dos variables, denominado regresión múltiple [Izenman, 2008], el análisis puede utilizarse para explorar y cuantificar la relación entre una variable llamada dependiente o criterio (y) y una o más variables llamadas independientes, predictoras o regresoras (x_1, x_2, …, x_m), así como para desarrollar una ecuación lineal con fines predictivos:

$$Ax = y \rightarrow y = w^T x + b,\ A \in R^{n\,x\,m}, y \in R^n, w \in R^m, b \in R, n \geq m \tag{4. 39}$$

Por otro lado, el reconocimiento de patrones, como se ha explicado en capítulos anteriores, puede emplearse para la detección rápida de fenomenología muy específica. Sin embargo, además de la necesaria utilidad de herramientas que faciliten los procesos de búsqueda y reconocimiento, la interpretación física de la información surge como un tema fundamental. Una alternativa a las técnicas de reconocimiento de patrones son los modelos orientados a los datos. En [González et al., 2012] se presenta la combinación de un método automático en la localización de transiciones L/H junto con diferentes estudios físicos acerca de la naturaleza y la implicación en la transición L/H de muchos otros parámetros físicos, todo ello utilizando técnicas de regresión estadística que exploran los datos para extraer leyes de escala y aportar predicciones con intervalos de confianza en sus coeficientes. Básicamente, la aportación del autor consistió en la estimación paramétrica del umbral de potencia en el instante de la transición L/H, haciendo uso de diferentes variables explicativas, aplicadas a diferentes métodos de regresión y comparando los puntos del instante real de la transición con el obtenido automáticamente por el clasificador.

4.3.1 Mínimos cuadrados ordinarios frente a regresión Ridge

En la aproximación clásica de mínimos cuadrados, (OLS, siglas del inglés), la matriz *A* de la ecuación (4.39) se asume que está libre de errores y/o correlaciones. Sin embargo

en aplicaciones de ingeniería y en física experimental, este supuesto es irreal. Debido a que los datos son obtenidos por sensores e instrumentación electrónica de medida, tanto la matriz *A* como el vector *y*, suele estar contaminado por ruido y por multi-colinealidad o sea, algún tipo de relación entre ellas, debido al acoplamiento de interferencias [Pearson, 2005]. De forma que los sistemas lineales, en muchas situaciones prácticas, se encuentran mal acondicionados. Este efecto puede ocasionar predicciones equivocadas, soluciones sin significado físico y elevadas incertidumbres en los coeficientes *w* de las variables independientes. Para mitigar el problema anterior se propuso la metodología denominada regresión Ridge (RR) [Saunders et al., 1998]. Este método consiste en agregar un parámetro sesgado a los estimadores de mínimos cuadrados ordinarios con la finalidad de reducir el error estándar de éstos que se comete a la hora de predecir el valor de la variable dependiente.

Las leyes de escala, en la mayoría de las ramas de las ciencias, suelen ser expresadas en términos de productos de monomios con diferentes exponentes. Para conseguir esto a partir de la ecuación (4.39), procederemos de la siguiente manera:

$$\sum_{i=1}^{m} w_i x_i + b = 0 \tag{4.40}$$

$$\sum_{i=1}^{m} w_i \log_{10} x_i + b = 0 \tag{4.41}$$

$$\sum_{i=1}^{m} \log_{10} x_i^{w_i} + b = 0 \tag{4.42}$$

$$\log_{10} \prod_{i=1}^{m} x_i^{w_i} = -b \tag{4.43}$$

$$\prod_{i=1}^{m} x_i^{w_i} = 10^{-b} \tag{4.44}$$

$$x_1^{w_1} x_2^{w_2} \dots x_j^{w_j} \dots x_m^{w_m} = 10^{-b} \tag{4.45}$$

$$x_j^{w_j} = 10^{-b} x_1^{-w_1} x_2^{-w_2} \dots x_m^{-w_m} \tag{4.46}$$

$$\left[x_j^{w_j}\right]^{\frac{1}{w_j}} = \left[10^{-b} x_1^{-w_1} x_2^{-w_2} \dots x_m^{-w_m}\right]^{\frac{1}{w_j}} \tag{4.47}$$

$$x_j = 10^{\frac{-b}{w_j}} x_1^{\frac{-w_1}{w_j}} x_2^{\frac{-w_2}{w_j}} \dots x_m^{\frac{-w_m}{w_j}} \tag{4.48}$$

$$x_j = 10^{k_0} x_1^{k_1} x_2^{k_2} \dots x_m^{k_m},\ para\ k_0 = -\frac{b}{w_j}, k_1 = -\frac{w_1}{w_j}, \dots, k_m = -\frac{w_m}{w_j} \tag{4.49}$$

Cualquier variable independiente de la ecuación lineal (4.39), la podemos despejar en función de las demás variables y en forma de productos de monomios aplicando los logaritmos correspondientes a los datos. La ventaja de utilizar ecuaciones de este tipo es que se puede observar el peso que tienen los coeficientes *k* sobre cada característica en una escala normalizada mediante logaritmos y muy útil para extrapolar resultados, el inconveniente principal es que el ajuste resultante es lineal pero en esa escala logarítmica, dejando de ser lineal cuando se trabaja con datos brutos reales. Por este motivo se hace también relevante dar información de las incertidumbres que existen en dichos coeficientes mediante los intervalos de confianza ($k \pm \Delta k$). Los límites de confianza para los coeficientes[29] ajustados por la regresión vienen dados por:

$$C = k \pm t\sqrt{S} \tag{4.50}$$

[29] Confidence and Prediction Bounds. Matlab Help.

Donde *t* es la inversa de la función de distribución de probabilidad *Student-t* dependiente del nivel de confianza de entrada y con un valor cercano a 1.96 para un nivel de confianza del 95%. En probabilidad y estadística, la distribución t de *Student* es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño. El error estándar de estimación *S* se calcula con la finalidad de medir la confiabilidad de la ecuación de predicción, siendo su valor, los elementos de la diagonal desde la matriz de covarianza. El término s^2 se conoce como el error cuadrático medio o MSE.

$$S = diag[(X^T X)^{-1} s^2] \tag{4. 51}$$

Para realizar un ajuste en términos de mínimos cuadrados ordinarios se trata de encontrar la función de aproximación lineal en la cual, la suma de los cuadrados de las diferencias entre los valores observados y esperados sea menor. Expresando esto en forma matricial, los coeficientes quedarían de la siguiente manera:

$$w = (X^T X)^{-1} X^T y \tag{4. 52}$$

Posteriormente, podemos predecir nuevas instancias de prueba, de la forma:

$$y' = X_{test} w \tag{4. 53}$$

De la expresión de los coeficientes se puede observar que si el determinante $X^T X$ es casi o muy cercano a cero, su inversa sería casi infinita. Esto ocurre cuando existen variables muy correlacionadas unas con otras. Los algoritmos de inversión de matrices pierden entonces precisión, al tener que dividir por un número muy pequeño. Esto originaría problemas de inestabilidad en los coeficientes, signos incorrectos en los mismos y elevados errores estándar. Para minimizar este problema, es necesario contraer los coeficientes *w* de OLS, logrando coeficientes ajustados con menor varianza, dando estabilidad así a la predicción del modelo. La matriz $X^T X$ es reemplazada por otra matriz numéricamente más estable debido a la agregación (suma) de un sesgo *s* con la finalidad de reducir el error estándar de éstos. Esta técnica se denomina regresión Ridge, y sus coeficientes se expresan de la siguiente manera:

$$w = (X^T X + sI)^{-1} X^T y \tag{4. 54}$$

De esta forma, la matriz $X^T X + sI$ es siempre invertible si el término *s* es mayor que cero, siendo *I* la matriz identidad. Al ser *s* un parámetro que introduce un sesgo en los estimadores, es deseable seleccionar el valor más pequeño de *s* por el cual se estabilizan los coeficientes de regresión *w*. Un valor muy alto sobre-regularizaría el ajuste y un valor igual a cero correspondería a una expresión OLS. La forma más práctica de conseguirlo es escoger un valor de *s* pequeño por el cual se consigue el menor error cuadrático medio en el ajuste de regresión. La estandarización y el escalado también son muy útiles para poder comparar los coeficientes *w* entre sí y en la misma escala, y por tanto poder comprobar donde se estabilizan éstos. El error estándar de estimación *S* (4.51) para un

ajuste mediante RR[30] y con la finalidad de medir la confiabilidad de la ecuación de predicción es:

$$S = diag[(X^T X + sI)^{-1} s^2] \quad (4.\ 55)$$

[30] Radial Basis Function Statistics. Pag 9-25. Matlab Help MBC toolbox

4.3.2 Leyes de escala para determinar el umbral de potencia en transiciones L/H

Las transiciones desde un modo de confinamiento estándar L a un modo de confinamiento mejorado H son observadas en la mayoría de los dispositivos tokamak. Se sabe que esta transición depende de muchos parámetros, no obstante es ampliamente aceptado que el umbral de potencia para que se produzca la transición L/H depende sobre todo de la densidad del plasma, del campo magnético toroidal y del tamaño del plasma [Martin et al., 2008][McDonald et al., 2006]. Con el objetivo de poder estimar paramétricamente el umbral de potencia de la transición L/H del Tokamak JET, en el trabajo [González et al., 2012] se publicaron resultados en este sentido mediante predicciones acotadas basadas en OLS y RR. Se trató de investigar si los tiempos determinados automáticamente por el predictor pueden ser utilizados directamente mediante sentido físico y sin intervención humana de forma automática. De esta forma se compararon las leyes de escala del umbral de potencia obtenida usando los tiempos de la transición determinados manualmente por los expertos y los obtenidos automáticamente por el predictor mediante la combinación de dos clasificadores SVM. En la siguiente tabla se puede observar dicha comparación para un total de 538 descargas del JET pertenecientes a las campañas C21 a C26.

	Regression method	Transition time determination	Scaling law	Mean square error
1	OLS	Experts	$P_{\text{Thresh}} = 10^{0.80\pm0.40} \cdot n_e^{0.56\pm0.12} \cdot B_T^{0.58\pm0.20} \cdot S^{0.39\pm0.60}$	7.082
2	OLS	Predictor	$P_{\text{Thresh}} = 10^{1.00\pm0.39} \cdot n_e^{0.64\pm0.08} \cdot B_T^{0.50\pm0.17} \cdot S^{0.23\pm0.60}$	8.635
3	RR	Experts	$P_{\text{Thresh}} = 10^{0.80\pm0.10} \cdot n_e^{0.55\pm0.08} \cdot B_T^{0.58\pm0.12} \cdot S^{0.39\pm0.14}$	7.077
4	RR	Predictor	$P_{\text{Thresh}} = 10^{1.00\pm0.10} \cdot n_e^{0.64\pm0.07} \cdot B_T^{0.50\pm0.11} \cdot S^{0.23\pm0.13}$	8.632

Tabla 4. 2. Resultados con la señal de la superficie del plasma del JET

Hay que recalcar que la incertidumbre de los coeficientes obtenida por el método de regresión mediante OLS es mucho mayor que la obtenida con RR para el mismo nivel de confianza del 95%, no existiendo diferencias significativas en lo que concierne a los propios coeficientes.

Resultados con el umbral de potencia utilizando el factor de seguridad q_{95} en lugar de la superficie del plasma se adjuntan a continuación.

	Regression method	Transition time determination	Scaling law	Mean square error
1	OLS	Experts	$P_{\text{Thresh}} = 10^{1.10\pm0.17} \cdot n_e^{0.55\pm0.12} \cdot B_T^{0.61\pm0.20} \cdot q_{95}^{-0.14\pm0.23}$	7.106
2	OLS	Predictor	$P_{\text{Thresh}} = 10^{1.19\pm0.13} \cdot n_e^{0.63\pm0.08} \cdot B_T^{0.52\pm0.17} \cdot q_{95}^{-0.10\pm0.22}$	8.657
3	RR ($k = 4$)	Experts	$P_{\text{Thresh}} = 10^{1.10\pm0.10} \cdot n_e^{0.54\pm0.08} \cdot B_T^{0.61\pm0.12} \cdot q_{95}^{-0.14\pm0.12}$	7.102
4	RR ($k = 4$)	Predictor	$P_{\text{Thresh}} = 10^{1.19\pm0.08} \cdot n_e^{0.63\pm0.07} \cdot B_T^{0.51\pm0.11} \cdot q_{95}^{-0.10\pm0.12}$	8.654

Tabla 4. 3. Resultados con la señal q95 del JET

Finalmente, se obtuvieron resultados separando las descargas que tienen un patrón morfológico claro en la señal D_α de aquellas otras que no lo tienen y utilizando solamente el método de RR.

	Regression method	Transition time determination	Scaling law	Mean square error
Transitions with clear signature				
1	RR ($k = 4$)	Experts	$P_{\text{Thresh}} = 10^{2.10\pm0.08} \cdot n_e^{0.92\pm0.08} \cdot B_T^{0.25\pm0.09} \cdot S^{-1.15\pm0.10}$	2.819
2	RR ($k = 4$)	Predictor	$P_{\text{Thresh}} = 10^{2.24\pm0.08} \cdot n_e^{0.98\pm0.07} \cdot B_T^{0.19\pm0.09} \cdot S^{-1.27\pm0.09}$	2.803
3	RR ($k = 4$)	Experts	$P_{\text{Thresh}} = 10^{1.39\pm0.07} \cdot n_e^{0.83\pm0.08} \cdot B_T^{0.26\pm0.09} \cdot q_{95}^{-0.18\pm0.10}$	2,890
4	RR ($k = 4$)	Predictor	$P_{\text{Thresh}} = 10^{1.43\pm0.07} \cdot n_e^{0.87\pm0.08} \cdot B_T^{0.21\pm0.09} \cdot q_{95}^{-0.14\pm0.10}$	2.866
Transitions with non-clear signature				
5	RR ($k = 4$)	Experts	$P_{\text{Thresh}} = 10^{0.81\pm0.12} \cdot n_e^{0.73\pm0.10} \cdot B_T^{0.50\pm0.13} \cdot S^{0.69\pm0.15}$	7.955
6	RR ($k = 4$)	Predictor	$P_{\text{Thresh}} = 10^{0.84\pm0.10} \cdot n_e^{0.63\pm0.08} \cdot B_T^{0.55\pm0.13} \cdot S^{0.49\pm0.14}$	9.792
7	RR ($k = 4$)	Experts	$P_{\text{Thresh}} = 10^{1.27\pm0.11} \cdot n_e^{0.71\pm0.10} \cdot B_T^{0.53\pm0.13} \cdot q_{95}^{-0.12\pm0.13}$	8.012
8	RR ($k = 4$)	Predictor	$P_{\text{Thresh}} = 10^{1.20\pm0.09} \cdot n_e^{0.62\pm0.08} \cdot B_T^{0.56\pm0.13} \cdot q_{95}^{-0.11\pm0.13}$	9.825

Tabla 4. 4. Resultado entre diferentes conjuntos de transiciones

Los resultados denotaron que, efectivamente, la base estadística de las señales es diferente en los dos grupos de datos, quedando reflejada su variabilidad en los coeficientes de cada característica.

4.3.3 Regresión conformal no paramétrica para transiciones L/H.

La usabilidad de predictores conformales es de gran importancia en problemas, no solo de clasificación, como hemos visto en apartados anteriores, sino también en problemas de regresión, sobre todo cuando se presentan dominios donde existen errores, tanto en los datos como en sus predicciones. En temas de clasificación, los predictores establecen niveles de confianza y credibilidad, aportando información añadida a dicha predicción. Para temas de regresión, se establece una barra de error estimada, que es tanto más grande, cuanto mayor es la confianza en el ajuste de la predicción. En la publicación [Vega et al., 2012], no solo se sintetizan las aplicaciones realizadas utilizando dichos predictores conformales, sino que se contribuye con una implementación conformal y no paramétrica para ajustes de regresión, este algoritmo fue implementado en Matlab siguiendo las indicaciones explicadas en la publicación [Nouretdinov et al., 2001]. En esta publicación se expone una aproximación alternativa a los modelos paramétricos de leyes de escala vistos en el capítulo anterior, conocida como regresión predictiva conformal. En dicho artículo, se facilita un pseudocódigo que implementa la técnica RR mediante funciones *kernel* para obtener barras de error en cada predicción a realizar en función de un nivel de confianza elegido a priori. De forma que, a mayor confianza, mayor es la longitud de la barra de error para dicha predicción. Las funciones *kernel* permiten obtener resultados sorprendentes en problemas no lineales utilizando solamente operaciones algebraicas sencillas, pudiéndose realizar regresión no lineal mediante la construcción de una función de regresión lineal en un espacio de características de más alta dimensión, ver Figura 4. 18.

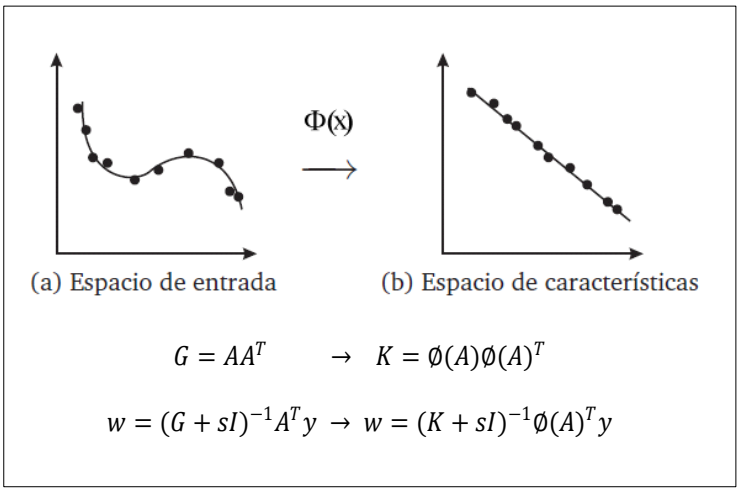


Figura 4. 18. Función kernel en regresión

Se pueden utilizar diferentes *kernels K* (lineal, polinomial, RBF, tangente hiperbólica, etc.)[Shawe-Taylor et al., 2004] junto con diferentes parámetros de regularización *s* con el objetivo de encontrar el mejor modelo explicativo en ese espacio de características y poder aplicarlo posteriormente a las aproximaciones a realizar para nuevos ejemplos de entrada. En la figura siguiente podemos observar diferentes ajustes para 100 puntos con una distribución de entrada a modo de sombrero mejicano. Se ha utilizado un *kernel*

polinómico de grado 2 y un *kernel* RBF, sigma = 2 para un nivel de confianza del 90% en ambos casos.

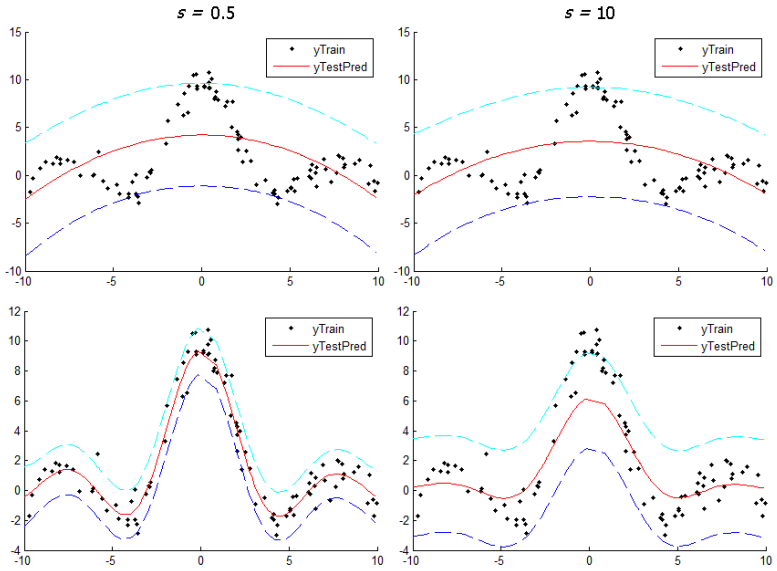


Figura 4. 19. Comparación ajuste polinomial y RBF

.

Utilizando la misma distribución del sombrero mejicano pero en tres dimensiones y para un total de 1000 observaciones, en la Figura 4. 20, observamos un ajuste mediante un *kernel* RBF (s = 0.5, sigma = 2) y un *kernel* tangente hiperbólica con parámetros (s = 15, η = 0.005), igualmente con un nivel de confianza del 90% .

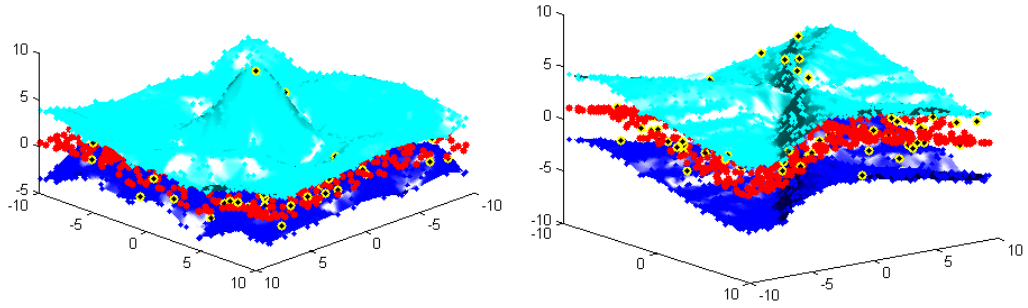


Figura 4. 20. Comparación ajuste RBF y tangente hiperbólica

En [Vega et al., 2012] se aplicaron estas técnicas para la obtención de barras de error sobre la predicción del umbral de potencia en las transiciones L/H del JET de forma no paramétrica y utilizando la misma base de datos que la utilizada en [González et al., 2012], donde se estimaron predicciones para el umbral de potencia pero de forma paramétrica.

En esta ocasión, para la estimación del modelo, se utilizaron las descargas más cercanas al centroide de las señales (Potencia, densidad, campo toroidal y superficie total del plasma) y se probaron las restantes descargas, obteniendo valores de predicción con sus correspondientes barras de error, para una confianza del 90%. Es necesario recalcar que el algoritmo conformal utilizado en este trabajo difiere del utilizado en el algoritmo del vecino más próximo conformal, en que no se va añadiendo la muestra a predecir al

conjunto de entrenamiento. El conjunto de entrenamiento es fijo y está formado por las 286 descargas más próximas al centroide de todas las señales.

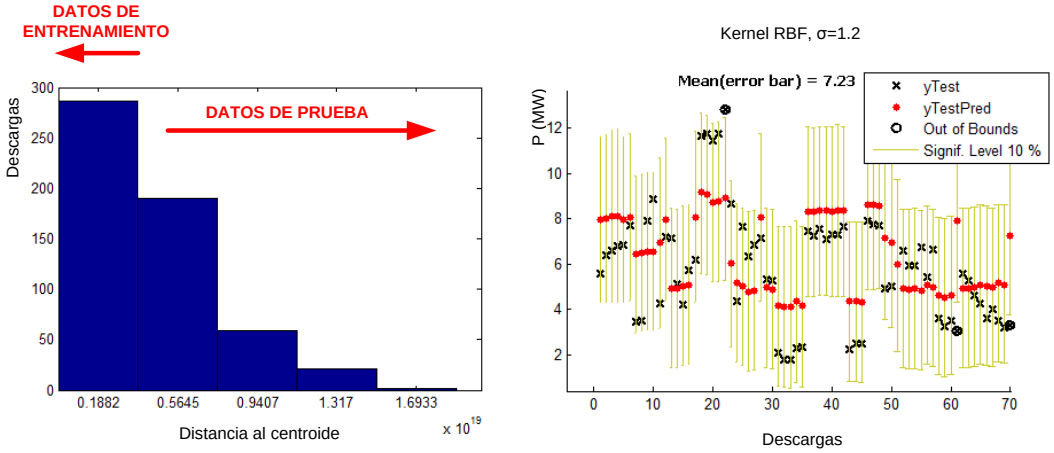


Figura 4. 21. Predicción transición L/H con regresión no paramétrica

Los resultados obtenidos se muestran en la Figura 4. 21. Se puede observar la predicción del umbral de potencia junto con su barra de error, se adjuntan también los valores observados para su comparación.

4.4 Metodologías de evaluación en el rendimiento de un predictor

La evaluación que se haga de la información que aportan las salidas de los algoritmos de aprendizaje es de extrema importancia. Tradicionalmente y como métrica más común para valorar los porcentajes de aciertos de un predictor se ha utilizado la relación, predicción de aciertos totales dividido entre número total de observaciones clasificadas. Imaginemos que un predictor nos proporciona un porcentaje de aciertos del 99.99%. A primera vista podría parecer que nuestro clasificador es muy válido y eficiente. La percepción cambia si añadimos la información de que se han evaluado 10000 observaciones de la clase A y 1 observación de la clase B y que nuestro predictor acierta todas las observaciones de la clase A pero falla a la hora de clasificar la única observación que se dispone de la clase B. Este ejemplo que puede parecer tan extremo, ocurre en muchas situaciones reales cotidianas. Los datos experimentales obtenidos por los diagnósticos de los dispositivos de fusión nuclear se adquieren en entornos muy hostiles, sometidos a elevados campos electromagnéticos que inducen interferencias y errores en los mismos, esto puede conllevar a clasificar datos erróneamente incluso por el experto o supervisor. Se trabajan con grandes conjuntos de datos de muy alta dimensionalidad, regularidades muy complejas y conocimiento a priori muy escaso. En muchas ocasiones la información disponible está fuertemente desbalanceada y solapada debido a que las clases no son separables y distinguibles. Ante este escenario, la evaluación que se haga de un predictor no puede quedarse en la métrica anteriormente expuesta y hay que ser más preciso aportando información con más detalle. Tanto en tareas de clasificación como en ajustes de regresión, es necesaria la aportación de otras métricas o evaluaciones que ponderen de diferente manera y por separado tanto los aciertos como los fallos, para tener una visión si cabe, más independiente de los mismos.

4.4.1 Métricas de rendimiento para evaluar clasificadores

Los resultados y salidas generados por un clasificador pueden organizarse por medio de una **matriz de confusión** como se ilustra en la Figura 4. 22. En dicha matriz, las columnas representan la clase actual a la que pertenecen y las filas representan la clase resultado de aplicar la predicción correspondiente.

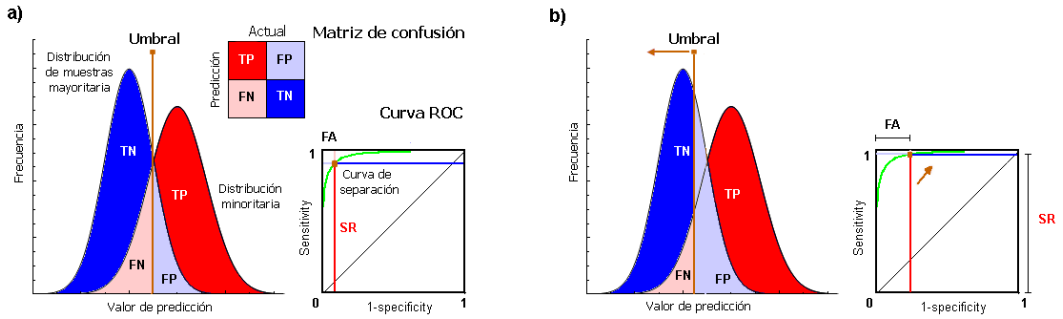


Figura 4. 22. Información útil para evaluar clasificadores

Los verdaderos negativos (TN, siglas del inglés), son el número de ejemplos u observaciones negativas correctamente clasificadas. Los falsos positivos o FP, son el número de ejemplos negativos incorrectamente clasificados como positivos. Los falsos negativos FN, son el número de ejemplos positivos incorrectamente clasificados como negativos y los verdaderos positivos TP son el número de ejemplos positivos correctamente clasificados. Tradicionalmente, la métrica más utilizada para medir los algoritmos de clasificación viene representada por un porcentaje global comúnmente conocida en su acepción inglesa como *Accuracy*.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4. 56}$$

Esta métrica refleja la suma del número de aciertos de cada clase respecto al número total de observaciones, igualmente sumando las dos clases. Por este motivo, no es apropiado ni conveniente utilizarla, cuando los ejemplos de prueba no están balanceados o los errores en cada clase varían notablemente, según se explica en [Chawla, 2005].

La curva ROC es una técnica muy conocida para reflejar gráficamente el rendimiento de un clasificador sobre un amplio rango de situaciones entre las tasas TP y FN de la distribución minoritaria. En un gráfico ROC, el eje de abscisas representa las tasas de falsas alarmas, esto es, el porcentaje de errores para la distribución mayoritaria.

$$FA = \frac{FP}{(TN+FP)} = 1 - \frac{TN}{(TN+FP)} = 1 - specificity \tag{4. 57}$$

El eje de ordenadas de un gráfico ROC refleja las tasas de acierto SR (conocida también con el nombre de *recall* o *sensitivity*, en inglés), esto es, el porcentaje de aciertos para la distribución de ejemplos minoritaria.

$$SR = \frac{TP}{(TP+FN)} \tag{4. 58}$$

El punto perfecto o ideal en una representación ROC sería (0,1), esto es, todos los ejemplos positivos están correctamente clasificados y ningún ejemplo negativo está incorrectamente clasificado como positivo, esto correspondería a decir que:

$$SR - FA = 1 \tag{4. 59}$$

Pero en la práctica, en la mayoría de los casos, esto no ocurre así. La precisión para la clase minoritaria se puede definir como todos los valores predictivos positivos:

$$precision = PPV = \frac{TP}{(TP+FP)} \tag{4. 60}$$

El principal objetivo es mejorar o aumentar el valor *sensitivity* sin perjudicar el valor de *precision*, pero este objetivo a menudo está en conflicto ya que, cuando se incrementa el valor de TP para la clase minoritaria, el valor de FP de los ejemplos mayoritarios también se verá incrementado, esto reducirá el valor *precision*.

La métrica denominada *F1-score* es una puntuación que combina los valores de *precision* y *sensitivity*.

$$F1 - score = \frac{2TP}{2TP+FP+FN} \tag{4. 61}$$

Por otro lado, tanto los valores de *Accuracy* como de la métrica *F1-score* a menudo presentan sesgos muy señalados debido principalmente a que estas puntuaciones ignoran el rendimiento, la importancia y el peso que tienen los ejemplos de la distribución mayoritaria y los valores negativos:

$$NPV = \frac{TN}{(TN+FN)} \tag{4. 62}$$

De esta manera, en el trabajo [David, 2011] se definen las métricas de *Informedness*, *Markedness* y el coeficiente de correlación de Matthew MCC, como medidas no sesgadas que evitan el sesgo comúnmente encontrado en las métricas *sensitivity*, *precision* y *Accuracy*, respectivamente.

$$Informedness = sensitivity + specificity - 1 \tag{4. 63}$$

$$Markedness = precision + NPV - 1 \tag{4. 64}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4. 65}$$

Independientemente de su idoneidad o conveniencia de uso, las cinco métricas o ecuaciones anteriormente explicadas, son contrastadas en el capítulo 7 como funciones de ajuste que pueden ser utilizadas y aportando mucha información de relevancia a los algoritmos genéticos.

4.4.2 Métricas de evaluación en ajustes de regresión

Procediendo con la misma justificación que para tareas de clasificación, después de haber ajustado un modelo de regresión, es importante contar con ciertos valores que nos ofrezcan información de cómo de importante es dicho ajuste con respecto a los datos [Rabinovich, 2005]. Existen muchos términos cuantitativos que nos dan información muy

valiosa respecto a dicha medición. No obstante, una vez obtenidos los coeficientes de regresión, se sugiere[31] el cálculo de las siguientes cantidades:

$$SST = \sum(y_i - \overline{y})^2 \tag{4. 66}$$

$$SSR = \sum(y_i' - \overline{y})^2 \tag{4. 67}$$

$$SSE = \sum(y_i - y_i')^2 \tag{4. 68}$$

Donde SST es el sumatorio de los cuadrados de las diferencias de la variable respuesta *y* respecto de su media. SSR representa la suma de los cuadrados de las diferencias de la variable predictiva *y'* respecto a la media de la variable observada *y*, finalmente SSE es el sumatorio de los cuadrados de los residuales. Los residuales son los errores observados entre las dos variables *y*, *y*'. Una relación fundamental entre estas variables es la siguiente:

$$SST = SSR + SSE \tag{4. 69}$$

En la siguiente figura se representan e ilustran gráficamente las relaciones existentes entre ellas.

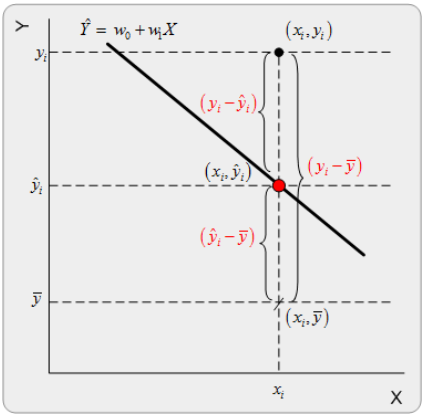


Figura 4. 23. Ilustración gráfica medición del ajuste

Una métrica importante que podemos obtener de estos valores es el denominado **coeficiente de determinación** R^2. Éste describe la proporción de varianza de la variable dependiente de los errores respecto del valor observado.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{4. 70}$$

Si el modelo de regresión es perfecto, SSE es cero, y R^2 es uno. Si SSE es casi igual a SST significaría que los errores serían muy altos y R^2 estaría muy próximo a cero.

Una vez introducidas las variables que hacen referencia a la suma de cuadrados, es necesario continuar con las variables que utilizan la media cuadrática, habitualmente utilizadas por el análisis de la varianza ANOVA en regresión múltiple. Está técnica

[31] Measuring the quality of fit. (Pag. 40). [Chatterjee, 2006]

estudia la igualdad de las medias para diferentes muestras poblacionales bajo la hipótesis de que éstas deben coincidir y por tanto el análisis de varianza sirve para comparar si los valores de un conjunto de datos numéricos son significativamente distintos a los valores de otro o más conjuntos de datos. No obstante la utilidad importante en un análisis de regresión respecto al análisis ANOVA son las variables medias cuadráticas que se utilizan frecuentemente como medida de comparación de los errores que se producen en los ajustes de regresión.

En la siguiente tabla se puede observar las expresiones de un análisis ANOVA y sus equivalencias entre variables.

Fuente	Suma de cuadrados	Media cuadrática	Cociente F
Regresión	*SSR*	*MSR = SSR / n*	*F = MSR / MSE*
Residuales	*SSE*	*MSE = SSE / n*	

Tabla 4. 5. Análisis ANOVA

Dónde MSE es la media del cuadrado debido al error de los residuales o también conocido como **error cuadrático medio** y MSR es la media del cuadrado debido a la regresión. El factor *F* es el cociente entre MSR y MSE y es la prueba de significación final en un análisis ANOVA. MSE representa la medición de comparación más común utilizada en los ajustes de regresión y es la que normalmente utilizaremos en los cálculos para la presente tesis.

Existen otras métricas de comparación, por ejemplo, el valor MAE, representa el valor medio absoluto de la diferencia de los valores observados y de los valores calculados en la regresión.

$$MAE = \frac{\Sigma|y_i - y_i'|}{n} \tag{4. 71}$$

Si el valor de MAE es igual a cero indicaría una predicción perfecta, en caso contrario, dicho valor se incrementaría proporcionalmente a las discrepancias en la predicción. MSE es mucho más sensible a los errores abruptos que MAE, debido a su componente cuadrática, penalizando a éstos cuanto mayor es el error. Una métrica muy parecida a MAE derivada desde el error cuadrático medio es:

$$RMSE = \sqrt{\frac{\Sigma(y_i - y_i')^2}{n}} \tag{4. 72}$$

RMSE es simplemente la raíz cuadrada del valor MSE. Mucho más útil puede llegar a ser el valor de NMSE[32], o sea el **error cuadrático medio normalizado**, donde se tiene en cuenta la varianza de toda la serie de valores observados de partida para así poder evaluar los errores de múltiples predicciones con respecto a esa varianza de referencia fija.

$$NMSE = \frac{MSE}{\sigma^2} = \frac{\frac{\Sigma(y_i - y_i')^2}{n}}{\frac{\Sigma(y_i - \bar{y})^2}{n}} \tag{4. 73}$$

[32] Parametric evaluation methods. (Pag. 276). [Lean et al., 2007]

Se puede observar que para un predictor perfecto, NMSE = *0*. Si NMSE = *1*, la predicción coincidiría con la media de la población inicial de partida observada, mientras que si NMSE > *1*, significaría que el rendimiento del predictor es mucho peor que el valor de la media de los valores de partida en cada ajuste de regresión efectuada.

Capítulo 5

Optimización de recursos en procesos de aprendizaje automático

Los procesos de aprendizaje que precisan ser automatizados, generalmente necesitan ser sincronizados previamente con otros procesos y a menudo tienen que quedar suspendidos en el tiempo a la espera de que se produzcan ciertos acontecimientos en su entorno para continuar con su funcionamiento habitual. Toda sincronización entre procesos implica también el acceso concurrente a recursos compartidos y no todos los sistemas operativos lo gestionan adecuadamente, y algunos no lo permiten, el mantenimiento efectivo de dicha concurrencia. Por tanto, se hace necesaria la implementación de recursos de sincronización y su emulación en cada uno de los sistemas operativos en donde se quiere administrar y automatizar la ejecución de dichos procesos. Además, la innovación y optimización de procesos y de sus interacciones con otros, debe ser llevada a cabo garantizando que si uno de ellos se queda esperando a una determinada acción en el sistema, éste no consuma recursos ni gastos de cómputo innecesarios en su entorno hasta la activación del evento por el que se está esperando. La forma de atacar el problema es diferente para cada sistema operativo, dependiendo de los recursos con los que se cuenta para cada uno de ellos. En este tema se solucionan los problemas de sincronización existentes en la operación experimental del TJ-II, que se fundamenta en ciclos pulsados y repetitivos, para la sincronización de procesos que se ejecutan en equipos diferentes y remotos. Se optimizan las tareas de gestión y administración entre subprocesos hijos, la comunicación con otros procesos concurrentes en el mismo ordenador y el entendimiento con otros aplicaciones que se ejecutan en otras máquinas muy distantes y que están esperando todos ellos a que se produzcan ciertos eventos, en un entorno de red de área local, para poder continuar con su ejecución ordenada en el tiempo. En la segunda parte del tema se incluye el diseño y los detalles técnicos realizados para una herramienta utilizada en el reconocimiento y la recuperación de patrones, basada en una arquitectura cliente/servidor. Dicha herramienta fue optimizada para incluir marcos embebidos de software como base de datos y un servidor web en la misma aplicación de escritorio, sin necesidad de instalaciones secundarias y complejas de las mismas, facilitando su portabilidad e instalación tanto en sistemas Windows, Linux y Mac.

5.1 Recursos de sincronización en entornos heterogéneos de computación

El Sistema de Distribución de Eventos Asíncronos SDEA del estellerator TJ-II [Vega et al., 2004] se ideó e implementó para proporcionar recursos de sincronización dentro de la red de área local del TJ-II. Este sistema, actualmente en operación, es fundamental en la misma y se encarga de distribuir todos los mensajes pertinentes a los sistemas experimentales durante la operación del TJ-II. Es un sistema software desarrollado para añadir capacidad de sincronización a la adquisición de datos, al control y a los entornos de análisis del TJ-II. Esto significa que SDEA no es un sistema de tiempo real. SDEA se basa en TCT/IP sobre redes ethernet.

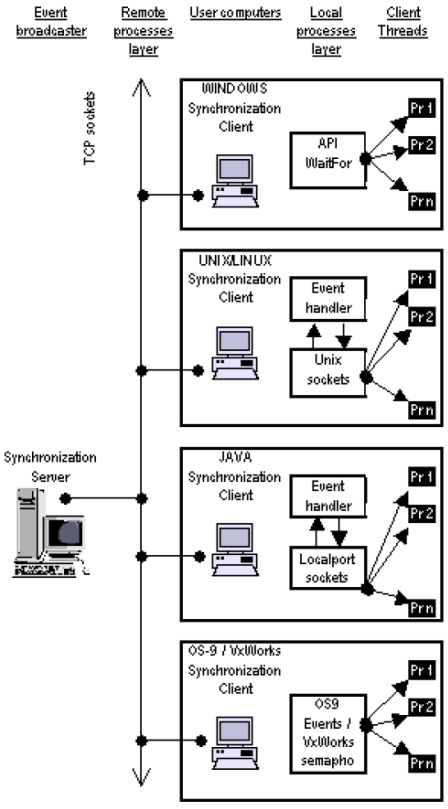


Figura 5. 1. Esquema de sincronización entre diferentes sistemas

Sin embargo, su tiempo de respuesta es adecuado cuando los requerimientos de sincronización pueden soportar algún retraso entre el inicio del evento y la recepción del mensaje. Los distribuidores de eventos, o servidores de sincronización en términos del SDEA, son ordenadores Windows. Ordenadores receptores, o clientes de sincronización, fueron también máquinas Windows en la primera versión de SDEA. Sin embargo, este hecho impuso una limitación importante sobre las capacidades de sincronización de muchos procesos de control, de aprendizaje automático, visualización, etc., que corren en otras máquinas con sistemas operativos muy diferentes. Llegados a esta situación, se añadieron al SDEA nuevos clientes de sincronización para diferentes entornos [Pereira et al., 2006], como son los sistemas operativos de tiempo compartido (UNIX y LINUX), sistemas operativos de tiempo real (OS-9 y VxWorks) y aplicaciones Java. Se necesitaba ampliar a estas plataformas las capacidades de sincronización desarrolladas. Las primitivas de sincronización que operan en estos sistemas son muy diferentes entre ellos y, por lo tanto, se han estudiado diferentes aproximaciones para realizar la misma funcionalidad en los diferentes entornos. La librería de hilos POSIX con sus primitivas de sincronización básicas, los cerrojos y las variables de condición, se utilizaron para poder realizarlo en sistemas UNIX/LINUX, mecanismos de comunicación interproceso para procesos concurrentes en sistemas de tiempo real como OS9 y VxWorks, y las primitivas "*synchronized - wait/notify"* en máquinas virtuales Java.

Cada cliente de sincronización (Linux, Unix, OS-9, VxWorks y Java), se desarrolla según un modelo de dos capas. La primera de ellas, **capa de procesos remotos**, se comunica con su servidor de eventos mediante sockets TCP y genera los recursos necesarios para el tratamiento de cada evento, el sistema de eventos es dinámico y admite definir nuevos eventos en cualquier momento y sin limitación en número. La segunda de las capas, **capa de procesos locales**, es la responsable de la sincronización de hilos locales mediante los pertinentes mecanismos de comunicación. A tal fin se ha desarrollado una librería de funciones que permite sincronizar hilos de ejecución locales con un número variable de eventos, relacionándolos con operaciones lógicas AND y OR. Cuando un evento se genera en el entorno de red del TJ-II, éste le llega a un servidor de sincronización basado en Windows; dicho servidor hace de distribuidor y lo propaga por la red ethernet (capa de procesos remotos) a todos los ordenadores de usuario que tengan corriendo un cliente de sincronización. Procesos o hilos locales son avisados (capa de procesos locales), mediante el cliente de sincronización, de que dicho evento se ha producido.

En mecanismos manejados por eventos, la ejecución asíncrona se diferencia de la síncrona sobre todo en que el trabajo realizado no consume recursos del núcleo, ni tiempo de CPU. Por tanto, cuando se produce un evento, la CPU no va sondeando los hilos o procesos que están esperando por ese determinado evento. De esta manera se pueden realizar otras tareas mientras el núcleo del sistema operativo no está esperando y que cuando se produce un evento los procesos encolados se activen. La familia de métodos *WaitFor* del sistema operativo Windows son las primitivas o mecanismos de sincronización existentes en estos sistemas y permiten sincronizar uno o más objetos. El método *WaitFor* es bloqueante hasta que ocurre después de un tiempo pre-especificado o cuando uno o todos los objetos que están esperando son señalizados y es no bloqueante si uno de los objetos está ya previamente señalizado. Por tanto, se detalla el problema y la necesidad de adoptar y desarrollar en otros sistemas, un mecanismo de eventos asíncronos e inter-proceso similar al que existe en plataformas Windows.

5.1.1 Sincronización en sistemas Unix para la clasificación de imágenes Thomson

La sincronización automática en la clasificación de las imágenes que genera el diagnóstico de esparcimiento Thomson del TJ-II con el entorno de operación experimental de descargas de dicho dispositivo, publicado en [Vega et al., 2005] y en [Makili et al., 2010], es llevada a cabo mediante un cliente de sincronización de procesos que espera a recibir los eventos que se producen en el entorno de red del TJ-II.

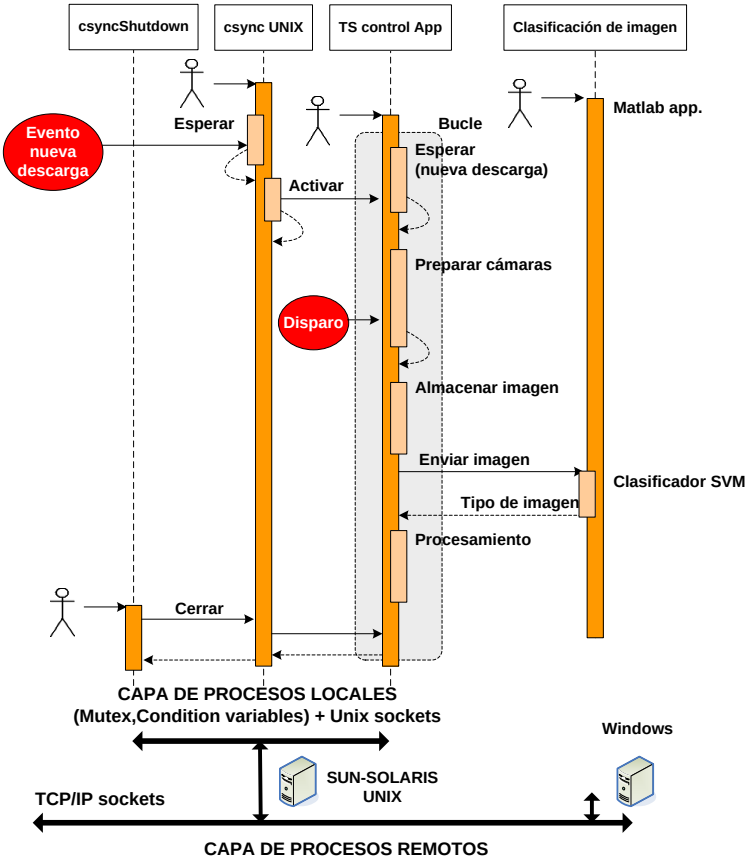


Figura 5. 2. Diagrama de secuencias clasificación de imágenes Thomson del TJ-II

Los procesos locales de adquisición de imágenes en el diagnóstico son administrados desde una estación de trabajo Sun-Solaris basada en Unix. Dichas aplicaciones quedan esperando hasta que el sistema SDEA notifica a dicha estación, el comienzo de la fase de pre-pulso de una descarga del TJ-II. Cuando ésta ocurre, un conjunto de tareas son activadas, como la adquisición de la imagen y su clasificación automática, finalmente, éstas vuelven a quedar suspendidas hasta que se produce la siguiente descarga. Es de destacar que antes de su automatización, tanto la sincronización en la generación de las imágenes como su posterior clasificación, se realizaba manualmente mediante supervisión humana. Igualmente y para estas publicaciones, se contribuyó también con la implementación cliente del sistema de control del reconocimiento de patrones, una herramienta en diferido de ordenamiento de imágenes y otra más de depuración para las imágenes clasificadas erróneamente por el sistema.

El cliente de sincronización realizado en UNIX fue recogido y explicado detalladamente en un informe técnico del CIEMAT [Pereira y Vega, 2005]. Para poder simular la familia de métodos *WaitFor* del sistema Windows en máquinas Unix/Linux se desarrolló un manipulador de eventos basado en un trabajo previo para entornos operativos Solaris [Nagarajayya y Gupta, 2000]. Dicho trabajo hace uso del **modelo de subscripció**n y consiste básicamente en que un proceso espera a que ocurra un determinado evento y se subscribe a una lista en la cual se queda latente hasta que el evento es señalizado por otro proceso. Para ello se recurre a las primitivas de sincronización que facilita la librería de hilos POSIX[33], disponible en la mayoría de sistemas Unix y Linux. Las librerias iniciales del trabajo original han sido modificadas y recompiladas con dos propósitos; el primero, para darle una funcionalidad inter-proceso de la cual carecía, solamente permitía comunicación inter-hilo, y el segundo, para hacer la portabilidad aún más extensa hacia otras plataformas no Solaris como hp-tru64 y kernel 2.6 y posteriores.

Las características principales del manipulador de eventos realizado son:

- Más de un hilo puede estar esperando en cualquier momento por el mismo evento.
- Falsas notificaciones no deben ser enviadas a los procesos.
- Garantía de que no existe peticiones continuas y por lo tanto no hay consumo de recursos y tiempos de CPU.
- Si un hilo o proceso que está esperando por un evento se muere, no bloquee todo el manipulador de eventos.

El soporte de notificación asíncrona realizado en el manipulador de eventos se fundamenta en dos primitivas POSIX, los cerrojos las variables de condición. La mayoría de los sistemas UNIX/Linux soportan estas variables de sincronización. Los cerrojos ofrecen las mismas posibilidades de sincronización que los semáforos binarios pero son menos costosos que éstos. La norma POSIX define un cerrojo[34] como un objeto de sincronización usado por los hilos para secuenciar sus accesos a los datos que comparten. Los cerrojos se utilizan normalmente para serializar el acceso a los recursos compartidos. Convierten al hilo que lo usa en propietario absoluto de la sección de código que engloban las llamadas de bloqueo y desbloqueo. Las variables de condición son objetos que le permiten a los hilos suspender su ejecución repetidas veces hasta que sea cierto un predicado asociado. Una variable condicional crea un entorno seguro para comprobar la veracidad de una condición. Cuando un hilo obtiene un cerrojo (las variables condicionales trabajan siempre asociadas a un cerrojo), se comprueba la condición bajo la protección del cerrojo. Ningún otro hilo podrá alterar ningún otro aspecto de la condición sin poseer el cerrojo. Si la condición es cierta, el hilo completa su trabajo y libera el cerrojo. Si la condición es falsa, el cerrojo se libera automáticamente, y el hilo se duerme esperando sobre la variable condicional. La lista de eventos se define como una tabla de punteros a una estructura establecida. Se podría haber implementado las estructura de eventos de tal forma que residiera en un espacio compartido para que pudieran ser accedidos por otros procesos y no solo por los hilos del proceso en el que reside el manipulador. Para realizar esto habría que determinar si nuestra biblioteca POSIX dispone de las funciones *pthread_mutexattr_getpshared* y *pthread_mutexattr_setpshared*

[33] POSIX. http://standards.ieee.org/develop/wg/POSIX.html
[34] IEEE P1003.1, Draft 3. Pag. 78. http://www.open-std.org/jtc1/sc22/open/n4217.pdf

con la posibilidad así de que hilos de diferentes procesos compartieran cerrojos habilitando el valor PTHREAD_PROCESS_SHARED de las funciones anteriores. Pero la realidad es que en la mayoria de los sistemas UNIX/LINUX esta posibilidad no está disponible[35] (símbolo _POSIX_THREAD_PROCESS_SHARED del fichero <unistd.h>). Por ello se ha diseñado este sistema de comunicación inter-proceso más estándar basado en sockets UNIX. En un proceso servidor residirá el manipulador de eventos y procesos clientes se comunican con estos *sockets* para subscribirse al manipulador de eventos y esperar a que estos sean señalizados; así podremos además disponer de tantos procesos servidores o manipuladores de eventos como necesitemos.
Una vez resuelto el problema de la sincronización inter-proceso de forma concurrente para las máquinas locales, finalmente, para conseguir el cliente de sincronización en ordenadores de tiempo compartido mediante sistemas operativos de la familia Unix-Linux se implementó un protocolo de comunicación con el servidor de sincronización central por medio de la red ethernet local, basándose en sockets TCP-IP.

5.1.2 Sincronización de algoritmos de aprendizaje en entornos de supercomputación Linux

En entornos de supercomputación y de programación paralela basados en el intercambio o paso de mensajes, como es el interfaz de pasos de mensajes MPI (siglas del inglés), los procesos son lanzados mediante la ejecución de scripts que a su vez hacen llamadas al sistema operativo para poder lanzar otros ficheros ejecutables.

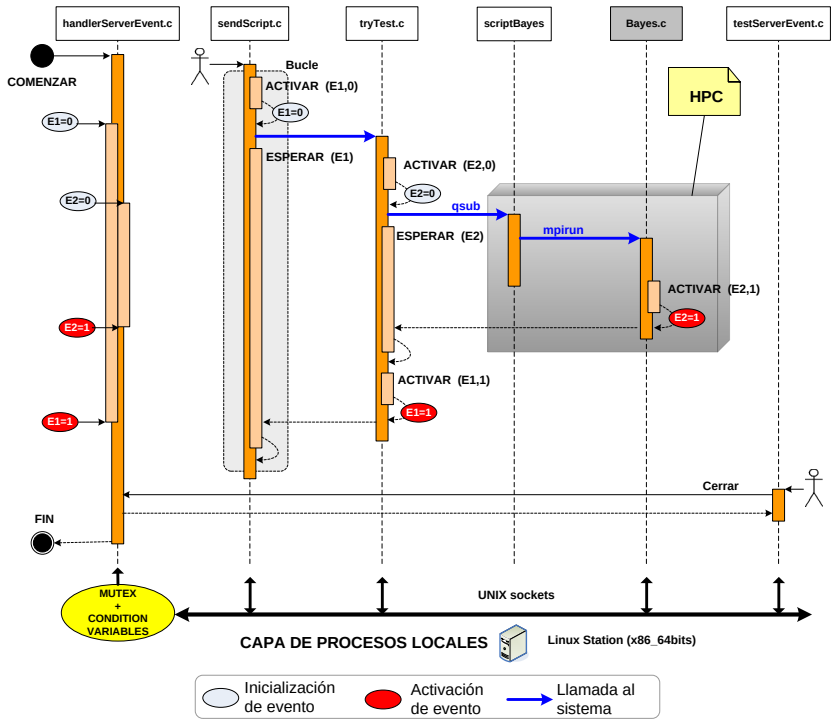


Figura 5. 3. Diagrama de secuencias para sincronizar procesos que hacen uso de llamadas al sistema

[35] Apartado 10.6.1 Atributos de un cerrojo. Pag. 403. [Márquez, 2004]

Todas estas llamadas al sistema tienen el inconveniente de la falta de sincronización entre la ejecución de la llamada y su finalización. Ante una llamada al sistema el proceso que lo lanza pierde el contacto con él y continúa su ejecución secuencial sin esperar a que termine el proceso lanzado. Si a este inconveniente le añadimos la dificultad de tener que repetir el proceso de ejecución varias veces hasta que pueda converger un cierto criterio, entonces, la sincronización entre dichos procesos se hace si cabe aún mucho más necesaria y relevante.

Aprovechando el entorno de sincronización realizado, éste se utilizó para poder sincronizar un algoritmo bayesiano, haciendo determinista su estado de finalización, para poder ejecutarlo otra vez repetidamente y con el objetivo de poder buscar el conjunto de observaciones de entrenamiento más óptimo que consiga las mejores tasas de acierto sobre un conjunto de observaciones de prueba muy elevado en cantidad y número de características.

La inclusión de una única línea de código en el programa fuente original con el objetivo de poder activar el evento de finalización de dicho programa paralelizado, es el único requisito en la utilización de la librería de sincronización realizada en lenguaje C y que fue recompilada para entornos de supercomputación Linux de 64 bits.

5.1.3 Monitorización de información sincronizada en aplicaciones JAVA

Siguiendo el mismo criterio que para entornos Unix/Linux, se ha conseguido obtener un cliente de sincronización Java que hace uso de unas librerías permitiendo emular las llamadas *WaitFor* de Windows y consiguiendo así facilitar un soporte nativo en ordenadores que tienen la máquina virtual Java instalada.

Para emular dichas llamadas se hizo uso de las primitivas "*synchronized*" y los métodos "*wait/notify*" de Java. El primero hace que un fragmento de código sea protegido de accesos concurrentes. El método *wait* provoca que un hilo se quede esperando y solamente puede ser invocado desde un código previamente sincronizado, dicho hilo se queda entonces dormido hasta que otro hilo ejecute el método *notify*(). Mediante la combinación de la palabra clave *syncronized* y los métodos *wait/notify* un manipulador de eventos inter-hilo puede ser programado en máquinas Java similar al obtenido en sistemas de la familia Unix. La comunicación interproceso dentro de un mismo ordenador se resuelve utilizando *sockets* de puerto local que facilita el lenguaje Java. El cliente de sincronización crea un *socket* servidor de puerto local y hilos clientes esperan, mediante el método *read*(), a que se les comuniquen la señalización de ciertos eventos. Cuando la notificación del evento le llega al cliente de sincronización, éste se lo comunica a todos sus hilos clientes por medio del método *write*(). Finalmente *sockets* TCP son utilizados para comunicarse con el servidor de sincronización que hace de distribuidor para todos sus clientes.

El paquete ha fue diseñado a imagen de lo conseguido para sistemas de tiempo compartido (Unix/Linux) en cuanto a la sincronización de procesos locales, por lo que la

mayoría de lo comentado al respecto, en cuanto a la estructura funcional, para dichos sistemas, es aplicable a lo conseguido en entornos Java.

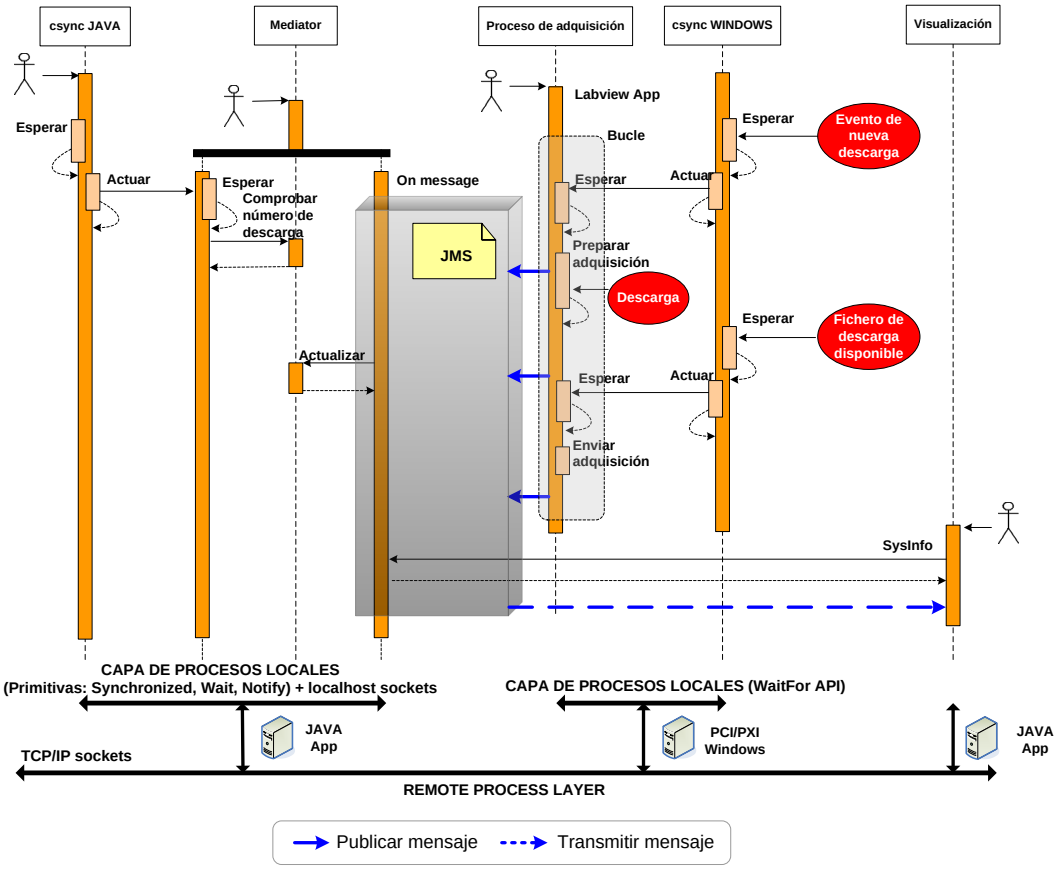


Figura 5. 4. Diagrama de secuencia para la sincronización de aplicaciones JAVA

En [Sánchez et al., 2006] se aporta una aplicación de visualización basada en JAVA y que utiliza también el presente sistema de sincronización de eventos, viniendo así a complementar el seguimiento de la operación en el entorno del TJ-II, que se fundamenta sobre todo en servidores web y aplicaciones de visualización JAVA que hacen uso de un servicio de mensajería. En el sistema de participación remota del TJ-II se colaboró también con este sistema de sincronización, recopilación realizada en la publicación [Vega et al., 2005b], otros trabajos consistieron en la implementación de una aplicación de mensajería, englobada dentro de una arquitectura orientada a mensajes JMS de JAVA y que funciona mediante un protocolo de **publicación-subscripción**, dado a conocer en [Sánchez et al., 2007b]. Las aplicaciones de participación remota instaladas para controlar el sistema de adquisición de datos del TJ-II han permitido comandar y seguir la operación de descargas del estellerator TJ-II desde Cadarache en Francia haciendo uso de estos sistemas, dicho seguimiento fue publicado en [Vega et al., 2006], y en la que también se ha participado. En la figura anterior se puede ver la utilidad del sistema de sincronización realizado para aplicaciones JAVA. Un proceso llamado 'Mediador' hace uso de la sincronización de eventos que se produce en el entorno de red de área local del TJ-II. Cuando se señaliza un evento de que va a producirse una nueva descarga en el dispositivo experimental, este proceso comprueba el número de descarga último y lo coteja con los que tiene registrados en su base de datos y que son actualizados constantemente por los mensajes que publican constantemente los sistemas de adquisición de datos. Si existe

algún equipo en el que no coincidiera dicho número significaría que hubo algún tipo de problema para dicho equipo. Un sistema de visualización y monitorización utilizado por los administradores de la operación, que aporta información del estado en el que se encuentra cada uno de los equipos de adquisición de datos, utiliza los mensajes que le envía el sistema de distribución de mensajes para actualizar constantemente el estado de dichos equipos. El proceso Mediador además aporta toda la información necesaria al proceso de monitorización cada vez que éste es ejecutado por primera vez, informándole desde ese momento de cuál es la situación de todos los equipos.

5.1.4 Sincronización de eventos del TJ-II en sistemas de tiempo real

En sistemas de tiempo real existen mecanismos interproceso de eventos mucho más flexibles y convenientemente implementados en el núcleo del sistema, los cuales se han aprovechado para conseguir obtener los clientes de sincronización. En sistemas de tiempo real como OS-9 y VxWorks no es necesario emular el conjunto de primitivas *WaitFor* característico de los sistemas operativos Windows, y en los que nos hemos ido basándonos hasta ahora para conseguir un manipulador de eventos inter-proceso. La razón principal es que estos sistemas de tiempo real se caracterizan porque utilizan variables globales accesibles por todos los procesos que corren en ellos; además facilitan otros mecanismos de comunicación inter-proceso particulares y muy útiles que no se encuentran en los sistemas operativos de tiempo compartido.

Los eventos del sistema OS-9 son variables globales mantenidas por el sistema operativo, se puede decir también que, son una clase especial de semáforos que aceptan múltiples valores, con ellos se pueden sincronizar procesos concurrentes que acceden a recursos compartidos. Un proceso cliente puede detener su ejecución esperando a que un determinado evento se genere y continúa su ejecución cuando otro proceso señaliza dicho evento.

En el sistema operativo de tiempo real VxWorks se utilizó la librería de semáforos *semLib* (#include "semLib.h") para obtener los clientes de sincronización.

Utilizado para sincronización de tareas, un semáforo representa una condición o evento por la cual la tarea se puede quedar esperando. Inicialmente el semáforo está vacio o no disponible. Una tarea espera por el semáforo mediante la rutina *semTake*(), otra tarea señaliza el evento mediante *semGive*(), esto permite que la primera tarea deje de estar esperando por el semáforo. Este modelo de utilizar eventos inter-proceso se recomienda en aplicaciones de tiempo real para VxWorks. La rutina *semFlush*() desbloquea todas las tareas que están esperando por un semáforo. Estas rutinas para comunicación inter-proceso han sido utilizadas para obtener el cliente de sincronización bajo VxWorks.

Con la implementación de los clientes de sincronización tanto para OS-9 como para VxWorks, estos sistemas pueden informarse de los eventos que se producen en el entorno de red del TJ-II con los procesos locales de dichas máquinas. Se da soporte así al sistema de control del TJ-II, basado en sistemas hardware VME con microprocesadores Motorola 68k y PowerPC, que puede subscribirse a los eventos que se producen en dicho entorno de red con el objetivo de poder ser informado y realizar ciertas tareas que deben estar sincronizadas con las tareas principales del TJ-II.

5.2 La herramienta de búsqueda de señales en la base de datos del TJ-II

El sistema de participación remota del TJ-II [Vega et al., 2005b] fue diseñado para que sirviera de punto común de acceso a todas las funcionalidades y herramientas que están disponibles para la monitorización y gestión tanto de la adquisición de datos como del control de diagnósticos. Este sistema engloba un conjunto de recursos que permite a un usuario monitorizar diferentes aspectos relacionados con la operación del TJ-II así como de la gestión histórica de las descargas que ya se han producido. También se puden configurar los canales de adquisición de todos los dispositivos que integran el sistema de adquisición de datos, siendo todos ellos configurables por los propios usuarios. Existe la posibilidad también de un registro electrónico para poder gestionar incidencias y comentarios sobre la gestión diaria de descargas durante la operación del TJ-II. Otros servicios que ofrece el sistema de participación remota es la de poder realizar reservas y salas de videoconferencia. Es de señalar que todos estos servicios son accesibles tanto desde la red de área local del TJ-II como desde fuera de las instalaciones del centro. Así por ejemplo en [Vega et al., 2006] se ha reflejado el seguimiento de la operación del TJ-II desde Cadarache (Francia) todo ello posible gracias a la actuación remota de los servicios implementados.

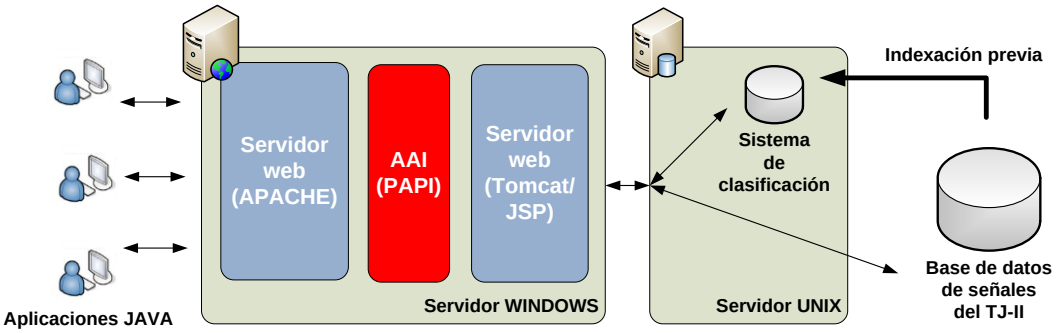


Figura 5. 5. Acceso a la herramienta de búsqueda de señales en el TJ-II

La seguridad y el servicio de control de acceso corre a cargo de una infraestructura de autenticación y autorización denominado PAPI [Castro y López, 2001] y que también fue instalado y configurado en la integración del sistema de participación remota del TJ-II como identidad federada de autenticación y autorización entre diferentes laboratorios europeos de investigación en Fusión [Castro et al., 2008][Castro et al., 2008b]. El entorno de participación del TJ-II esta basado en una arquitectura multicapa de tres niveles [Vega et al., 2004b], capa cliente, capa intermedia, y capa de integración de datos. Su objetivo es la de poder separar procedimientos para en el caso de que se necesite hacer alguna modificación no influya a los servicios implementados en otros niveles. Las aplicaciones

cliente utilizadas para acceder al sistema de participación remota del TJ-II se pueden dividir en dos grupos: navegadores web y aplicaciones JAVA. Éstas comparten el protocolo HTTP como protocolo base de comunicación entre los clientes y la capa intermedia. Un ejemplos de aplicación cliente que hace uso de este protocolo de comunicación es la aplicación de búsqueda se señales realizada en JAVA y que implementa los métodos de búsqueda explicados en el capítulo 3. Esta aplicación utiliza el sistema de gestión de versiones JWS (Java Web Start). Este sistema permite tanto el arranque de aplicaciones de una forma fácil para el usuario, como una actualización de la aplicación de forma totalmente automática y completamente transparente para el usuario. La tecnología JWS se basa en ficheros JNLP (JAVA Network Launching Protocol), de forma que cuando existe una nueva versión de la aplicación, los usuarios del sistema de participación remota del TJ-II verán actualizada su aplicación automáticamente.

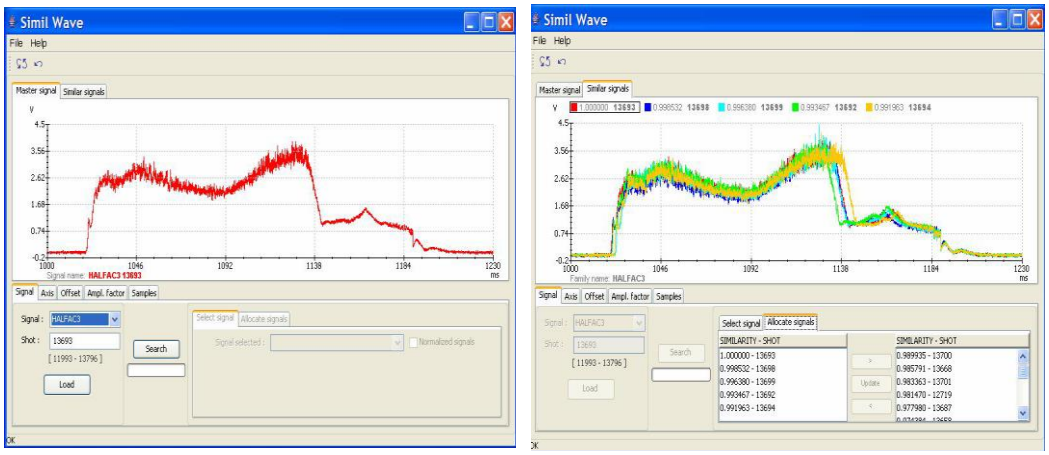


Figura 5. 6. Primera versión para la búsqueda de señales instalada en el TJ-II

En la figura de arriba se puede observar la primera versión de la aplicación de búsqueda de patrones instalada en el sistema de participación remota del TJ-II. En esta primera versión solamente está disponible el reconocimiento de señales completas.

La parte servidora que atiende a las peticiones de la aplicación JAVA reside en un servidor UNIX y fue programado en lenguaje C. A su vez las señales son descargadas de la base de datos del TJ-II mediante sockets TCP/IP. La aplicación servidora, integrada en la tercera capa del sistema de participación remota, devuelve el contenido dinámico explícitamente peticionado por los clientes JAVA.

5.3 La herramienta de reconocimiento de patrones en el JET

Una vez desarrolladas las técnicas de búsqueda de patrones dentro de las señales, se implementaron estos métodos en el cluster de ordenadores para análisis del JET (JAC, abreviatura en inglés). En el JET, el concepto de seguridad informática es muy acusado y las instalaciones software que se hacen en sus equipos están muy controladas. Existen normas muy restrictivas al respecto y con fuertes limitaciones para poder realizar comunicaciones desde localizaciones externas al recinto de investigación. Todas estas restricciones obligaron a realizar modificaciones en el software del reconocimiento de patrones inicialmente instaladas en el sistema de participación remota del TJ-II. Es de resaltar que el desarrollo que se realizó en el JET, para poder dar servicio en la búsqueda y reconocimiento de patrones, fue un desarrollo muy ajustado a los requerimientos del JET.

Las comunicaciones HTTP de la aplicación cliente se substituyeron por comunicaciones mediante sockets TCP/IP y la aplicación servidora se adaptó desde código C para sistemas UNIX hacia sistemas LINUX. El sistema de base de datos relacional necesario para poder indexar las señales se realizó mediante el software PosgreSQL. El acceso y lectura de las señales brutas se realizaron mediante una librería de acceso a datos facilitada por el JET.

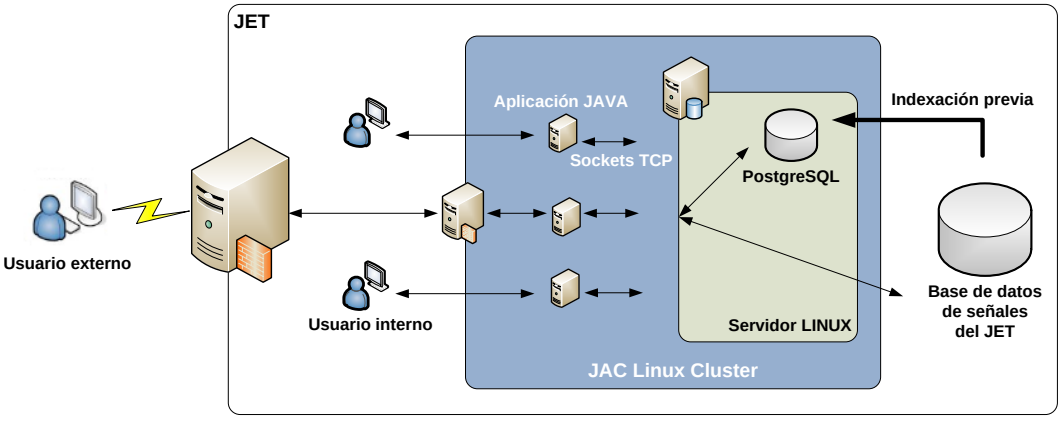


Figura 5. 7. Acceso a la herramienta de búsqueda de señales en el JET

A continuación se muestran las pruebas realizadas mediante la búsqueda de formas de onda completas en dos señales diferentes del JET pertenecientes a la campaña C17 [Vega, 2007]. Con un ordenador cuyas características son Pentium-Pro 200 MHz, 1 procesador, 1GB RAM, se obtuvieron los siguientes resultados:

Señal	Disparos (Campaña C17)	Tamaño (MB)	Tamaño adicional (MB)	Tiempo de búsqueda (ms)
KG1V/LID3	**689**	**10.51**	**1.82 (17%)**	**25**
BOLO/TOPI	**692**	**67.58**	**1.88 (3%)**	**20**

Tabla 5. 1. Tiempo de búsqueda de señales completas en el JET

Figura 5. 8. Segunda versión de la herramienta de búsqueda instalada en el JET

Las pruebas de tiempo realizadas para la búsqueda de patrones dentro de señales utilizando la técnica de primitivas de longitud adaptable, arrojaron los siguientes resultados:

Señal	Disparos (Campaña C17)	Tamaño (MB)	Tamaño adicional (MB)	Tiempo de búsqueda (ms)
EFIT/WDIA	**706**	**1.23**	**0.80 (65%)**	**300**
KK3/TE13	**501**	**48.93**	**0.7 (1%)**	**90**

Tabla 5. 2. Tiempos de búsqueda para patrones dentro de señales en el JET

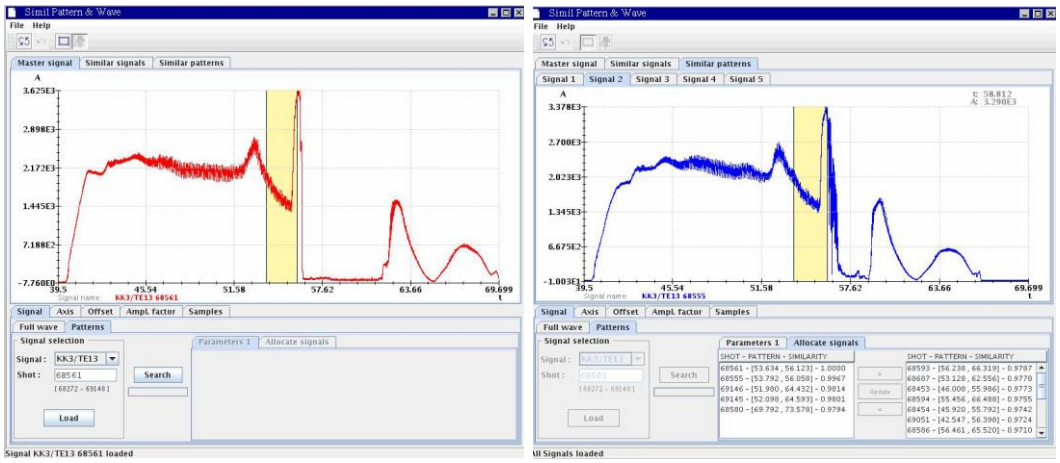


Figura 5. 9. Herramienta para la búsqueda de patrones en el JET

Una vez instalada la aplicación en el JAC cluster del JET, ésta ha servido para poder realizar estudios y extraer conocimiento a partir de ciertos patrones de comportamiento que se han observado en algunas señales. Por ejemplo en [Rattá et al., 2008] para la investigación de fenomenología física reflejada en el comportamiento de ciertos patrones morfológicos muy característicos.

5.4 Distribución abierta y remota para la recuperación de patrones

En la publicación [Pereira et al., 2010] se da a conocer la implementación de una herramienta software multipropósito, no solo para señales del JET o del TJ-II, multiplataforma (Windows, Linux, Mac) y que hace uso de una arquitectura distribuida en un entorno de participación remota mediante internet. El objetivo es dar visibilidad, complementar y publicitar el trabajo principal de la búsqueda de patrones. La particularidad de esta implementación es el uso de marcos embebidos de software que evitan configuraciones e instalaciones complejas y ajenas de otros servicios que también son necesarios en la utilización de la herramienta, como son las bases de datos y los servidores web, encapsulando todo ello en una única aplicación de escritorio tanto para la aplicación cliente como para la aplicación servidora. Una dificultad importante a solventar en este trabajo fue la recuperación de grandes cantidades de información y el envío de las señales brutas originales para poder ser visualizadas entre diferentes ordenadores muy distantes y remotos. Para ello se han aplicado diferentes técnicas de compresión de datos sin pérdida de información recopiladas en [Vega et al., 2007b] y que algunas fueron revisadas para su utilización en tareas de tiempo real y orientadas a descargas de pulso largo, consiguiendo migrar satisfactoriamente dichos algoritmos hacia librerías JAVA. De esta manera, el envío de todas las señales se realiza de forma comprimida y en la recepción de los mismos es descomprimida por los diferentes clientes remotos, haciendo así más transparente y ágil el trasiego de grandes cantidades de información. El entorno distribuido y abierto realizado, complementa la recuperación y la disponibilidad de los datos y puede ejecutarse en cualquier entorno de computación y sistema operativo, con el único requerimiento de que se disponga de la máquina virtual JAVA versión 5 instalada. Además, el uso de componentes embebidos o embarcados como son la base de datos relacional apache-DERBY y el contenedor de servlets JETTY, favorece la simplicidad y la distribución de la herramienta sin necesidad de complejas configuraciones de software. Todo ello conforma una herramienta muy útil, fácil de instalar con autoinstalador incluído y ejecutable tanto en entornos locales como en entornos de participación remota y de alcance más amplio.

5.4.1 Arquitectura multicapa basada en 3 niveles

Los sistemas abiertos y distribuidos constituyen hoy en día la base sobre la que se construyen las aplicaciones software, debido fundamentalmente a la gran difusión y creciente potencia de los ordenadores personales y a la aparición de redes de cobertura

global como Internet. Por ello, las aplicaciones han dejado de ejecutarse de una forma aislada y de manera monolítica, para pasar a ejecutarse en entornos distribuidos y a tener que interactuar con sistemas externos desarrollados por otras organizaciones (proveedores, clientes, usuarios, etc.). La mayoría de las propuestas actuales para describir la arquitectura global de los sistemas distribuidos se basan en la identificación y separación de puntos de vista independientes. Cada uno de estos puntos de vista se centra en una serie de aspectos concretos, abstrayéndose del resto, y simplificando por tanto el diseño. Se ha realizado una aplicación cliente/servidora distribuida y basada en una arquitectura de tres capas o niveles e implementada en JAVA. Mediante aplicaciones desarrolladas en este marco, garantizamos que el código implementado es capaz de correr y de ejecutarse en cualquier entorno operativo con el único requisito de que dicho entorno tenga instalado la máquina virtual JAVA correspondiente. Para el desarrollo realizado en este proyecto es requisito necesario para su correcto funcionamiento que esté instalada la máquina virtual Java versión 5 o posterior. Los tres niveles de la arquitectura distribuida sobre la que se realizará el diseño correspondiente son:

1. Nivel de presentación.

Éste es el nivel encargado de generar la interfaz de usuario en función de las acciones llevadas a cabo por el mismo. La capa de presentación contiene los componentes necesarios y más ligeros para habilitar la interacción del usuario con la aplicación. Los componentes de la interfaz de usuario deben mostrar las señales al usuario, obtener y validar los datos procedentes del mismo e interpretar las acciones de éste que indican que desea realizar una operación con los datos, la petición de búsqueda de una señal y la visualización gráfica de las señales o patrones más parecidos. Asimismo, la interfaz debe filtrar las acciones disponibles con el fin de permitir al usuario realizar sólo aquellas operaciones que le sean permitidas en un momento determinado.

2. Nivel de negocio.

Contiene la lógica que modela los procesos de negocio y es donde se realiza todo el procesamiento necesario para atender las peticiones del usuario. Aplicaciones del lado del servidor ("*servlets*") se encargarán de atender las peticiones de los clientes, gestionar la concurrencia de dichas peticiones y responder adecuadamente a cada uno de ellos en base a contenido dinámico y específico.

3. Nivel de integración de datos.

Es el encargado de hacer persistente toda la información, así como de suministrar y almacenar la información para el nivel de negocio. Casi todas las aplicaciones y servicios necesitan almacenar y obtener acceso a un determinado tipo de datos.

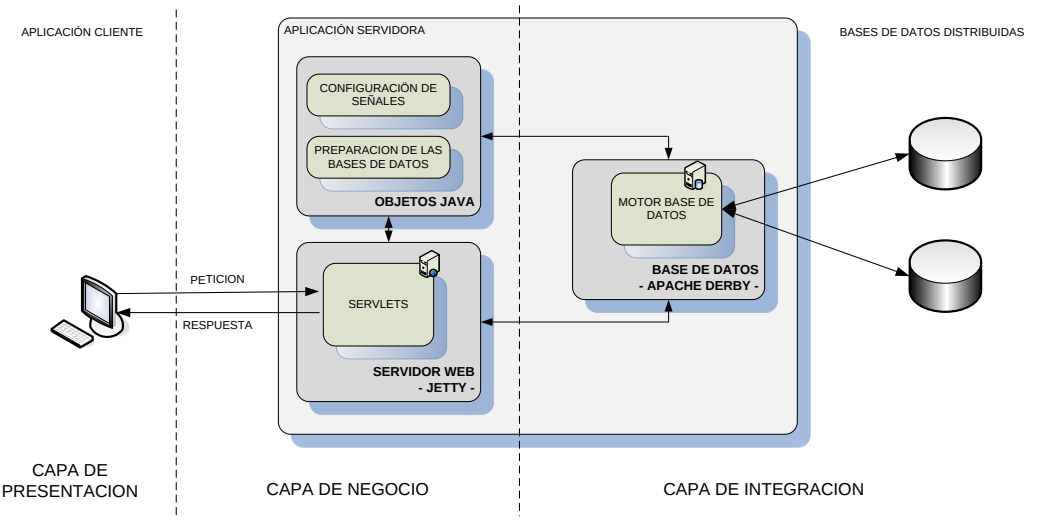


Figura 5. 10. Arquitectura distribuida para la búsqueda de patrones

Tomando como base esta arquitectura, se utilizan entornos y marcos de desarrollo como son los sistemas embarcados tanto en el nivel de la lógica de negocio, un contenedor web embarcado denominado JETTY, como en la persistencia de los datos, una base de datos embarcada denominada APACHE-DERBY. El objetivo que se persigue con tal decisión es la de poder facilitar una herramienta lo más simple y sencilla en cuanto al proceso de instalación, distribución y configuración para el usuario y sin perder de vista ninguna funcionalidad.

La plataforma de desarrollo y despliegue de componentes utilizados es J2EE ("*Java 2 Enterprise Edition"*). J2EE propone una arquitectura multi-capa como estilo arquitectónico para las aplicaciones que se desarrollen bajo esta plataforma. Esto quiere decir que para crear una aplicación típica J2EE es necesario dividirla en capas, desarrollar los componentes que sean necesarios y colocarlos en la capa correspondiente. Las aplicaciones J2EE se construyen ensamblando componentes y desplegándolos en un contenedor. Las aplicaciones están formadas por componentes que se ejecutan dentro de contenedores. Los contenedores proporcionan un entorno de ejecución y el acceso a un conjunto de servicios de bajo nivel a los componentes que forman la aplicación. J2EE define un modelo de componentes y contenedores abierto y estándar. Esto quiere decir que los contratos entre los componentes, los contenedores y los servicios que tienen que proporcionar los define una especificación estándar. De esta manera, se consigue que cualquier fabricante pueda desarrollar contenedores capaces de ejecutar componentes J2EE; sólo tiene que cumplir los contratos estándar e implementar los servicios requeridos por la especificación. Se ha optado por utilizar el contenedor de aplicaciones embarcadas JETTY. Éste es un contenedor de aplicaciones enteramente implementado en JAVA con la particularidad de que puede ser incrustado en cualquier aplicación y código desarrollado en dicho entorno. En la capa de integración de datos hacemos uso de un marco de mapeo objeto/relacional denominado APACHE-DERBY. Al igual que el servidor web utilizado, DERBY es una base de datos relacional realizada enteramente en JAVA con capacidad igualmente de poder ser embarcada en cualquier aplicación de este tipo. El hecho de poder incrustar tanto un servidor web como una base de datos relacional en código nativo, nos brinda la posibilidad de poder aprovechar toda la potencia de estos marcos de trabajo con la flexibilidad de encapsular acciones y detalles que tienen que ver

con la distribución y la instalación del software, integrando así estas acciones en una sola aplicación lista para ser ejecutada sin necesidad de instalar y configurar ningún otro componente secundario. Completando la lógica de negocio, tenemos la implementación de los objetos JAVA que no son más que simples clases Java necesarias para el acondicionamiento de las señales y para la preparación de la base de datos. Estas clases se encargarán de convertir y comprimir las señales en un formato binario para ser almacenadas en un repositorio y de preparar la base de datos de las señales pre-procesadas para que puedan ser accedidas posteriormente por medio de consultas a petición de los clientes remotos.

5.4.2 Protocolos de comunicación utilizados

En la siguiente figura, se puede observar el protocolo de comunicación utilizado en cada transmisión realizada entre todas las partes integrantes de la aplicación distribuida.

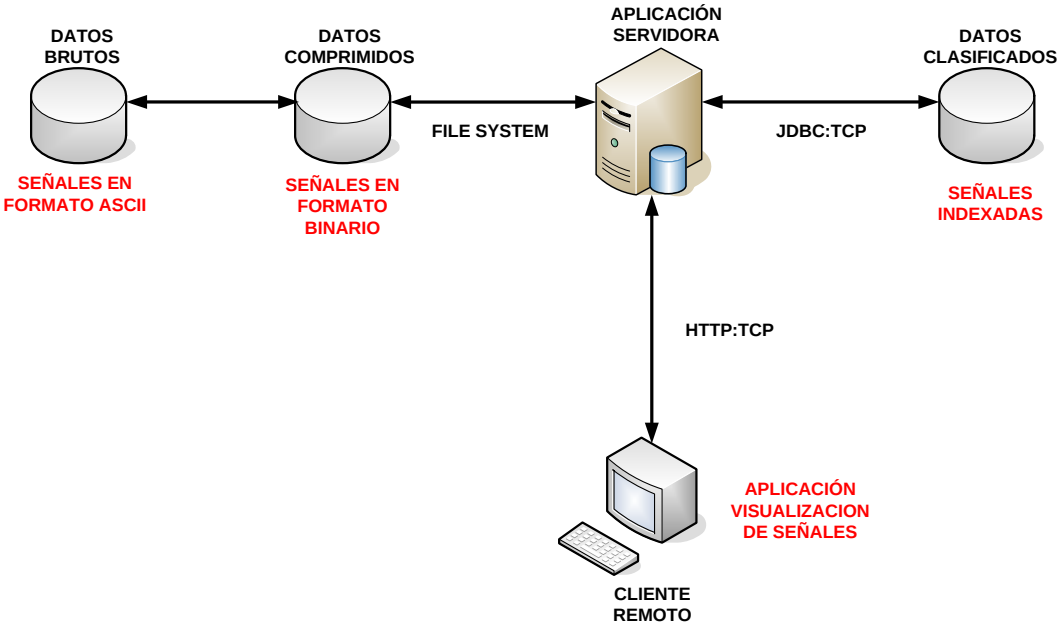


Figura 5. 11. Protocolos de comunicación

Las señales originales y comprimidas estarán almacenadas en forma de ficheros binarios en carpetas bien definidas y gestionadas por el sistema operativo. Cada vez que la aplicación servidora necesita hacer uso de una señal, éste accederá a ella por medio de rutinas de acceso a ficheros localizados en una carpeta determinada bien sea en el propio sistema donde se ejecuta el programa servidor o por medio de un acceso de sistema de ficheros en red NFS habilitado por el sistema operativo para tal propósito sobre directorios de equipos remotos. La información almacenada por las bases de datos puede estar igualmente distribuida en diferentes plataformas. La comunicación entre la aplicación servidora y estos datos remotos se realizará por medio de conexiones JDBC de JAVA sobre TCP en algún directorio concreto del equipo remoto o cabe la posibilidad también de que las bases de datos residan igualmente en el mismo equipo local que el programa servidor. La comunicación entre las aplicaciones cliente y la aplicación

servidora se realiza por medio de peticiones HTTP sobre TCP, dando la posibilidad de que la elección del puerto sea configurable por el administrador del sistema. Una de las grandes ventajas de escoger este tipo de protocolo es que podemos realizar comunicaciones HTTP sobre el puerto 80, facilitando así el paso de muchos cortafuegos que suelen dejar este puerto abierto para el acceso a Internet. Desde el punto de vista del desarrollo, el paquete java.net maneja la clase URL, que representa una dirección de Internet. Esta clase tiene métodos generalizados muy útiles para la comunicación y el intercambio remoto sobre el protocolo HTTP.

5.4.3 Entorno operativo

El proceso final para la selección de un patrón, búsqueda y recuperación de la información más parecida o similar, se describe en el esquema de la figura siguiente.

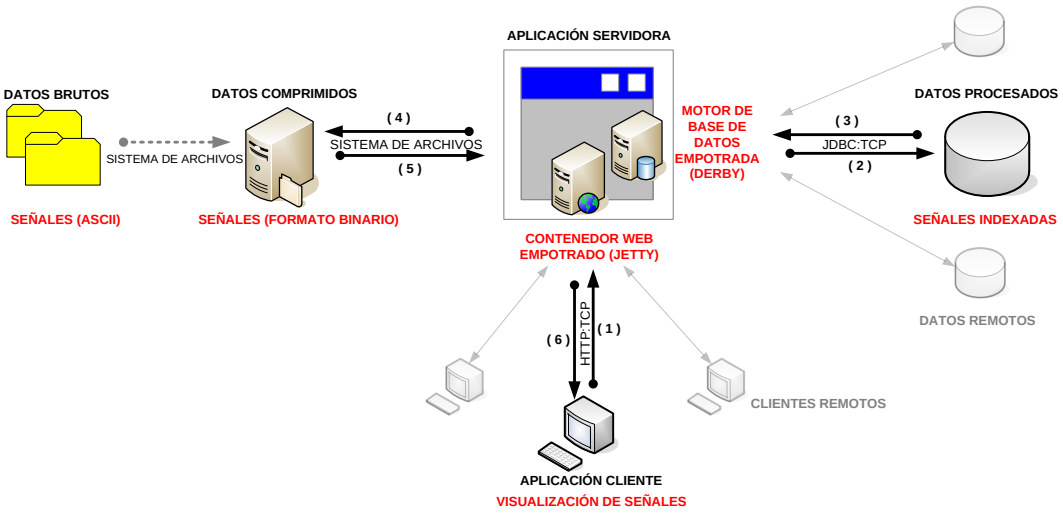


Figura 5. 12. Proceso operativo básico para la recuperación y búsqueda de señales

En el paso (1), un usuario selecciona un patrón de una señal y ordena una petición de búsqueda. Toda la información relativa a dicha operación se codifica en una sentencia URL (Uniform Resource Locator, del inglés), también se especifica en dicha sentencia en que base de datos se debe buscar. En el paso (2) y una vez llegada la consulta a la aplicación servidora, se confecciona una consulta hacia la base de datos indicada y formulada por el usuario. En (3) se devuelven todas las filas de la base de datos que casan con la petición formulada anteriormente haciendo uso del protocolo JDBC. Posteriormente en (4) y (5) se recuperan los datos comprimidos de las señales indicadas en la consulta devuelta por la base de datos. Finalmente en (6) se devuelve al cliente final cada señal con toda la información disponible mediante contenido dinámico y cambiante para cada respuesta.

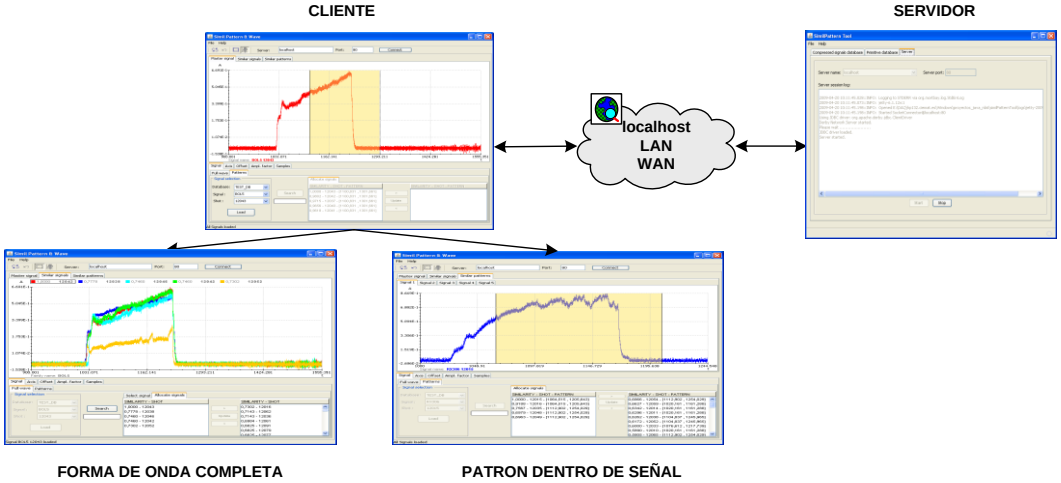


Figura 5. 13. Aplicaciones de usuario gráficas implementadas

El producto final es una herramienta (Figura 5. 13) basada en una aplicación servidora central y varias aplicaciones clientes que pueden ejecutarse en diferentes sistemas operativos y localizados en entornos físicos distantes y remotos.

5.5 Conclusiones

En sistemas de tiempo compartido para máquinas de la familia Unix-Linux se ha descrito una notificación genérica de eventos asíncronos e interproceso, recogido en un manipulador de eventos, que facilita un soporte nativo en dichos sistemas. El soporte ha sido realizado para dar compatibilidad al sistema existente de sincronización distribuida de la red de área local del TJ-II hacia máquinas UNIX, mediante clientes de sincronización, y está basado en la emulación de la familia de rutinas *WaitFor*, que es el mecanismo de sincronización básico en sistemas operativos Windows. El sistema de notificación asíncrona realizado es un soporte eficiente y válido, ampliamente probado y que asegura una auténtica comunicación interproceso apoyado en primitivas POSIX estándares que emulan la sincronización de Windows. Asimismo, se ha conseguido un manipulador de eventos JAVA apoyándonos en las primitivas de sincronización (*wait/notify* y *synchronized*. En sistemas de tiempo real existen mecanismos interproceso de eventos, los cuales se han aprovechado para conseguir implementar los clientes de sincronización. Diferentes clientes de sincronización fueron desarrollados para múltiples plataformas del TJ-II:

- UNIX (Sun-Solaris y hp-tru64)
- LINUX kernel 2.6.4 y posteriores (SuSe, Red Hat, 32 y 64 bits)
- JAVA (J2 sdk 1.4.2_07 y posteriores)
- OS-9 (M68k-VME)
- VxWorks (PowerPC-VME diskless, tarjetas MVME5500)

Métodos de búsqueda y reconocimiento de patrones en señales de evolución temporal han sido aplicados a bases de datos masivas en el campo de la fusión nuclear. Dicho reconocimiento se ha tratado de abordar en dos fases, la primera de ellas atiende a lo que se denomina forma de onda completa y consiste en encontrar las señales completas más parecidas a una de referencia. La segunda aproximación se refiere a formas estructurales contiguas dentro de la señal o patrones, consistente igualmente en indicar y en encontrar donde y en que señales se repiten esas similaridades para un patrón de referencia seleccionado por el usuario. Paralelamente a estas técnicas se ha implementado un entorno distribuido y abierto que complementa la recuperación y la disponibilidad de los datos y que puede ejecutarse en cualquier entorno de computación y sistema operativo. Además, el uso de componentes embebidos o embarcados (base de datos relacional, contenedor de *servlets*) favorece la simplicidad y la distribución de la herramienta sin necesidad de complejas configuraciones de software. Todo ello conforma una herramienta muy útil, fácil de instalar y ejecutable tanto en entornos locales como en entornos de participación remota y distribuida. Como valor añadido a la aplicación final, se ha incluido un instalador multiplataforma basado también en Java que permite instalar y seleccionar que módulos se incluirán para cada usuario final.

5.6 Síntesis de publicaciones

Las principales publicaciones que recogen los trabajos explicados en el presente capítulo son las siguientes:

Tema y aplicación principal	Aportación al trabajo	Publicación
Sincronización de la operación del TJ-II con procesos locales y remotos que se ejecutan en diferentes entornos de computación no basados en Windows	- Recursos de sincronización en sistemas UNIX, Linux, JAVA, OS-9 y VxWorks	[Pereira et al., 2006] [Pereira y Vega, 2005]
Clasificación automática de las imágenes generadas por el diagnóstico Scattering Thomson durante la operación pulsada del TJ-II	- Clientes de sincronización para entornos Sun-Solaris. - Aplicación principal de control para el reconocimiento de las imágenes. - Herramienta off-line de depuración de imágenes clasificadas erróneamente durante la operación automática.	[Vega et al., 2005] [Makili et al., 2010]
Monitorización automática del estado de los sistemas de adquisición de datos del TJ-II haciendo uso de una arquitectura orientada a mensajes mediante JMS	- Aplicación de sincronización para aplicaciones JAVA. - Proceso centralizado de control para almacenar el estado de los sistemas de adquisición.	[Sánchez et al., 2006] [Sánchez et al., 2006b] [Sánchez et al., 2007b] [Sánchez et al., 2008]
Sistema de participación remota del TJ-II	- Instalación y mantenimiento de los diferentes clientes de sincronización. - Seguimiento remoto de la operación de adquisición de datos del TJ-II.	[Vega et al., 2005b] [Vega et al., 2006]
Diseño e implementación de una arquitectura distribuida y abierta para la recuperación de datos y el reconocimiento de patrones dentro de señales	- Herramienta para el reconocimiento de patrones estructurales en señales mediante el desarrollo de aplicaciones visuales cliente/servidor. - Migración de las rutinas de compresión de datos hacia entornos JAVA.	[Pereira et al., 2010] [Vega et al., 2007b]

Tabla 5. 3. Síntesis de publicaciones capítulo 5

Capítulo 6

Reconocimiento morfológico en señales e imágenes del TJ-II y JET

La exploración visual de señales e imágenes es de especial relevancia para poder entender el comportamiento del plasma. La mayoría de los eventos físicos que suceden en el plasma quedan reflejados mediante formas de onda muy característicos. Patrones semejantes, a menudo, reflejan comportamientos físicos muy similares. Las técnicas de reconocimiento de patrones explicados en el capítulo 3 fueron instaladas en el TJ-II y en el JET para dar soporte en la búsqueda de fenomenología física concreta. En [Rattá et al., 2008] se utilizó la herramienta desarrollada en [Vega et al., 2008b] para un primer estudio sobre los cortes de amplitud en los canales de temperatura del JET. En segundo lugar se determinó el instante de la transición L/H mediante una múltiple búsqueda de patrones, combinando las formas morfológicas de las señales Dα y la densidad integrada de línea del JET. La aplicación de reconocimiento morfológico de patrones dentro de señales, al proveer resultados con gran velocidad, son metodologías particularmente apropiadas para fusión, donde las bases de datos guardan cantidades cada vez mayores de información. Proporcionan una solución rápida e intuitiva y además pueden ser extendidos a cualquier otra fenomenología cuyo comportamiento sea posible describir morfológicamente. La optimización que se hizo en [Pereira et al., 2010], [Pereira et al., 2010b], de las estrategias de búsqueda de patrones, permitieron mejorar la recuperación de formas de onda estructurales, independientemente de la longitud de las mismas y de la cantidad de señales almacenadas.

Se explica también en este capítulo la herramienta diseñada para la búsqueda y recuperación de patrones semejantes en imágenes del JET. Se adjuntan los resultados obtenidos para una base de datos compuesta por 25798 imágenes y 5.82 Gb de información almacenada.

6.1 Recuperación de señales y formas de onda

Los métodos de recuperación de formas de onda quedaron plasmados en la construcción de una herramienta software a modo de buscador, tanto de patrones, mediante formas estructurales específicas dentro de una señal, como de señales enteras, mediante formas de onda completa [Pereira et al., 2010]. Se ha pretendido desde el principio que, el producto fuera lo más abierto posible, que todos los componentes pudieran ejecutarse en un solo ordenador o bien que, tanto la aplicación cliente, la aplicación servidora y las bases de datos, pudieran residir en diferentes ordenadores y conectados por red.

Las tareas globales más importantes realizadas fueron:

- Desarrollo de la aplicación servidora:

  - Generación de señales en formato comprimido a partir de las señales originales brutas (dadas en formato ASCII).
  - Generación de la base de datos de señales clasificadas a partir de las señales comprimidas creadas anteriormente y mediante la aplicación de diferentes técnicas de minería de datos y de reconocimiento de patrones.
  - Creación de un servidor que atienda las peticiones de búsqueda de formas de onda completas y de patrones dentro de señales a petición de clientes remotos.

- Desarrollo de la aplicación cliente:

  - Interfaz de usuario gráfico que muestre una señal y poder realizar sobre ella ciertas operaciones.
  - Posibilidad de seleccionar un patrón con el ratón para posteriormente poder aplicar la acción de búsqueda de ese patrón o señal seleccionada.
  - Posibilidad de mostrar los resultados de las señales o patrones más parecidos a uno de referencia en la misma aplicación de visualización.

En general, los conjuntos de datos y señales con los que se va a tratar, partirán de un elevado número de observaciones y con una alta dimensionalidad. El tema relacionado con el descubrimiento de patrones tiene que ver con la aplicación de métodos de reducción de dimensionalidad como la transformada wavelet y un sistema de indexación-clasificación efectivo, apoyándonos en la potencia de un motor de base de datos relacional y técnicas de reconocimiento de patrones, para posteriormente aplicar una medida de similaridad y comparar como de similares son las señales encontradas. Se pretende reducir drásticamente la dimensionalidad de tal forma que, habiendo eliminado la información menos representativa, se manifiesten sin embargo, patrones característicos. También fue necesario implementar un conjunto amplio de métodos y

funciones gráficas con los que poder visualizar de manera sencilla el agrupamiento y estructura de los datos.

El diseño e implementación se enfocó dentro de una arquitectura cliente-servidor de tres capas, separando lo que es la capa de presentación, la lógica de negocio y la capa de manejo de datos.

A continuación se explica la herramienta desarrollada utilizando señales del TJ-II y explicando las posibilidades técnicas de la misma. Seguidamente se muestran resultados en la búsqueda de patrones morfológicos de la transición L/H. Se indexan en la base de datos el mismo número de señales utilizadas en [González et al., 2012]. En aquella ocasión se estudiaron las transiciones L/H, pero utilizando modelos orientados a datos y correlacionando diferentes señales para obtener el instante de tiempo de la transición L/H. La herramienta para la búsqueda de patrones morfológicos descrita en los párrafos siguientes, no hace uso de ninguna función de relación que modele o utilice los datos experimentales de otras señales ajenas. Exclusivamente se hace uso de los datos almacenados para las descargas de la misma señal. El concepto de similaridad recae en la distancia entre vectores de datos pertenecientes a la misma señal y no entre correlaciones de diferentes señales.

6.1.1 Posibilidades técnicas y científicas de la herramienta en su versión distribuíble

La primera tarea a realizar consiste en la indexación de señales en formato comprimido a partir de las señales originales brutas, dadas en formato ASCII. Esta tarea es necesaria por dos motivos principales, el primero es poder almacenar las muestras originales de las señales con el menor tamaño posible y el segundo motivo es el de poder transmitir la señal entera hacia un cliente remoto en el menor tiempo posible. Lógicamente si las señales están comprimidas se tardará menos tiempo en realizar esta operación.

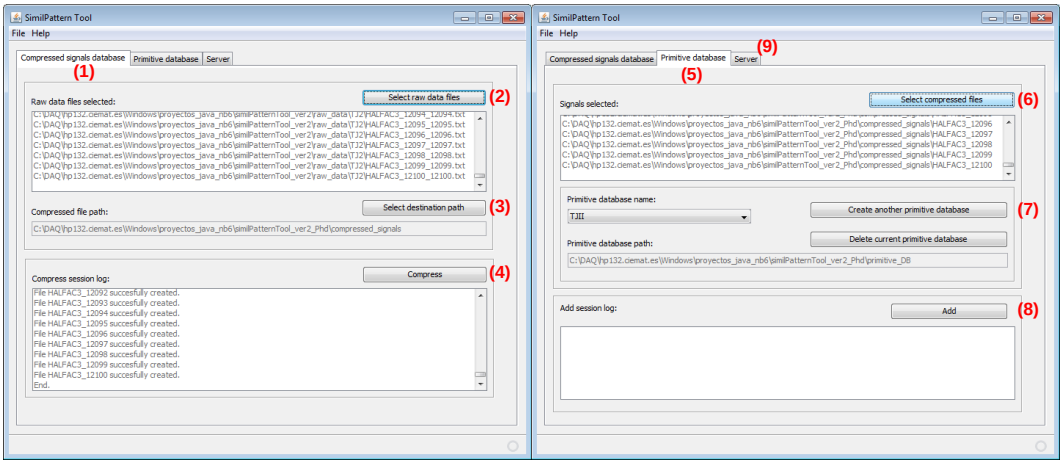


Figura 6. 1. Configuración de señales en la aplicación servidora

En la Figura 6. 1 se detalla el proceso a seguir en la indexación y preparación de las bases de datos. Desde la aplicación servidora y en su primera pestaña (1), un usuario selecciona los ficheros de entrada de las señales en formato ASCII (2), posteriormente selecciona una carpeta de destino (3) y procede a comprimir dichas señales (4), depositándolas en el directorio seleccionado. En la segunda pestaña de la aplicación (5), el usuario selecciona (6) las señales comprimidas generadas anteriormente. Puede crear y eliminar diferentes bases de datos de primitivas y en diferentes lugares (7) para posteriormente añadirlas (8) a la base de datos seleccionada. Finalmente en la tercera pestaña de la aplicación servidora (9), se puede inicializar el servidor para que espere por peticiones de los diferentes clientes remotos.

Los diferentes métodos de compresión utilizados [Vega et al., 2007b], permiten alcanzar tasas cercanas al 80%, comprimiendo notablemente el espacio almacenado por las señales brutas originales. La posterior reducción de datos, aplicada mediante las funciones wavelet-Haar, han permitido deshacerse de información que es irrelevante para la comparación entre señales y la búsqueda de patrones. Los coeficientes de aproximación wavelet son almacenados en una base de datos relacional. El motor de búsqueda relacional es el encargado de realizar posteriores búsquedas de una forma transparente y rápida. Básicamente, los pasos (1) a (8) explicados anteriormente, sirven para acondicionar y preparar los datos de entrada y poder almacenarlos en las tablas de la base de datos relacional, como se indica en la siguiente figura.

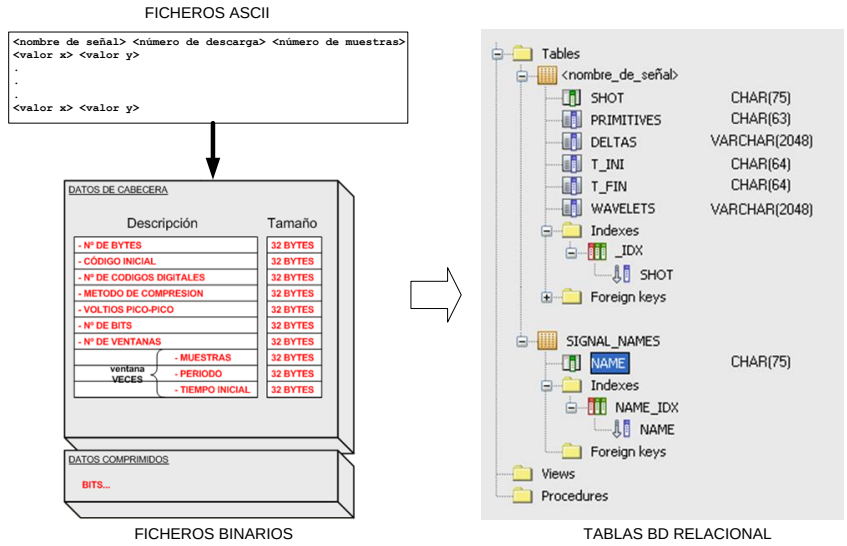


Figura 6. 2. Formato de los ficheros y tablas de la base de datos relacional

Finalmente, la aplicación servidora hace uso de un servidor web y un contenedor de *Servlets* escritos en Java. Se ha optado por utilizar el servidor JETTY, el cual se publica como un proyecto de software libre bajo la licencia Apache 2.0. Debido a su pequeño tamaño, JETTY se complementa para ofrecer servicios http en una aplicación Java empotrada. La ventaja principal de poder embarcar un servidor http en la aplicación servidora es la de evitar configuraciones e instalaciones de terceros componentes manteniendo las virtudes y los beneficios de trabajar con este protocolo en entornos de programación Java.

Paralelamente se ha desarrollado una aplicación cliente enteramente en Java cuya finalidad es servir de entorno gráfico para la visualización de las señales y la posterior

recuperación de los patrones. El usuario interacciona directamente en la aplicación estableciendo una comunicación y conexión con el servidor de datos. El servidor está localizado por un nombre de servidor DNS y por un puerto de escucha. Una vez establecida la conexión, el servidor le comunica al cliente de que información dispone en su dominio, esto es, bases de datos, señales en cada base de datos y números de descarga para cada señal.

Por medio de diferentes selectores de la aplicación cliente, se elige una señal de referencia y se peticiona al servidor. Todas las muestras de las señales enviadas por la aplicación servidora están en formato comprimido y posteriormente son descomprimidas en el lado cliente por la aplicación de visualización.

Una característica muy importante de esta aplicación es la implementación que se hizo para flexibilizar las consultas a la base de datos, Figura 6. 3. Esta utilidad permite encontrar multitud de patrones semejantes independientemente de la longitud de los mismos.

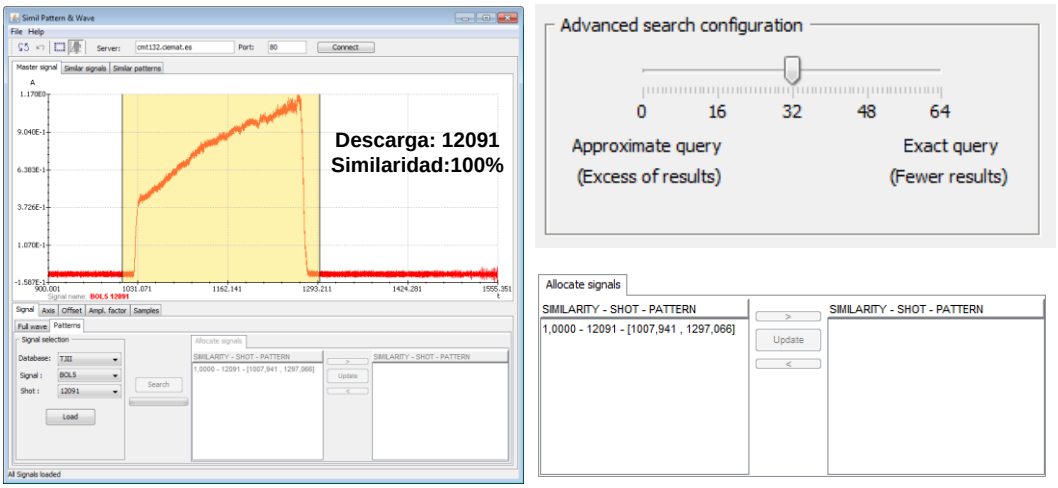


Figura 6. 3. Recuperación mediante consulta muy restrictiva

La aplicación cliente dispone de un configurador para poder ajustar las búsquedas. Una consulta muy estricta a la base de datos obtiene pocos resultados si nuestra base de datos es pequeña. En el ejemplo de la figura anterior existen 109 señales pertenecientes a la señal BOL5 del TJ-II. Realizando una consulta para una forma de onda muy alargada y sin apenas flexibilizar la búsqueda, ocasiona una muy pobre recuperación de patrones similares. Solamente se obtiene la señal y el patrón original de referencia. Si flexibilizamos la consulta y no la hacemos tan estricta se obtienen 38 recuperaciones muy similares al patrón de referencia. En la imagen de la Figura 6. 4 se pueden observar diferentes recuperaciones junto con sus valores de similaridad respecto al patrón original. El modo de funcionamiento de este configurador se explica detalladamente en el capítulo 3.

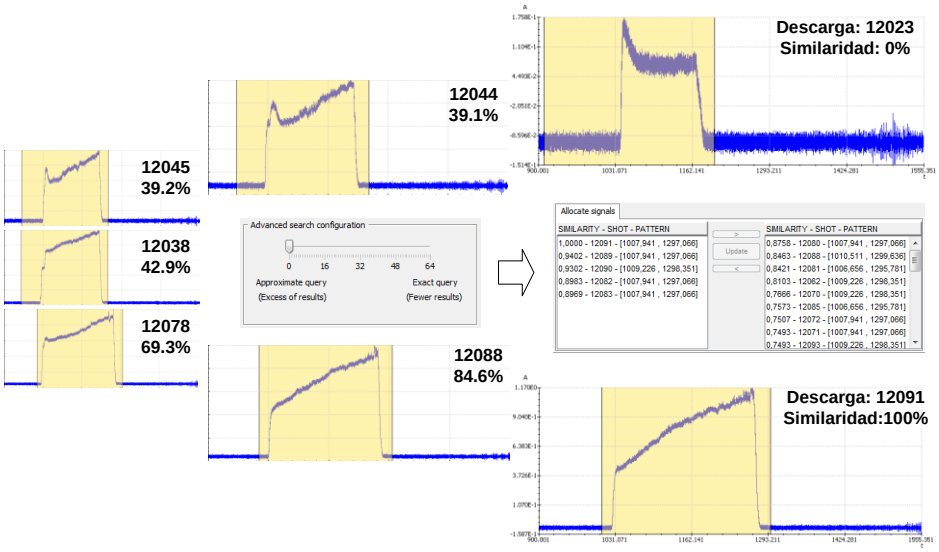


Figura 6. 4. Recuperación mediante consulta más flexible

Mediante la flexibilización de esta estrategia de búsqueda, el número de patrones similares a recuperar se hace independiente tanto de la longitud del patrón de referencia como de la cantidad de señales indexadas en la base de datos. Pudiendo ajustar la cantidad y calidad de las recuperaciones al requerimiento que más nos interese.

Otra de las grandes posibilidades de la herramienta es la búsqueda de señales completas, esto es, la forma de onda más larga posible. Además, la métrica utilizada para comparar la similitud entre vectores de datos, cuando se trata de señales completas, permite no hacer distinción en la polaridad de la señal. Señales muy diferentes, pero simétricas respecto del eje de abscisas son recuperadas con similaridad pareja e independientemente de los límites de interpolación de la señal, Figura 6. 5.

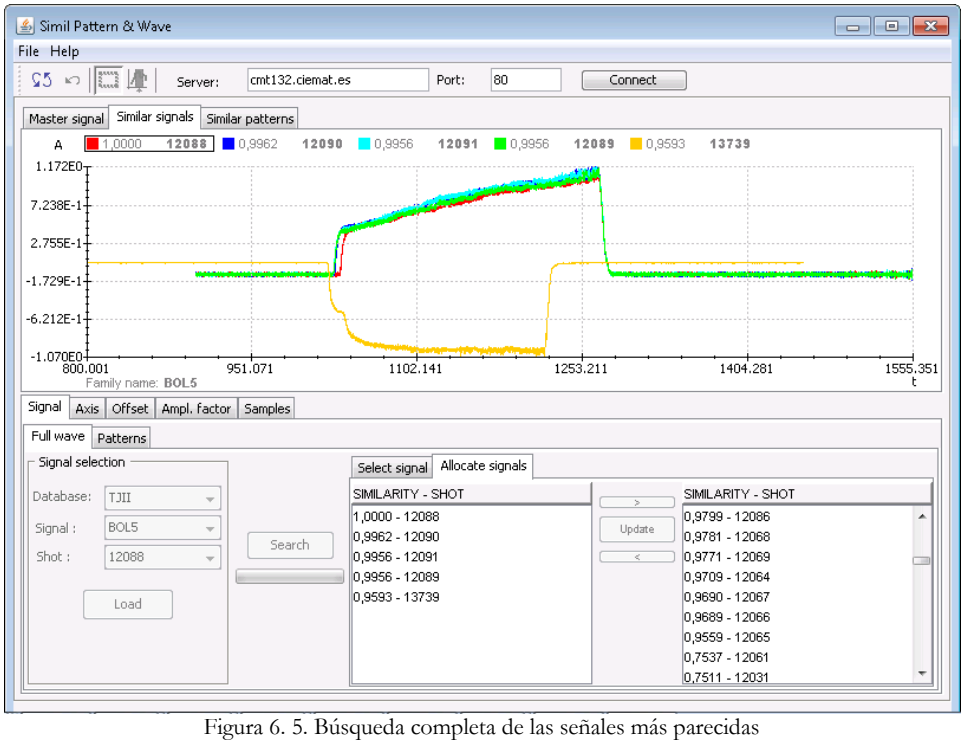


Figura 6. 5. Búsqueda completa de las señales más parecidas

6.1.2 Aplicación práctica en la búsqueda de patrones que identifica la transición L/H

El análisis exploratorio y visual de las señales es fundamental antes de emprender cualquier tipo de acción sobre los datos. El estudio de la evolución temporal y gráfica de cada señal individualmente, permite comprender el comportamiento de las mismas, ante la constatación de cierta fenomenología física. De esta forma, el analista consigue un primer entendimiento básico de los datos y de las relaciones existentes entre las variables visualizadas. Muy frecuentemente, los patrones de comportamiento repetitivos que evidencian los datos, no son lo suficientemente uniformes, tanto espacialmente como temporalmente. Llegados a este punto, las herramientas de visualización de señales se hacen imprescindibles, en una primera aproximación, para poder identificar dichos comportamientos característicos. La transición L/H suele identificarse visualmente en la mayoría de los casos mediante una inspección minuciosa de esta señal. La señal suele evidenciar la transición mediante una caída en su amplitud en un periodo de tiempo de pocos milisegundos (Figura 6. 6). Sin embargo, esta rápida disminución de la magnitud de la señal no siempre es notoria y en muchos casos puede confundirse con el mismo ruido que afecta la señal (Figura 6. 6, descarga 76199).

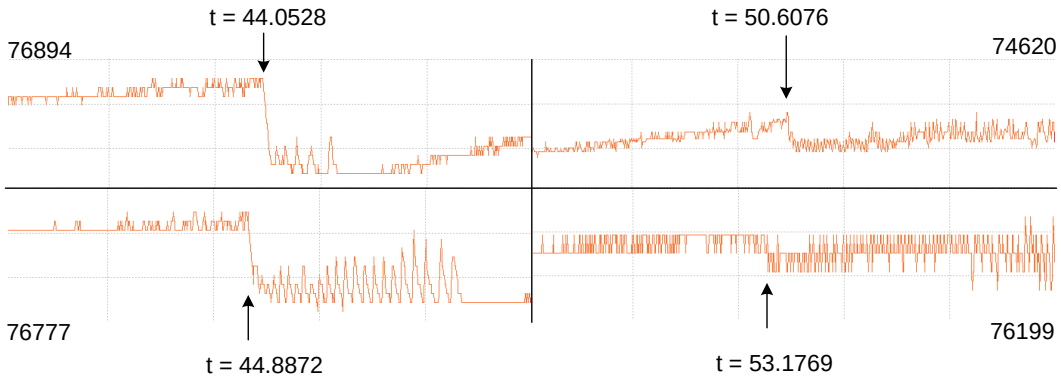


Figura 6. 6. Diferentes manifestaciones gráficas de la transición L/H

Antes de comenzar los trabajos expuestos en el apartado 4.2, se realizaron varios análisis gráficos del instante de la transición L/H en la señal Dα (AD36) del JET mediante la herramienta de recuperación de patrones y alrededor de la transición L/H. El procedimiento llevado a cabo consistía en seleccionar señales con patrones bien definidos en el instante de la transición L/H y examinar las recuperaciones aportadas. La herramienta de recuperación de patrones permite realizar a priori diferentes configuraciones. Una configuración de búsqueda muy restrictiva obtiene pocos resultados, pero éstos son muy exactos o coincidentes con el patrón de referencia. Una búsqueda menos restrictiva encuentra prácticamente patrones coincidentes en todas las señales, pero también incrementa el valor de los falsos positivos visuales. Estas recuperaciones equivocadas son debidas a que el patrón que caracteriza la transición L/H no siempre aparece con la misma representación morfológica. No obstante, la flexibilidad de la herramienta desarrollada para la búsqueda de patrones permite realizar múltiples

recuperaciones a partir de diferentes formas de onda de partida que caracteriza la transición L/H, tanto en su duración temporal, como en sus valores de amplitud correspondientes.

Se indexaron en la base de datos 528 descargas pertenecientes a la señal AD36 del JET. La señal AD36 mide la radiación emitida por el plasma en contacto con la cara interior del divertor del JET. Un divertor es un componente de una máquina toroidal de fusión cuya misión es dirigir mediante campos magnéticos las partículas del borde del plasma a una cámara separada donde chocan con unas placas y son neutralizadas. Cuando se pasa de modo de confinamiento L a un modo mejorado H, la señal AD36 recoge una caída brusca de amplitud en su evolución temporal. En los dos ejemplos de la Figura 6. 7, se muestras las búsquedas de dos patrones característicos de la transición L/H para la descargas 76843 y 76929. Mediante la herramienta software disponible, realizamos una configuración de forma que nuestra primera consulta de búsqueda sea muy restrictiva. Un valor correspondiente a la desviación de los valores de amplitud de todas las señales almacenadas dividido por 64. Este umbral, como se ha explicado en el capítulo 3, equivale a la frontera donde se determina para cada primitiva si puede tomar el valor indiferente '?', bien sea la primitiva 'a' o la 'b', que corresponden a las pendientes positivas o negativas de sus correspondientes segmentos.

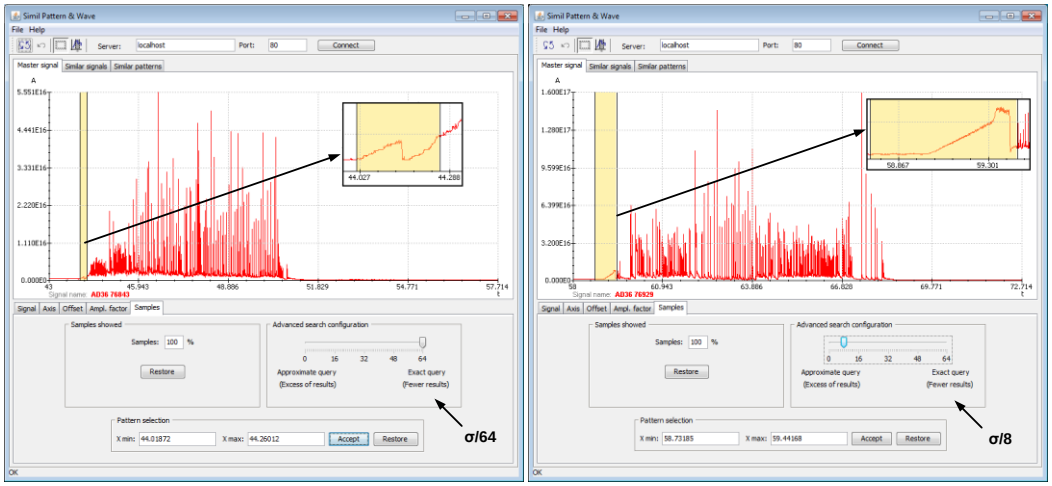


Figura 6. 7. Dos formas de onda diferentes para buscar patrones L/H coincidentes

La búsqueda de dicho patrón para la descarga 76843, obtiene 143 recuperaciones con patrones muy semejantes al de referencia. De este subconjunto, 83 de ellas son coincidentes con el instante de la transición L/H supervisada por el experto y en las 60 restantes, el instante de la transición es no coincidente dentro de los intervalos recuperados (Figura 6. 8).

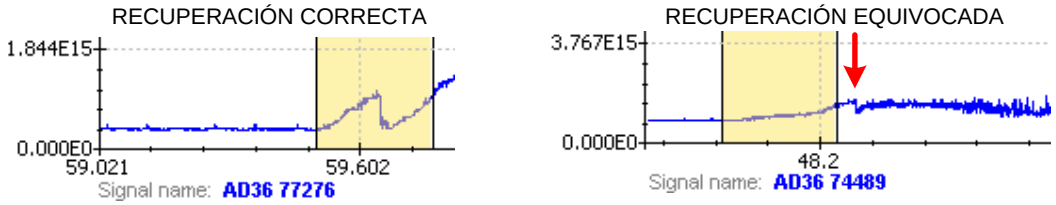


Figura 6. 8. Ejemplo de recuperación correcta y de recuperación no coincidente

Una búsqueda muy restrictiva en el segundo ejemplo (descarga 76929), solamente obtiene dos recuperaciones coincidentes. Esto es debido a la dificultad de encontrar

coincidencias muy exactas para un patrón de longitud más largo y con solamente 528 señales indexadas en la base de datos. Dificultad de búsqueda que se acrecienta a medida que aumenta la longitud del patrón de referencia. Cuanto mayor sea la longitud del patrón a buscar y cuanto menor sea la información almacenada en la base de datos, menores recuperaciones de patrones coincidentes se van a obtener. No obstante, si intentamos realizar la búsqueda del mismo patrón pero esta vez flexibilizando la búsqueda a un valor *σ/8*, conseguimos obtener 329 coincidencias para un total de 528 recuperaciones.

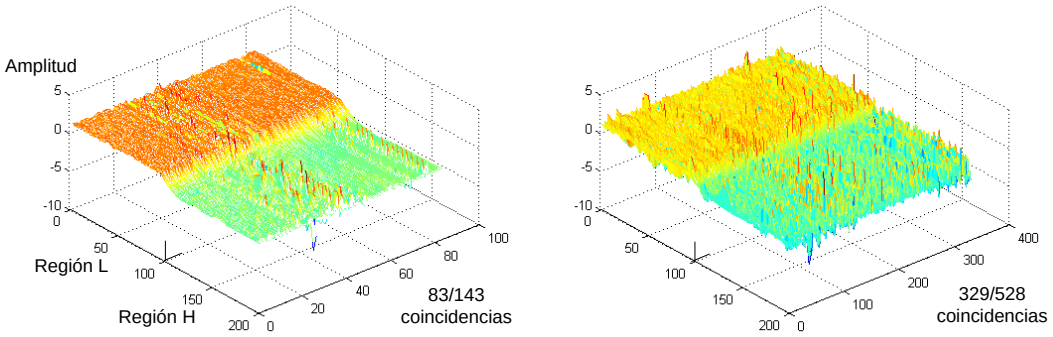


Figura 6. 9. Recuperaciones coincidentes con los dos patrones de búsqueda

En la Figura 6. 9 se pueden observar gráficamente todas las recuperaciones coincidentes para las dos búsquedas realizadas. Para una mejor distinción de las imágenes, se ha representado para cada patrón ±100 muestras alrededor de la transición L/H, normalizando las amplitudes de la señal con una resolución temporal de 0.1 ms. Comparando las dos figuras, se puede observar como las amplitudes en los patrones de la imagen de la izquierda son más homogéneas, existiendo menos uniformidad en los valores de las amplitudes para la gráfica de la derecha. En dicha gráfica, aparecen más saltos o picos en los valores de la amplitud a ambos lados del instante de la transición L/H, tanto en la región L como en la región H. Esas discontinuidades son las causantes de que no se puedan encontrar patrones muy coincidentes para configuraciones de búsqueda muy restrictivas en los valores de las primitivas almacenadas en la base de datos.

Con estos ejemplos, se muestra la utilidad de las diferentes estrategias de configuración implementadas para la búsqueda y recuperación de patrones gráficos en señales de evolución temporal. No obstante, mediante la observación visual del patrón característico de la transición L/H se llegó a la conclusión de que se hace todavía imprecisa la correcta determinación del instante de la transición, debido a la obtención de falsos positivos visuales. Inevitablemente es necesario añadir más información al análisis, para aumentar las tasas de acierto y poder alcanzar niveles de precisión más elevados en la identificación de dichas transiciones. Por este motivo, en el apartado 4.2 se presentan métodos orientados a datos que interrelacionan varias variables para la determinación del instante de tiempo de la transición L/H. Por ejemplo, se añadió el análisis detallado de la señal de la densidad integrada de línea de JET (LAD4), buscando dentro de ella pendientes positivas. Esto se debe a que un incremento en la densidad suele ser una consecuencia también de la transición, al mejorarse el confinamiento, menos partículas se pierden y la densidad del plasma aumenta.

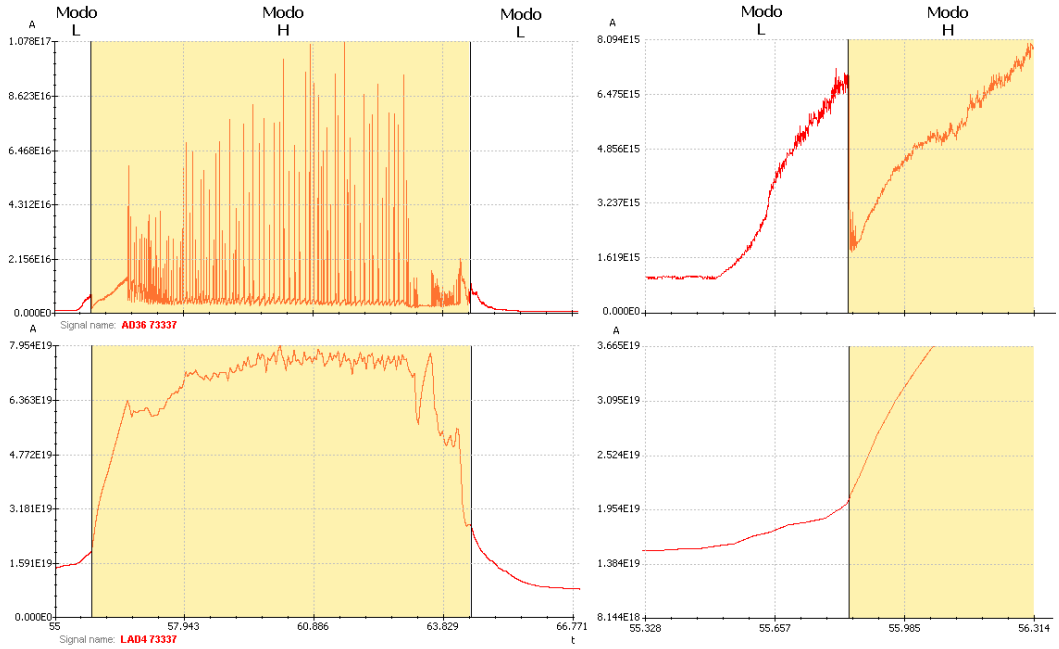


Figura 6. 10. Aumento de la densidad en el instante de la transición L/H

En la Figura 6. 10 se puede observar el incremento de la densidad que se origina en el instante de la transición L/H. La señal LAD4 del JET aporta información relevante para poder utilizarla en la determinación del instante de la transición. Esta señal ha resultado ser una de las más importantes para la predicción del instante de la transición L/H con elevadas tasas de acierto (ver apartado 4.2.1).

6.2 Recuperación de patrones gráficos semejantes en imágenes

En el trabajo [Vega et al., 2008b] se exponen los resultados obtenidos en la indexación, búsqueda y recuperación de diferentes patrones de imágenes pertenecientes a 16 video-películas de la cámara del espectro visible del JET. Con un almacenamiento bruto cercano a los 6 Gb de datos y un total de 25798 imágenes. La indexación de todas esas imágenes se realiza mediante la ejecución de varios procesos que preparan los datos convenientemente antes de su volcado a la base de datos relacional. Como se ha visto en apartados anteriores, los datos brutos son reducidos mediante la aplicación de la transformada 2D-wavelet-Haar al nivel x de descomposición, posteriormente se umbraliza cada una de las imágenes para eliminar ruido residual de fondo y las intensidades resultantes se discretizan en primitivas. Las pruebas realizadas consistieron en una categorización de intensidades a 4 y 2 primitivas. Finalmente, un último proceso introducirá las primitivas en la base de datos relacional.

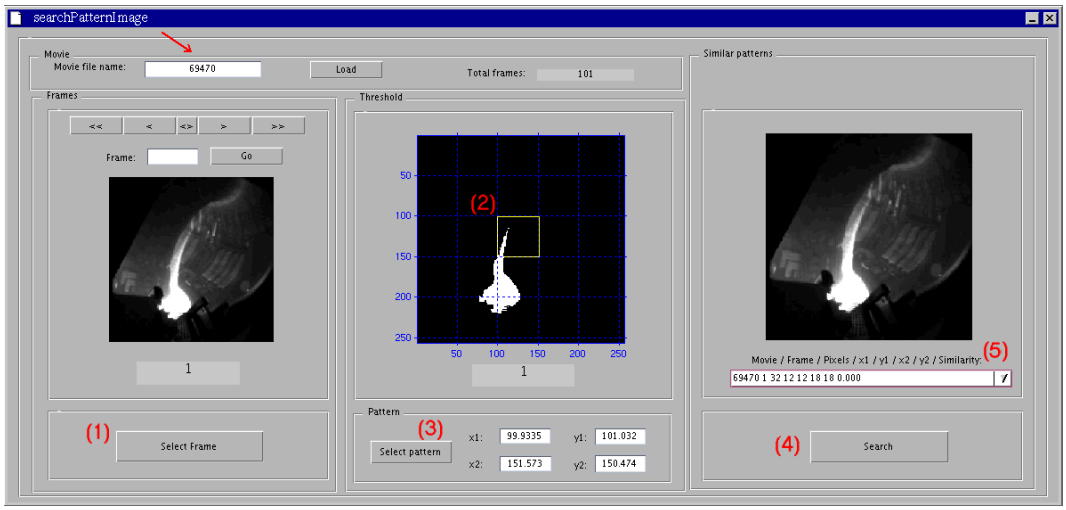


Figura 6. 11. Aplicación de usuario gráfica para la búsqueda de patrones en imágenes

Mediante una interfaz de usuario gráfica realizada en Matlab (Figura 6. 11) e instalada en el JAC-Linux Cluster del JET, un usuario puede interactuar con las películas y las imágenes, escoger el número de ventana de una película para poder visualizarlas con el umbral correspondiente (1), seleccionar un patrón gráfico interactuando con el puntero del ratón sobre la imagen (2), extraer las cotas de dicho patrón gráfico (3) para poder enviárselas (4) a la base de datos con el fin de poder buscar en toda la base la repetición de ese patrón para finalmente, una vez recuperadas dichas imágenes, poder visualizarlas con un orden de similaridad (5).

Con esta herramienta se realizaron diferentes pruebas tanto con bases de datos indexadas para un tamaño final de 32x32 pixeles como para imágenes de 16x16. En el ejemplo de la figura se pude observar la búsqueda de un patrón en la imagen número 1 de la película 69903, la cual está formada por un total de 999 imágenes. La búsqueda es para la base de datos indexada mediante imágenes de 16x16 y con solamente 2 primitivas. Se recuperan un total de 67 imágenes similares con el mismo patrón. Se puede apreciar cómo no solamente se encuentran patrones similares pertenecientes a la misma película sino que aparecen esos mismos patrones en diferentes películas, en la Figura 6. 12 se indica la imagen número 38 de la película 69470 con sus cotas y un valor de similaridad respecto al patrón de búsqueda.

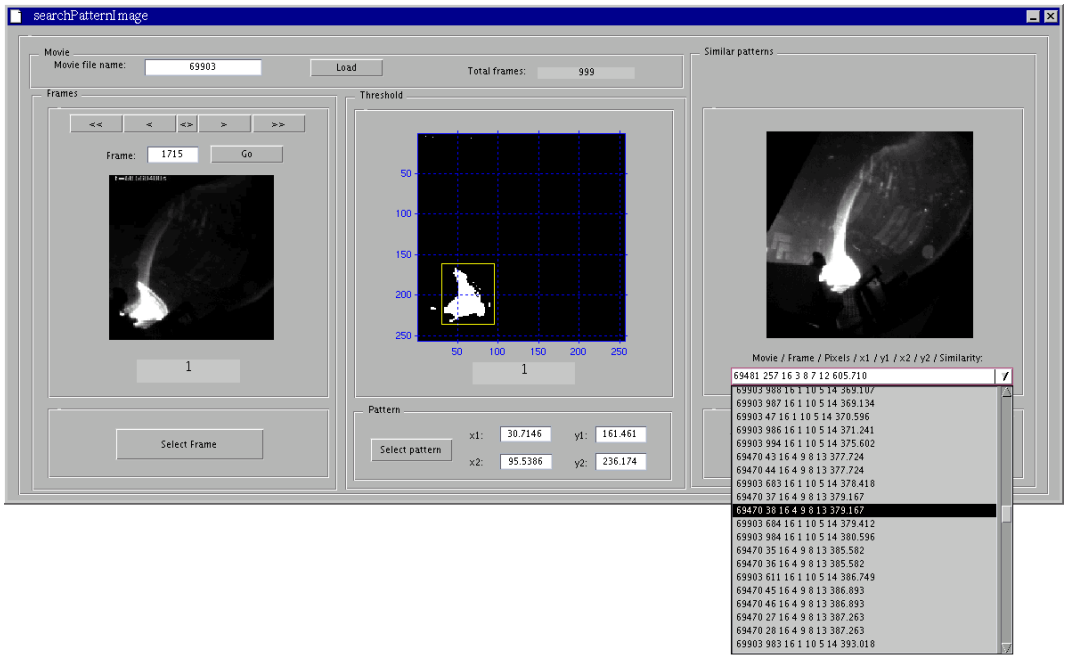


Figura 6. 12. Recuperación de múltiples patrones gráficos en imágenes

Las diferentes pruebas realizadas en un ordenador del JAC Linux Cluster del JET (Pentium-Pro 200 MHz, 1 procesador, 1GB RAM) son las siguientes:

INDEXING			SEARCH		RETRIEVAL	
MBytes	Movie	n Frames	32 x 32			
147,48	68842.avi	626		Pattern Length 32x32	32x32 4 primitives	32x32 2 primitives
432,77	69331.avi	1839				
26,62	69470.avi	101	69903, Frame 1	12x12	1 F, 10 s	1 F, 304 s
156,58	69481.avi	599	69331, Frame 1707	15x18	2 F, 112 s	2 F, 625 s
541,20	69734.avi	2300	69331, Frame 1715	8x11	2 F, 197 s	2 F, 1035 s
91,35	69787.avi	920	69470, Frame 50	13x20	2 F, 134 s	2 F, 561 s
500,04	69904.avi	2125	69904, Frame 1200	17x9	2 F, 37 s	60 F, 69 s
499,81	69905.avi	2124				
364,33	69925.avi	1548				
552,73	69927.avi	2349				
576,25	69932.avi	2449	16 x 16			
552,73	69933.avi	2349				
234,74	69995.avi	997		Pattern Length 16x16	16x16 4 primitives	16x16 2 primitives
527,79	69997.avi	2243				
524,74	69998.avi	2230	69903, Frame 1	6x7	4 F, 7 s	67 F, 619 s
235,21	69903.avi	999	69331, Frame 1707	6x9	2 F, 11 s	2 F, 465 s
			69331, Frame 1715	5x5	2 F, 47 s	446 F, 485 s
Gigas	Total Movies	Total Frames	69470, Frame 50	7x11	22 F, 81 s	46 F, 407 s
5,82	16	25798	69904, Frame 1200	9x5	246 F, 199 s	3248 F, 378 s

Tabla 6. 1. Resultados de patrones en imágenes del JET

Se puede apreciar en dicha tabla también los tiempos empleados en la búsqueda de cada patrón.

6.3 Conclusiones

Dos aplicaciones de usuario gráficas fueron implementadas para el reconocimiento morfológico de señales e imágenes. Dichas aplicaciones hacen uso de los métodos y funciones explicadas en el capítulo 3 de la presente tesis. Han sido instaladas satisfactoriamente en ordenadores del JET y del TJ-II para la inspección visual de información perteneciente a dichos dispositivos. En la versión definitiva para la aplicación de búsqueda de señales, se hace uso de software que utiliza marcos embebidos para entornos informáticos remotos. El uso de este enfoque, que está destinado a proporcionar una mayor flexibilidad en un entorno más abierto, evita instalaciones de software complejas, como son los sistemas de gestión de base de datos y los servidores web. Esto permite encapsular todo el software necesario en una sola aplicación de escritorio. Con ello se consigue no solo ocultar estos módulos sino también otros detalles relacionados con la configuración de los mismos, tales como, el poder impedir el acceso a opciones de configuración no autorizadas, la configuración de argumentos incorrectos, etc.

Tanto la visualización como la búsqueda de patrones en señales han servido para poder inspeccionar visualmente fenomenología física del plasma, tanto del JET como del TJ-II.

6.4 Síntesis de publicaciones

CONJUNTO	MÉTODO	DATOS	CLASIFICADOR	MEDIDA SIMILARIDAD	PUBLICACIÓN
Señales TJ-II	Señal completa	Wavelet Haar	Cluster multi-capa	Producto escalar normalizado	[Vega et al., 2008]
Señales TJ-II	Patrones en señales	Discretización primitivas longitud constante (5 umbrales)	Base de datos relacional	Error del ajuste	[Dormido-Canto et al., 2006]
Señales JET	Señal completa	Wavelet Haar	Cluster multi-capa	Producto escalar normalizado	[Vega et al., 2007] [Vega et al., 2008c] [Dormido-Canto et al., 2008]
Señales JET	Patrones en señales	Discretización primitivas longitud adaptable (varios umbrales)	Base de datos relacional	Media del producto escalar normalizado	
Señales genéricas	Señal completa	Discretización primitivas longitud constante (2 umbrales)	Base de datos relacional	Distancia de Hamming	[Pereira, 2009]
Señales genéricas	Señal completa	Wavelet Haar	Base de datos relacional	Producto escalar normalizado	[Pereira et al., 2010b] [Pereira et al., 2010]
Señales genéricas	Patrones en señales	Discretización primitivas longitud constante (2 umbrales)	Base de datos relacional	Distancia euclídea	

Tabla 6. 2. Síntesis de publicaciones para la búsqueda de señales

CONJUNTO	MÉTODO	DATOS	CLASIFICADOR	MEDIDA SIMILARIDAD	PUBLICACIÓN
Imágenes Thomson TJ-II	Imagen completa	2D Wavelet Haar	SVM	-	[Vega et al., 2005] [Makili et al., 2010]
Imágenes Thomson TJ-II	Imagen completa	2D Wavelet Haar	Vecino más próximo (conformal)	-	[Vega et al., 2010]
Imágenes JET	Imagen completa	2D Wavelet Haar	Base de datos relacional	Distancia euclídea	[Vega et al., 2008b] [Vega et al., 2009]
Imágenes JET	Patrones en imágenes	2D Wavelet Haar	Base de datos relacional	Distancia euclídea	
Imágenes Thomson TJ-II	Imagen completa	2D Wavelet Haar	SVM (conformal)	-	[González et al., 2012b]
Imágenes JET	Patrones en imágenes	2D Wavelet Haar	Base de datos relacional distribuida	Distancia euclídea	[Vico, 2010]

Tabla 6. 3. Sístesis de publicaciones para la búsqueda de imágenes

Capítulo 7

Selección de características para la predicción de disrupciones del JET

En el trabajo realizado en [Vega et al., 2014], un clasificador probabilístico, basado en los predictores Venn, fue implementado en el JET para ser usado como predictor de disrupciones mediante una aproximación empezando desde el principio ("*from scratch*", en inglés). Fue aplicado a una base de datos de 1237 descargas, pertenecientes a las campañas C28 a C30 del JET, las cuales, 201 descargas eran de tipo disruptivo y el resto, 1036 eran descargas no disruptivas. Se alcanzaron tasas de acierto en las descargas disruptivas del 94% y con un porcentaje de falsas alarmas en las no disruptivas del 4.21%. Para seleccionar las mejores características de entre un conjunto inicial de partida, formado por un total de 14 señales, se utilizó un análisis combinatorio completo no exhaustivo (ver apartado 4.1.2). Se analizaron todas las posibles combinaciones entre 2 y 7 características. En total fueron evaluados 9893 predictores diferentes, lo que supuso un tiempo de cómputo total de 1731 horas, equivalente a casi dos meses y medio de cálculos para completar todas esas combinaciones. Con el objetivo de alcanzar estas buenas tasas de acierto pero reduciendo todo lo posible el tiempo de cálculo al mínimo, se investigó la posibilidad de utilizar los algoritmos evolutivos explicados en el apartado 4.1.4. Cinco métricas diferentes de evaluación fueron utilizadas (ver apartado 4.4.1) como funciones de ajuste del algoritmo genético empezando con el mismo conjunto de individuos aleatorios y la misma generación de partida. Los mejores resultados, consistentes en encontrar las mejores características con las mejores tasas de acierto, fueron conseguidos con la función de evaluación llamada *Informedness* (la diferencia entre tasas de acierto y falsas alarmas), con unos cómputos finales de evaluación de solamente 168 predictores en un tiempo de 29.4 horas (poco más de un día).

7.1 Antecedentes

Como se ha explicado en esta tesis, la disrupción del plasma en un tokamak es una pérdida repentina de la corriente del plasma y de su confinamiento. Ello puede acarrear un peligro para los componentes y accesorios mecánicos del dispositivo, debido a la generación de fuertes cargas producidas por la expansión abrupta de la energía contenida en el interior del plasma y dirigida hacia el exterior del dispositivo. Detectar e impedir eventos disruptivos con antelación es extremadamente importante en dispositivos experimentales de tipo tokamak. En el JET, diferentes algoritmos de aprendizaje se han utilizado para implementar sistemas automáticos que son capaces de predecir disrupciones con antelación. El objetivo principal de un predictor de disrupciones es detectar con suficiente tiempo de antelación el comportamiento del plasma como disruptivo durante el transcurso de cualquier descarga en operación. El predictor de disrupciones avanzado APODIS fue desarrollado mediante un modelo apoyado en los datos y utilizando la combinación de tres clasificadores SVM. Fue instalado satisfactoriamente en la red de tiempo real del JET [López et al., 2012] y obtuvo muy buenos resultados en las primeras campañas de la pared metálica del JET [Vega et al., 2013d]. Una actualización del sistema APODIS que se hizo para predicciones empezando y entrenando con muy pocas descargas 'desde cero', obtuvo también muy buenas tasas de acierto [Dormido-Canto et al., 2013]. La terminología 'desde cero', implica que existe un déficit de información durante los primeros procesos de aprendizaje y el predictor tiene que aprender desde el principio sin casi ningún tipo de conocimiento acerca de cómo poder discernir que es disruptivo y que no lo es. Esto ocurre cuando una máquina experimental comienza su operación después de su construcción o su actualización, de gran importancia pensando en el dispositivo ITER. Continuando con el paradigma 'desde cero', un predictor probabilístico basado en clasificadores Venn fue desarrollado en el JET [Vega et al., 2014]. Los predictores Venn facilitan información probabilística añadida, además del tipo de clase a la que pertenecen las observaciones a predecir. En contraposición a SVM, que solamente informa del tipo de clase en la predicción. Además, en la implementación llevada a cabo en dicho trabajo fue posible reducir el espacio de entrada de todas las observaciones, tanto para el entrenamiento como para la posterior predicción, utilizando agrupaciones de datos basadas en una taxonomía del centroide más próximo. Esta característica mejoró notablemente la rapidez del predictor frente a nuevas descargas. Como se ha comentado anteriormente, tasas de acierto del 94% y falsas alarmas del 4.21% (ver Figura 7. 1) fueron los resultados conseguidos, para una base de datos formada por un total de 1237 descargas pertenecientes a la pared metálica del JET.

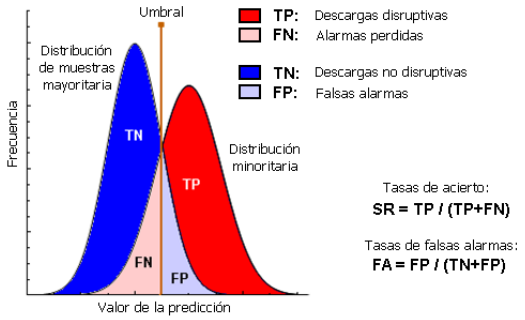


Figura 7. 1. Elementos para describir las tasas de acierto

Para su implementación, se utilizaron 7 señales diferentes del JET, con el objetivo de poder caracterizar y extraer la información disruptiva y no disruptiva. Estas señales se procesan secuencialmente utilizando ventanas temporales de 32 ms de tamaño. Para cada ventana temporal, se obtiene tanto la media de las muestras como la desviación estándar del espectro de Fourier, eliminando la componente continua de la misma. Procediendo de esta manera, se obtienen dos características por cada señal, lo que equivale a un total de 14 características diferentes (Tabla 7. 1). Para asegurarse la selección y la importancia de los mejores atributos de entre estas 14 características, se realizó un análisis combinatorio completo no exhaustivo. Un total de 9893 predictores diferentes fueron analizados, lo que conllevó un tiempo total empleado de casi dos meses y medio. Este proceso tan costoso fue el principal incoveniente del trabajo realizado, pero por otro lado, se aseguraba poder encontrar las características más valiosas para poder determinar un predictor con las mejores tasas de acierto.

Signal name	Label	Units	Id	Definition
Plasma current	Ip	A	1	mean(Ip)
			2	std(|fft(Ip)|)
Mode locked amplitude	ML	T	3	mean(ML)
			4	std(|fft(ML)|)
Plasma internal inductance	LI		5	mean(LI)
			6	std(|fft(LI)|)
Plasma density	Ne	m^3	7	mean(Ne)
			8	std(|fft(Ne)|)
Diamagnetic energy derivative	dW/dt	W	9	mean(dW/dt)
			10	std(|fft(dW/dt)|)
Radiated power	Pout	W	11	mean(Pout)
			12	std(|fft(Pout)|)
Total input power	Pin	W	13	mean(Pin)
			14	std(|fft(Pin)|)

Feature Id.														SR	FA
1	2	3	4	5	6	7	8	9	10	11	12	13	14	%	%
		x	x											94.00	4.70
x			x											92.50	5.09
x	x	x												94.00	4.31
		x	x							x				94.00	4.70
x	x	x	x											94.00	4.21
x	x	x								x				94.00	4.21
x	x	x	x							x				94.00	4.21
x	x	x				x	x							94.00	4.21
x	x	x				x	x			x				94.00	4.21
x	x	x	x							x	x			94.00	4.31
x	x	x	x			x	x			x				94.00	4.31
x	x	x				x	x			x	x			94.00	4.31

Tabla 7. 1. Lista de señales y tasas de acierto

Con el objetivo de poder reproducir estas tasas de acierto pero reduciendo el tiempo empleado en ello, se investigó en [Pereira et al., 2014] un método de selección de características basado en algoritmos genéticos junto con los predictores Venn. Diferentes métricas de evaluación para el predictor fueron utilizadas [Powers, 2011], demostrando que es realmente muy significante la selección correcta de la métrica elegida para poder obtener resultados rápidos y precisos a la vez.

7.2 Técnica combinada mediante algoritmos genéticos y predictores Venn

El proceso de selección de características ha consistido en la extracción de las características más importantes de entre todo el conjunto de partida. En el caso que nos ocupa las 14 señales. Se tienen que eliminar todas las redundantes y las que sean irrelevantes. Los atributos redundantes son aquellos que no aportan más información relevante al modelo de entre las que se consideren en cada momento y los atributos irrelevantes no aportan información útil en ningún contexto. Ambos atributos, si se incluyeran al modelo, tendrían efectos negativos a la hora de realizar las predicciones, incrementando los tiempos de cálculo y reduciendo la precisión y las tasas de aciertos. Los AG han demostrado su validez para identificar variables importantes y poder seleccionar características. Por otro lado, para encontrar los mejores atributos de todo el conjunto, la única posibilidad segura es probar haciendo combinaciones sin repetición hasta agotar todo el espacio muestral disponible. Esto requiere mucho tiempo y un coste computacional muy elevado. Se presenta a continuación un método ágil para extraer las características más importantes utilizando los AG.

El algoritmo empieza con un conjunto inicial de soluciones aleatorias llamado población. Una población aleatoria de 28 individuos es generada. Cada individuo se conoce también como cromosoma y está formado por 14 genes (o características). La calidad de cada cromosoma es estimada por un clasificador. En esta ocasión se utiliza un clasificador probabilístico basado en los predictores Venn. Los predictores Venn realizan predicciones directamente desde los datos mediante transducción, sin ninguna generación de regla o modelo previo, en vez de repetidamente estar entrenando clasificadores para generar modelos. Previamente a esto, el espacio de muestras o de observaciones de entrada (toda la información disruptiva y no disruptiva) se condensa siguiendo el criterio del centroide más próximo para cada clase. Resultando así, de la obtención de un clasificador más rápido y ágil para testear y probar todos los individuos que integran la población. En cada generación o iteración del AG, los 28 cromosomas son evaludados usando la función de ajuste (Figura 7. 2). Esta función juega el papel más importante en la búsqueda genética ya que la descendencia futura para la siguiente generación se determina mediante la puntuación obtenida por dicha función, que refleja como de óptima es la solución. Esta función tiene que evaluar la bondad de las tasas de error numéricas generadas por el clasificador. De esta manera, la salida del clasificador se convierte en la entrada de la función de ajuste y ésta obtiene una evaluación numérica analizando la bondad del subconjunto de características.

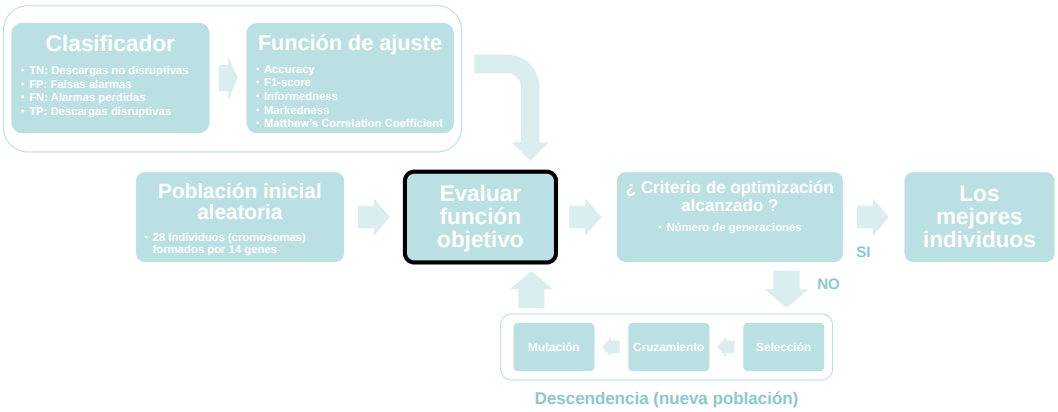


Figura 7. 2. Descripción detallada del algoritmo genético

Cinco métricas de rendimiento, explicadas en el apartado 4.4.1, fueron utilizadas como medidas en la función de ajuste del AG utilizado (Figura 7. 3). Los mejores individuos con las mejores puntuaciones son seleccionados para crear la siguiente generación.

$$- \boldsymbol{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$
$$- \boldsymbol{F1\text{-}score} = \frac{2TP}{2TP + FP + FN}$$
$$- \boldsymbol{Informedness} = sensitivity + specificity - 1 = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1$$
$$- \boldsymbol{Markedness} = precision + NPV - 1 = \frac{TP}{TP+FP} + \frac{TN}{TN+FN} - 1$$
$$- \boldsymbol{MCC} = {(TP * TN - FP * FN)} \Big/ {\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Figura 7. 3. Métricas utilizadas para la función de ajuste

Finalmente, dos operadores genéticos tales como el cruzamiento y la mutación explorarán nuevas regiones del espacio de búsqueda mediante la combinación y el reemplazamiento de los genes mientras se conserva al mismo tiempo la mayoría de la información de los individuos padres. El proceso generacional finaliza cuando el criterio de parada se satisface, por ejemplo, en nuestro caso se ha optado por realizar un total de 18 generaciones para la comparación de las cinco métricas de la función de ajuste. Las características ganadoras, esto es las más importantes, se corresponderán con los mejores individuos encontrados en todas las generaciones realizadas.

La lista de señales utilizadas se explica en la Tabla 7. 1, que coinciden y son las mismas que las utilizadas en el trabajo [Vega et al., 2014], integradas por 14 características pertenecientes a 7 señales del plasma del JET. Los mejores resultados alcanzados fueron igualmente del 94% y del 4.21% en términos de tasas de acierto y falsas alarmas respectivamente. En la investigación que nos ocupa y publicada en [Pereira et al., 2014], una población aleatoria de 28 individuos (dos veces el número de características), es generada y conservada para testear como población inicial. La búsqueda genética es ejecutada cinco veces, una por cada función de ajuste ó métrica

diferente utilizada. Una predicción completa 'desde cero' para un individuo y todo el conjunto de datos (las 1237 descargas) tarda 10.5 min, por lo tanto, una generación de 28 individuos dura 4.9 horas.

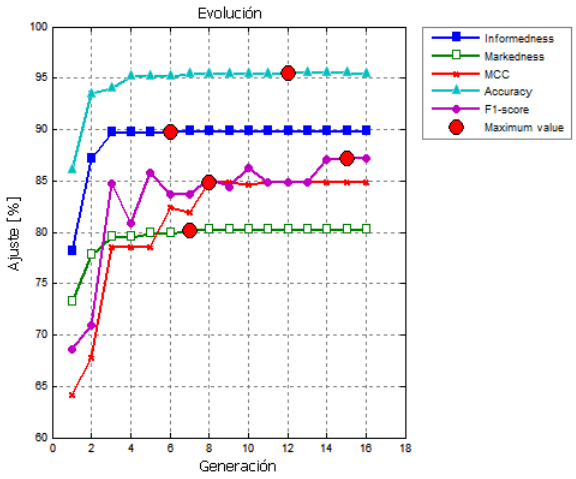


Figura 7. 4. Evolución del ajuste para las diferentes métricas utilizadas

La evaluación de las cinco métricas utilizando la combinación de AG y los predictores Venn se presentan en el gráfico de la Figura 7. 4.

MÉTRICA	PREDICTORES EVALUADOS	TIEMPO TRANSCURRIDO (horas)	GENERACIONES
Informedness	168	29.4	6
Markedness	196	34.3	7
MCC	224	39.2	8
Accuracy	336	58.8	12
F1-score	420	73.5	15

Tabla 7. 2. Tiempo transcurrido en encontrar la mejor solución para diferentes métricas

En la Tabla 7. 2 se pueden observar todos los resultados obtenidos por las diferentes métricas utilizadas. *Informedness* alcanzó los mejores resultados en solamente 6 generaciones, esto equivale a decir que 168 predictores fueron evaluados en tan sólo 29.4 horas. En el otro extremo, la elección de *F1-score* como función de ajuste necesitó 15 generaciones, equivalente a evaluar 420 predictores en 73.5 horas.

Informedness	SR	FA	Features													Generation
%	%	%	1	2	3	4	5	6	7	8	9	10	11	12	13	14	n
89.79	94.00	4.21	0	1	1	1	0	0	1	1	0	0	1	0	0	0	6
89.79	94.00	4.21	0	1	1	1	0	0	1	1	0	0	0	0	0	0	9
89.79	94.00	4.21	0	1	1	1	1	0	0	0	0	0	1	0	0	0	13
89.79	94.00	4.21	0	1	1	1	0	0	0	0	0	0	1	0	0	0	18
89.79	94.00	4.21	0	1	1	1	1	0	0	0	0	0	0	0	0	0	19
89.69	94.00	4.31	0	1	1	1	0	0	0	0	0	0	1	1	0	0	6
89.69	94.00	4.31	0	1	1	1	0	0	0	1	0	0	0	0	0	0	7
89.69	94.00	4.31	0	1	1	1	0	0	1	1	0	0	0	1	0	0	10
89.69	94.00	4.31	0	1	1	1	0	0	1	1	0	0	1	1	0	0	11
89.69	94.00	4.31	0	1	1	1	1	0	1	1	0	0	1	0	0	0	11
89.69	94.00	4.31	0	1	1	1	0	0	1	0	0	0	1	0	0	0	14
89.69	94.00	4.31	0	1	1	1	1	0	1	0	0	0	1	0	0	0	15
89.69	94.00	4.31	0	1	1	1	0	0	0	1	0	0	1	0	0	0	15
89.69	94.00	4.31	0	1	1	1	1	0	1	1	0	0	0	0	0	0	19
89.69	94.00	4.31	0	1	1	1	0	0	0	0	0	0	0	0	0	0	19
89.69	94.00	4.31	0	1	1	1	1	0	0	0	0	0	1	1	0	0	25
89.69	94.00	4.31	0	1	1	1	1	0	0	1	0	0	1	0	0	0	27
89.69	94.00	4.31	0	1	1	1	0	0	1	0	0	0	0	0	0	0	29
89.69	94.00	4.31	0	1	1	1	1	0	1	0	0	0	0	0	0	0	35
89.69	94.00	4.31	0	1	1	1	1	0	0	1	0	0	0	0	0	0	47

Tabla 7. 3. Resultados obtenidos con la métrica Informedness

En la tabla de la Figura 7. 3 se muestran las mejores características utilizando la medida de evaluación *Informedness* como función de ajuste para 50 generaciones. Puede apreciarse en la Figura 7. 5, como la primera de las mejores tasas de acierto aparecen en la generación sexta y la última, de entre todas las mejores tasas de acierto, aparece en la generación número 19 (532 predictores en 93.1 horas empleadas para ello).

Figura 7. 5. Evolución de los mejores individuos y de la media de la población

7.3 Conclusiones

En la publicación [Pereira et al., 2014], cinco medidas de evaluación del rendimiento para tareas de clasificación fueron evaluadas como funciones de ajuste en AG. Las características más relevantes obtenidas en este análisis son consistentes y coincidentes con las alcanzadas en un trabajo anterior prévio, pero con la diferencia de reducir significativamente el tiempo empleado para ello (1731 vs 29.4 horas), una mejora del 98.31%. *Informedness*, *Markedness* y *MCC* muestran mejor rendimiento que las métricas *Accuracy* y *F1-score*, encontrando en menos tiempo las variables más impactantes e importantes para usar en la predicción de disrupciones con elevadas tasas de acierto. Las tres primeras métricas nombradas, que están catalogadas como medidas sin sesgo, son más objetivas manejando ejemplos incorrectamente clasificados y las medidas están mejor ponderadas. Por lo tanto, se puede decir que es realmente significante e importante la selección adecuada de la función de ajuste en los AG para obtener resultados rápidos y precisos.

7.4 Síntesis de publicaciones

Método Selección de características	**Clasificador**	**Métrica de evaluación**	**Publicación**
Búsqueda completa no exhaustiva	Predictores Venn	SR, FA	[Vega et al., 2014]
Algoritmos Genéticos	Predictores Venn	Informedness Markedness MCC F1-Score Accuracy	[Pereira et al., 2014]

Tabla 7. 4. Sístesis de publicaciones capítulo 7

Conclusiones finales de la tesis

La extracción de conocimiento oculto en bases de datos masivas de fusión nuclear, requiere el uso de herramientas y técnicas automáticas de análisis que faciliten la generación de modelos predictivos eficientes y con elevado poder explicativo. En esta tesis, se han presentado diferentes estrategias de búsqueda de patrones que incrementan las recuperaciones de imágenes y formas de onda muy similares. Además, estos métodos han sido implementados en sendas herramientas de usuario gráficas que facilitan su manejabilidad para poder realizar consultas optimizadas y flexibles en la recuperación de patrones. Estas herramientas también fueron instaladas de forma permanente en los ordenadores de los dispositivos experimentales TJ-II y JET.

Los datos generados por estos dispositivos también fueron analizados para seleccionar las características más importantes que intervienen en la clasificación, detección y predicción de diferentes procesos físicos. Distintas aplicaciones experimentales fueron llevadas a cabo en este sentido. En el TJ-II, un sistema permite clasificar las imágenes del diagnóstico de esparcimiento Thomson de forma automática con la operación de dicho dispositivo. Para el reconocimiento en diferido de dichas imágenes, también se aplicaron técnicas de aprendizaje basadas en el algoritmo conformal del vecino más próximo y con elevadas tasas de acierto. También se han desarrollado diferentes rutinas que implementan métodos de sincronización de procesos para tareas de aprendizaje en entornos de supercomputación Linux, aplicaciones JAVA y en sistemas de tiempo real. Se facilitan así recursos de sincronización entre equipos muy heterogéneos que se ejecutan durante la operación del TJ-II.

En el JET, además de la recuperación de patrones gráficos en imágenes, se ha trabajado en la identificación de eventos físicos relevantes, como son las transiciones L/H y las disrupciones del plasma. La aplicación visual para la búsqueda de patrones que identifica la transición L/H en plasmas del JET, ha permitido una primera comprensión de dicha problemática. A partir de aquí, técnicas de clasificación basadas en SVM fueron aplicadas para encontrar que subconjunto de características son las más relevantes para la determinación precisa del instante de tiempo de la transición L/H. Métodos de regresión paramétrica y no paramétrica también se utilizaron en la búsqueda del umbral de potencia cuando el plasma del JET transita del modo L al modo H.

Finalmente, se ha presentado un método que selecciona las características más importantes en la predicción de disrupciones del JET. Basado en la combinación de algoritmos genéticos y predictores probabilísticos Venn. Se complementa este trabajo con métricas de evaluación de clasificadores que mejoran notablemente el tiempo de búsqueda, en la localización de las mejores características.

La presente tesis doctoral es el resultado aglutinador de 5 publicaciones principales y 38 artículos más en los que se colaboró a modo de coautor y que fueron publicados en diferentes revistas científicas indexadas con índice de impacto. Además, se realizaron 3 informes técnicos CIEMAT y 9 artículos de divulgación científica. De los trabajos anteriores, 35 de ellos fueron dados a conocer y presentados en diferentes congresos nacionales e internacionales.

Bibliografía

[Alander, 1992] J.T. Alander. **On optimal population size of genetic algorithms**. Proceedings Computer Systems and Software Engineering. *6th Annual European Computer Conference, 65-70*. 1992

[Alfven, 1942] H. Alfven. **Existence of electromagnetic-hydrodynamic waves** Nature, Vol. 150, pp. 405-406. 1942

[An et al., 2007] S. An, W. Liu, S. Venkatesh. **Face recognition using KRR**. IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07. 17-22 June 2007. ISSN: 1063-6919

[Baker, 1987] J.E. Baker. **Reducing bias and ineffiency in the selection algorithm**. Proceedings of the Second International Conference on genetic Algorithms and Their Applications, 14-21.

[Castro y López, 2001] R. Castro, D. López. **The PAPI system: point of access to providers of information**. *Computer Networks, Volume 37, Issue 6, December 2001, Pages 703-710*. ISSN: 1389-1286

[Castro et al., 2008] R. Castro, J. Vega, A. Portas, A. Pereira, D. R. López. **EFDA-Fed: Una federación internacional para investigación en fusión basada en PAPI**. Jornadas Técnicas RedIRIS 2007, Mieres (Asturias), 19-23 de noviembre de 2007. Publicación de la red nacional de I+D. Boletín RedIris, nº 82-83, pp 17-23,Abril 2008. ISSN: 1139-207X
http://www.rediris.es/difusion/publicaciones/boletin/82-83/ponencia1.1B.pdf

[Castro et al., 2008b] R. Castro, J. Vega, A. Portas, A. Pereira, C. Rodriguez, S. Balme, J. M. Theis, J. Signoret, P. Lebourg, K. Purahoo, K.Thomsen, H. Fernandes, A. Neto, A. Duarte, F. Oliveira, F. Reis, J. Kadlecsik. **EFDA-fed: European federation among fusion energy research laboratories.** Terena networking conference. Bruges (Belgium) 19-22 May 2008. *Campus-Wide Information Systems. Vol 25, Number 5, 2008, pp. 359-373, Emerald Group Publishing Limited.* ISSN: 1065-0741
http://dx.doi.org/10.1108/10650740810921493

[Castro et al., 2009] R. Castro, J. Vega, A. Pereira, A. Portas. **Data distribution architecture based on standard Real Time Protocol.** Proceedings of the 25th Symposium on Fusion Technology - SOFT-25. Rostock, Germany 15–19 September. *Fusion Engineering and Design. Volume 84, Issues 2-6, June 2009, Pages 565-568.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2008.12.074

[Castro et al., 2010] R. Castro, J. Vega, T. Fredian, K. Purahoo, A. Pereira, A. Portas. **Securing MDSplus in a Multi-organization Environment.** Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 15 - 19 June 2009, Aix-en-Provence, France. *Fusion Engineering and Design. Volume 85, Issues 3–4, July 2010, Pages 614–617.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2010.03.016

[Castro et al., 2010b] R. Castro, K. Kneupner, J. Vega, G. De Arcas, J.M. López, K. Purahoo, A. Murari, A. Fonseca, A. Pereira, A. Portas and JET-EFDA Contributors. **Real-time remote diagnostic monitoring test-bed in JET.** Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 15 - 19 June 2009, Aix-en-Provence, France. *Fusion Engineering and Design. Volume 85, Issues 3–4, July 2010, Pages 598–60.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2010.03.050

[Chatterjee, 2006] S. Chatterjee, A. Hadi. **Regression analysis by example**. 2006. Fourth Edition. Wiley-Interscience. ISBN: 100471746967

[Chawla, 2005] N.V. Chawla. **Data Mining for Imbalanced Datasets: An Overview**. *Data Mining and Knowledge Discovery Handbook 2005, pp 853-867*. ISBN 978-0-387-24435-8

[Cherkassky y Mulier, 2007] V. Cherkassky, F. Mulier. **Learning from Data. Concepts, Theory, and Methods**. 2nd Edition. John Wiley & Sons, Inc. 2007. ISBN 978-0-471-68182-3

[David, 2011] M.W. David. **Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation**. *Journal of Machine Learning Technologies. vol 2, Issue 1, 2011, pp 37-63*. ISSN: 2229-3981.

[Dashevskiy et al., 2008] M. Dashevskiy, Z. Luo. **Reliable Probabilistic Classification and Its Application to Internet Traffic**. *Advanced Intelligent Computing Theories and Applications with Aspects of Theoretical and Methodological Issues. Lecture Notes in Computer Science Volume 5226, pp 380-388*. Springer, Heidelberg (2008)

[Dormido-Canto et al., 2006] S. Dormido-Canto, G. Farias, J. Veja, R. Dormido, J. Sánchez, N. Duro, M. Santos, J.A. Martin, G. Pajares. **Search and retrieval of plasma wave forms: Structural pattern recognition approach**. *Review of Scientific Instruments. 77, 10F514. 2006*

[Dormido-Canto et al., 2008] S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, N. Duro, H. Vargas, G. Rattá, A. Pereira, A. Portas. **Structural pattern recognition methods based on string comparison for fusion databases**. Proceedings of the 6th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. Inuyama, Japan 4–8 June 2007. *Fusion Engineering and Design. Vol 83, Issues 2-3, pp 421-424. April 2008.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2007.11.009

[Dormido-Canto et al., 2013] S. Dormido-Canto, J. Vega, J. M. Ramírez, A. Murari, R. Moreno, J. M. López, A. Pereira and JET-EFDA Contributors. **Development of an efficient real-time disruption predictor from scratch on jet and implications for ITER.** *Nuclear Fusion, 53, Issue 11, 8pp, November 2013.* ISSN: 0029-5515
http://dx.doi.org/10.1088/0029-5515/53/11/113001

[Farias et al., 2006] G. Farias, S. Dormido-Canto, J. Vega, J. Sánchez, N. Duro, R. Dormido, M. Ochando, M. Santos, G. Pajares. **Searching for patterns in TJ-II time evolution signals**. *Fusion Engineering and Design. Volume 81, Issues 15–17, July 2006, Pages 1993–1997*. ISSN: 0920-3796

[Farias et al., 2012] G. Farias, J. Vega, S. González, A. Pereira, X. Lee, D. Schissel, P. Gohil. **Automatic Determination of L/H Transition Times in DIII-D Through a Collaborative Distributed Environment.** Proceedings of the 8th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Francisco, California, United States of America 20-24 June 2011. *Fusion Engineering and Design. Vol. 87, Issue 12, December 2012, Pages 2081–2083.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2012.02.038

[Forster, 2012] M. Forster. **Runaway Electrons in Disruptions and Perturbed Magnetic Topologies of Tokamak Plasmas**. Dissertation. *Dusseldorf University, June 2012*.

[Gammerman et al., 1998] A. Gammerman, V. Vovk, V. Vapnik. **Learning by transduction**. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. 1998

[Gammerman et al., 2007] A. Gammerman, V. Vovk. **Hedging predictions in machine learning**. *The Computer Journal. Volume 50 Issue 2, March 2007. Pages 151-163* Oxford University Press Oxford, UK.

[García et al., 2001] P. García, A. Barbolla, M. Romero, C. Alejaldre, E. González, J. Jorcano. **Tecnologías energéticas e impacto ambiental**. *McGraw Hill, Primera edición. Madrid 2001*. ISBN: 84-481-3360-9

[García et al., 2006] J. García, G. Chagolla, S. Noriega. **Efectos de la colinealidad en el modelado de la regresión y su solución**. *Cultura Científica y Tecnológica. 16. 23-34*. 2006

[González et al., 2010] S. González, J. Vega, A. Murari, A. Pereira, J.M. Ramírez, S. Dormido-Canto and JET-EFDA contributors. **Support vector machine based feature extractor for L/H transitions in JET**. Proceedings of the 18th High Temperature Plasma Diagnostics (HTPD) conference, May 16-20, 2010, Wildwood, New Jersey, USA. *Review of Scientific Instruments. Vol. 81 Issue 10. 10E123 (2010).* ISSN: 0034-6748, E-ISSN: 1089-7623
http://dx.doi.org/10.1063/1.3502327

[González et al., 2010b] S. González, J. Vega, A. Murari, A. Pereira, M. Beurskens and JET-EFDA contributors. **Automatic ELM location in JET using a universal multi-event locator.** 6th Fusion Data Validation Workshop 2010. January 25-27, 2010, Madrid, Spain. *Fusion Science and Technology, Vol. 58, nº 3, pp 755-762. Nov 2010.* ISSN: 1535-1055
http://www.new.ans.org/pubs/journals/fst/a_10924

[González et al., 2012] S. González, J. Vega, A. Murari, A. Pereira, S. Dormido-Canto, J.M. Ramírez and JET-EFDA contributors. **Automatic location of L/H transition times for physical studies with a large statistical basis.** *Plasma Physics and Controlled Fusion 54, 065009, Issue 6, 19 pp. June 2012.* ISSN 0741-3335
http://dx.doi.org/10.1088/0741-3335/54/6/065009

[González et al., 2012b] S. González, J. Vega, A. Pereira, I. Pastor. **Region selection and image classification methodology using a non-conformity measure.** *Progress in Artificial Intelligence. September 2012, Vol. 1, Issue 3, pp 215-222.* ISSN: 2192-6352 (Print) 2192-6360 (Online).
http://dx.doi.org/10.1007/s13748-012-0020-z

[González et al., 2012c] S. González, J. Vega, A. Murari, A. Pereira and JET-EFDA contributors. **Automated Analysis of Edge Pedestal Gradient Degradation During ELMs.** 7th Workshop on Fusion, Data Processing, Validation and Analysis. March 26-28, 2012,ENEA, Frascati, Italy. *Fusion Science and Technology. Volume 62. Number 3. November 2012. Pages 403-408.* ISSN: 1535-1055
http://www.ans.org/pubs/journals/fst/a_15339

[González et al., 2012d] S. González, J. Vega, A. Murari, A. Pereira, S. Dormido-Canto, J.M. Ramírez and JET EFDA contributors. **H/L transition time estimation in JET using conformal predictors.** Proceedings of the 8th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Francisco, California, United States of America 20-24 June 2011. *Fusion Engineering and Design. Vol. 87, Issue 12, December 2012, Pages 2084–2086.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2012.02.126

[Greenwald et al., 2005] M. Greenwald, D. Schissel, J. R. Burruss, T. Fredian, J. Lister, J. Stillerman. **Visions for data management and remote collaboration for ITER**. ICALEPCS 2005. Geneva, Switzerland.

[Haupt et al., 2007] R. L. Haupt, D. H. Werner. **Genetic Algorithms in Electromagnetics**. IEEE Press, Wiley Interscience. 2007. ISBN: 978-0-471-48889-7

[Hidalgo et al., 2005] C. Hidalgo, C. Alejaldre, A. Alonso, J. Alonso, L. Almoguera, F. de Aragón, E. Ascasíbar, A. Baciero, R. Balbín, E. Blanco, J. Botija, B. Brañas, E. Calderón, A. Cappa, J.A. Carmona, R. Carrasco, F. Castejón, J.R. Cepero, A.A. Chmyga, J. Doncel, N.B. Dreval, S. Eguilior, L. Eliseev, T. Estrada, J.A. Ferreira, A. Fernández, J.M. Fontdecaba, C. Fuentes, A. García, I. García-Cortés, B. Gonçalves, J. Guasp, J. Herranz, A. Hidalgo, R. Jiménez, J.A. Jiménez, D. Jiménez-Rey, I. Kirpitchev, S.M. Khrebtov, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Rázola, A. López-Sánchez, E. de la Luna, G. Marcon, R. Martín, K.J. McCarthy, F. Medina, M. Medrano, A.V. Melnikov, P. Méndez, B. van Milligen, I.S. Nedzelskiy, M. Ochando, O. Orozco, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, A. Petrov, S. Petrov, A. Portas, D. Rapisarda, L. Rodríguez-Rodrigo, E. Rodríguez-Solano, J. Romero, A. Salas, E. Sánchez, J. Sánchez, M. Sánchez, K. Sarksian, C. Silva, S. Schchepetov, N. Skvortsova, F. Tabarés, D. Tafalla, A. Tolkachev, V. Tribaldos, I. Vargas, J. Vega, G. Wolfers, B. Zurro. **Overview of TJ-II experiments.** *Nuclear Fusion, Vol. 45, nº 10, pp 266-275. October 2005.* ISSN: 0029-5515
http://dx.doi.org/10.1088/0029-5515/45/10/S22

[Holland, 1975] J.H. Holland. **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence**. University of Michigan Press, 1975 - 183 páginas. ISBN: 0472084607.

[Hosmer et al., 2000] D. W. Hosmer, S. Lemeshow. **Applied Logistic Regression**, 2nd ed. New York; Chichester, Wiley. ISBN 0-471-35632-8.

[Izenman, 2008] A. J. Izenman. **Modern Multivariate Statistical Techniques**. Springer. ISSN: 1431-875X

[Jong, 1975] K.A. De Jong. **An analysis of the behaviour of a class of genetic adaptive systems**. Tesis doctoral 1975, University of Michigan.

[Kumar et al., 2014] V. Kumar, B. Kumar. **Genetic algorithm: an overview and its application**. *International Journal of Advanced Studies in Computer Science and Engineering. Vol. 3, Issue 2, 2014.*

[Lean et al., 2007] Y. Lean, S. Wang, K. Lai. **Foreign-Exchange-Rate Forecasting with Artificial Neural Networks**. *International series in operational research and management Science*. Springer 2007. ISBN-13: 978-0-387-71719-7

[López et al., 2012] J. M. López, J. Vega, D. Alves, Member, S. Dormido-Canto, A. Murari, J.M. Ramírez, R. Felton, M. Ruiz, G. de Arcas and JET EFDA contributors **Implementation of the disruption predictor APODIS in JET real-time network using the MARTe framework**. *18th Real-Time Conference. 11th June – 15th June, 2012. Berkeley, CA (USA).* Submitted to IEEE Trans. on Nuclear Science

[Marqués de Sa, 2001] J.P. Marqués de Sá. **Pattern Recognition**. Springer. ISBN 978-3-642-56651-6

[McDonald et al., 2006] D. McDonald, A. J. Meakins, J. Svensson, A. Kirk, Y. Andrew, J. G. Cordey and ITPA H-mode Threshold Database WG. **The impact of statistical models on scalings derived from multi-machine H-mode threshold experiments**. *Plasma Phys. Control. Fusion 48 (2006) A439–A447.*

[Makili et al., 2010] L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, A. Portas, D. Pérez-Risco, M.C. Rodríguez-Fernández, P. Busch. **Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: new image recognition classifier and fault condition detection.** Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 15 - 19 June 2009, Aix-en-Provence, France. *Fusion Engineering and Design. Volume 85, Issues 3-4, July 2010, Pages 415-418.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2009.10.004

[Márquez, 2004] F. Márquez. **UNIX Programación avanzada**. 3ª edición, Editorial Ra-Ma, (2004), ISBN 84-7897-603-5, Pag 403

[Martin et al., 2008] Y. Martin, T. Takizuka and ITPA CDBM H-mode Threshold Database Working Group. **Power requirements for accessing the H mode in ITER**. 11th IAEA Technical Meeting on H-mode Physics and Transport Barriers. *Journal of Physics: Conference Series 123 (2008) 012033*.

[McCracken, 2005] G. M. McCracken, Peter Stott. **Fusion, the energy of the universe**. *Elsevier Academic Press, 2005*. ISBN: 0-12-481851-X

[Miyamoto, 2011] K. Miyamoto. **Plasma Physics and Controlled Nuclear Fusion**. Third edition. *Research Report NIFS-PROC-88 Series, 2011*.

[Nagarajayya y Gupta, 2000] N. Nagarajayya, A. Gupta., **Porting of Win32 API WaitFor to Solaris**, 2000.
http://developers.sun.com/solaris/articles/waitfor_api.html

[Nouretdinov et al., 2001] I. Noretdinov, T. Melluish, V. Vovk. **Ridge Regression Confidence Machine**. *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001.

[Nouretdinov et al., 2012] I. Nouretdinov, D. Devetyarov, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss. **Multiprobabilistic Venn Predictors with Logistic Regression**. *AIAI 2012 Workshops, IFIP AICT 382, pp. 224-233*. Springer, 2012

[Papadopoulos, 2013] H. Papadopoulos. **Reliable Probabilistic Classification with Neural Networks**. *Timely Neural Networks Applications in Engineering. Selected Papers from the 12th EANN International Conference*, 2011. Neurocomputing. Volume 107, 1 May 2013, Pages 59–68.

[Pearson, 2005] K. Pearson. Mining imperfect data. SIAM. ISBN 0-89871-582-2

[Pereira, 2000] A. Pereira. **Fundamentos de los criptosistemas de clave pública para internet.** *Revista Técnica Industrial, nº 238, Septiembre 2000, pp 40-44.* ISSN: 0040-1838.
http://dialnet.unirioja.es/servlet/articulo?codigo=25234

[Pereira, 2001] A. Pereira. **Control de transductores de posición en tiempo real.** *Revista Técnica Industrial, nº 242, Septiembre 2001, pp 70-73.*ISSN: 0040-1838.
http://dialnet.unirioja.es/servlet/articulo?codigo=264558

[Pereira, 2002] A. Pereira. **Seguridad y privacidad en los sistemas de pagos.** *Revista Técnica Industrial, nº 244, Marzo 2002, pp 70-75.* ISSN: 0040-1838.
http://www.tecnicaindustrial.es/TIAdmin/Numeros/1/47/a47.pdf

[Pereira, 2009] A. Pereira. **Búsqueda y reconocimiento de patrones en señales de evolución temporal**. Proyecto Fin de Carrera. Ingeniería en Informática. UOC, 2009.

[Pereira, 2010] A. Pereira. **Análisis predictivo de datos mediante técnicas de regresión estadística**. *Master's Thesis, Máster en Investigación en Informática*. Facultad de Informática, Departamento de Arquitectura de Computadores y Automática, Universidad Complutense de Madrid, 2009-2010.
http://eprints.ucm.es/11389/

[Pereira et al., 2004] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **Control Electronics and Data Acquisition for the Spanish CRG Beamline SpLine at the E.S.R.F.** *Proceedings 1ª Reunión Nacional de Usuarios de Radiación Sincrotrón, Torremolinos (Málaga), 5-6 Febrero 2004.*

[Pereira et al., 2004b] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **Compiling the Linux kernel 2.2.12 on VME-PowerPC architecture**. *Linux Gazette, Issue 103, June 2004.*
http://web.archive.org/web/20050320005058/www.linuxgazette.com/node/9024

[Pereira et al., 2004c] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **Configuring a VME-Linux Diskless System on Motorola MVME24xx**. *Linux Gazette, Issue 103, June 2004.*
http://web.archive.org/web/20050320004541/www.linuxgazette.com/node/9011

[Pereira et al., 2004d] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **A flexible VME-PowerPC-Linux Device Driver for MEN A201S carrier board and MEN M37 mezzanine module**. *Linux Gazette, Issue 107, October 2004.*
http://web.archive.org/web/20050416194353/www.linuxgazette.com/node/9436

[Pereira et al., 2004e] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **Adding 16 bits-ADC extensions on VME bus with MEN M36 Linux Device Driver**. *Linux Gazette, Issue 107, October 2004.*
http://web.archive.org/web/20050416194358/http://www.linuxgazette.com/node/9437

[Pereira et al., 2005] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **Control and data acquisition system for the Spanish Beamline (BM25) at the ESRF**. *Informes Técnicos Ciemat, nº 1056, Abril 2005, 108 pp.* ISSN: 1135-9420. http://publicacionesoficiales.boe.es/detail.php?id=2251265405-0001 http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/36/083/36083410.pdf

[Pereira et al., 2005b] A. Pereira, C. Olalla, J. Sánchez, G.R. Castro. **Control y adquisición de datos en la línea española del sincrotrón de Grenoble.** *Revista Técnica Industrial, nº 260, Diciembre 2005, pp 40-43.* ISSN: 0040-1838. http://www.tecnicaindustrial.es/TIAdmin/Numeros/20/39/a39.pdf

[Pereira et al., 2006] A. Pereira, J. Vega, L. Pacios, E. Sánchez, A. Portas. **Synchronization resources in heterogeneous environments: Time-sharing, real-time and Java.** Proceedings of the 5th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 12 - 15 July 2005; Budapest (Hungary). *Fusion Engineering and Design, Volume 81, Issues 15-17, July 2006. Pages 1869-1872.* ISSN: 0920-3796. http://dx.doi.org/10.1016/j.fusengdes.2006.04.016

[Pereira et al., 2010] A. Pereira, J. Vega, R. Castro, A. Portas. **Distributed open environment for data retrieval based on pattern recognition techniques.** Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research, 15-19 June 2009. Aix-In-Provence (France). *Fusion Engineering and Design. Volume 85, Issues 3-4, July 2010, Pages 595-597.*ISSN: 0920-3796 http://dx.doi.org/10.1016/j.fusengdes.2009.12.009

[Pereira et al., 2010b] A. Pereira, J. Vega, A. Portas, R. Castro, A. Murari and JET/EFDA Contributors. **Optimized search strategies to improve structural pattern recognition techniques.** Proceedings of the 8th International FLINS Conference. September 21-24, 2008, Madrid (Spain). *World Scientific Proceedings Series on Computer Engineering and Information Science. Computational Intelligence in Decision and Control. Vol. 1. (pp 405-410). September 2008.* ISBN-13: 978-981-279-946-3,ISBN-10:981-279-946-X. http://dx.doi.org/10.1142/9789812799470_0066 *International Journal of Nuclear Knowledge Management. Vol. 4, No.1 pp. 18 – 24, 2010.* ISSN: 1479-540X http://dx.doi.org/10.1504/IJNKM.2010.031151

[Pereira et al., 2014] A. Pereira, J. Vega, R. Moreno, S. Dormido-Canto, G. Rattá and JET/EFDA Contributors. **Feature selection for disruption prediction from scratch in JET by using genetic algorithms and probabilistic predictors**. Proceedings of the 28th Symposium on Fusion Technology. San Sebastián, Spain. September 29th – October 3rd. SOFT 2014. *Fusion Engineering and Design*. ISSN: 0920-3796 (*Accepted for publication*) http://dx.doi.org/10.1016/j.fusengdes.2015.04.040

[Pereira y Vega, 2005] A. Pereira, J. Vega. **Desarrollo de entornos cliente para un sistema de sincronización basado en eventos**. *Informes Técnicos Ciemat, nº 1064, Octubre 2005, 172 pp.* ISSN: 1135-9420.

http://publicacionesoficiales.boe.es/detail.php?id=2251365405-0001

[Powers, 2011] D. Powers. **Evaluation, from precision, recall and F-measure to ROC, informedness, markedness and correlation**. *Journal of Machine Learning Technologies. Volume 2, Issue 1, 2011, pp-37-63*. ISSN: 2229-3981

[Rabinovich, 2005] S. Ravinobich. **Measurements Errors and Uncertainties**. Third edition. 2005. Springer. ISBN-10: 0-387-25358-0

[Ramírez et al., 2010] J. Ramírez, S. Dormido-Canto, J. Vega. **Parallelization of automatic classification systems based on support vector machines: Comparison and application to JET database**. *Fusion Engineering and Design. 85 (2010) 425–427.*

[Rattá et al., 2008] G. A. Rattá, J. Vega, A. Pereira, A. Portas, E. de la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari. **First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET**. Proceedings of the 6th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. Inuyama, Japan 4–8 June 2007. *Fusion Engineering and Design. Vol 83, Issues 2-3, pp 467-470. April 2008.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2007.09.008

[Rattá et al., 2012] G.A. Rattá. **Improved feature selection based on genetic algorithms for real time disruption prediction on JET**. *Fusion Engineering and Design. Volume 87, Issue 9, September 2012, Pages 1670–1678.*

[Reux, 2010] C. Reux. **Study of a disruption mitigation method for tokamak plasmas**. Dissertation. *Plasma Physics. Ecole Polytechnique X, 2010*. French.

[Ruiz., 2006] R. Ruiz. **Heurísticas de selección de atributos para datos de gran dimensionalidad**. *Tesis doctoral em informática. Universidad de Sevilla 2006.*

[Sánchez et al., 2006] E. Sánchez , A. Portas, A. Pereira, J. Vega. **Applying a message oriented middleware architecture to the TJ-II remote participation system.** 5th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 12 - 15 July 2005 , Budapest , Hungary. *Fusion Engineering and Design, Vol 81, Issues 15-17, pp 2063–2067. July 2006.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2006.04.057

[Sánchez et al., 2006b] E. Sánchez, A. Portas, A. Pereira, J. Vega. **Visual data analysis in the TJ-II remote participation system**. *Informes Técnicos Ciemat, nº 1094, Noviembre 2006, 20 pp,* ISSN: 1135-9420.
http://publicacionesoficiales.boe.es/detail.php?id=069065407-0001
http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/38/061/38061226.pdf

[Sánchez et al., 2007] J. Sánchez, M. Acedo, A. Alonso, J. Alonso, P. Alvarez, F. de Aragón, E. Ascasíbar, A. Baciero, R. Balbín, L. Barrera, E. Blanco, J. Botija, B.

Brañas, E. de la Cal, E. Calderón, I. Calvo, A. Cappa, J.A. Carmona, B.A. Carreras, R. Carrasco, F. Castejón, G. Catalán, A.A. Chmyga, N.B. Dreval, M. Chamorro, S. Eguilior, J. Encabo, L. Eliseev, T. Estrada, A. Fernández, R. Fernández, J.A. Ferreira, J.M. Fontdecaba, C. Fuentes, J. de la Gama, A. García, L. García, I. García-Cortés, J.M. García-Regaña, B. Gonçalves, J. Guasp, J. Herranz, A. Hidalgo, C. Hidalgo, R. Jiménez-Gómez, J.A. Jiménez, D. Jiménez, I. Kirpitchev, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Rázola, A. López-Sánchez, E. de la Luna, G. Marcon, F. Martín, L. Martínez-Fresno, K.J. McCarthy, F. Medina, M. Medrano, A.V. Melnikov, P. Méndez, E. Mirones, B. van Milligen, I.S. Nedzelskiy, M. Ochando, J. Olivares, R. Orozco, P. Ortiz, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, D. Pérez-Risco, A. Petrov, S. Petrov, A. Portas, D. Rapisarda, L. Ríos, C. Rodríguez, L. Rodríguez-Rodrigo, E. Rodríguez-Solano, J. Romero, A. Ros, A. Salas, E. Sánchez, M. Sánchez, E. Sánchez-Sarabia, X. Sarasola, K. Sarksian, C. Silva, S. Schchepetov, N. Skvortsova, A. Soleto, F. Tabarés, D. Tafalla, J. Tera, A. Tolkachev, V. Tribaldos, V.I. Vargas, J. Vega, G. Velasco, M. Weber, G. Wolfers, S.J. Zweben, B. Zurro. **Overview of TJ-II experiments.** *Nuclear Fusion, Vol. 47, nº 10, pp 677-685. October 2007.* ISSN: 0029-5515
http://dx.doi.org/10.1088/0029-5515/47/10/S16

[Sánchez et al., 2007b] E. Sánchez, A. Portas, A. Pereira, J. Vega, I. Kirpitchev. **Remote control of data acquisition devices by means of message oriented middleware.** Proceedings of the 24th Symposium on Fusion Technology - SOFT-24, Warsaw, Poland, 11-15 September 2006. *Fusion Engineering and Design. Vol . 82, Issues 5-14, pp 1365-1371. October 2007.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2007.03.002

[Sánchez et al., 2008] E. Sánchez, A. de la Peña, A. Portas, A. Pereira, J. Vega, A. Neto, H. Fernandes. **An event-oriented database for continuous data flows in the TJ-II environment**. Proceedings of the 6th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. Inuyama, Japan 4–8 June 2007. *Fusion Engineering and Design. Vol 83, Issues 2-3, pp 413-416. April 2008.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2007.11.007

[Sánchez et al., 2009] J. Sánchez, M. Acedo, A. Alonso, J. Alonso, P. Alvarez, E. Ascasíbar, A. Baciero, R. Balbín, L. Barrera, E. Blanco, J. Botija, A. de Bustos, E. de la Cal, I. Calvo, A. Cappa, J.M. Carmona, D. Carralero, R. Carrasco, B.A. Carreras, F. Castejón, R. Castro, G. Catalán, A.A. Chmyga, M. Chamorro, L. Eliseev, L. Esteban, T. Estrada, A. Fernández, R. Fernández-Gavilán, J.A. Ferreira, J.M. Fontdecaba, C. Fuentes, L. García, I. García-Cortés, R. García-Gómez, J.M. García-Regaña, J. Guasp, L. Guimarais, T. Happel, J. Hernanz, J. Herranz, C. Hidalgo, J.A. Jiménez, A. Jiménez-Denche, R. Jiménez-Gómez, D. Jiménez-Rey, I. Kirpitchev, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Rázola, A. López-Sánchez, S. Lysenko, G. Marcon, F. Martín, V. Maurin, K.J. McCarthy, F. Medina, M. Medrano, A.V. Melnikov, P. Méndez, B. van Milligen, E. Mirones, I.S. Nedzelskiy, M. Ochando, J. Olivares, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, G. Pérez, D. Pérez-Risco, A. Petrov, S. Petrov, A. Portas, D. Pretty, D. Rapisarda, G. Rattá, J.M. Reynolds, E. Rincón, L.

Ríos, C. Rodríguez, J.A. Romero, A. Ros, A. Salas, M. Sánchez, E. Sánchez, E. Sánchez-Sarabia, K. Sarksian, J.A. Sebastián, C. Silva, S. Schchepetov, N. Skvortsova, E.R. Solano, A. Soleto, F. Tabarés, D. Tafalla, A. Tarancón, Yu. Taschev, J. Tera, A. Tolkachev, V. Tribaldos, V.I. Vargas, J. Vega, G. Velasco, J.L. Velasco, M. Weber, G. Wolfers, B. Zurro. **Confinement transitions in TJ-II under Li-coated wall conditions.** *Nuclear Fusion, Vol. 49, Issue 10, 104018. October 2009.* ISSN: 0029-5515
http://dx.doi.org/10.1088/0029-5515/49/10/104018

[Sánchez et al., 2011] J. Sánchez, M. Acedo, D. Alegre, A. Alonso, J. Alonso, P. Álvarez, J. Arévalo, E. Ascasíbar, A. Baciero, D. Baiao, L. Barrera, E. Blanco, J. Botija, A. Bustos, E. de la Cal, I. Calvo, A. Cappa, D. Carralero, R. Carrasco, B.A. Carreras, F. Castejón, R. Castro, G. Catalán, A.A. Chmyga, M. Chamorro, L. Eliseev, L. Esteban, T. Estrada, J.A. Ferreira, J.M. Fontdecaba, L. García, R. García-Gómez, J.M. García-Regaña, P. García-Sánchez, A. Gómez-Iglesias, S. González, J. Guasp, T. Happel, J. Hernanz, J. Herranz, C. Hidalgo, J.A. Jiménez, A. Jiménez-Denche, R. Jiménez-Gómez, I. Kirpitchev, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Razola, T. Madeira, F. Martín-Díaz, F. Martín-Hernández, A.B. Martín-Rojo, J. Martínez-Fernández, K.J. McCarthy, F. Medina, M. Medrano, L. Melón, A.V. Melnikov, P. Méndez, B. van Milligen, E. Mirones, A. Molinero, M. Navarro, I.S. Nedzelskiy, M. Ochando, J. Olivares, E. Oyarzábal, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, A. Petrov, S. Petrov, A. Portas, G. Rattá, J.M. Reynolds, E. Rincón, L. Ríos, C. Rodríguez, B. Rojo, J.A. Romero, A. Ros, M. Sánchez, E. Sánchez, G. Sánchez-Burillo, E. Sánchez-Sarabia, K. Sarksian, J.A. Sebastián, C. Silva, E.R. Solano, A. Soleto, F. Tabarés, D. Tafalla, J. Tera, A. Tolkachev, J. Vega, G. Velasco, J.L. Velasco, M. Weber, G. Wolfers and B. Zurro. **Overview of TJ-II experiments.** *Nuclear Fusion. Vol. 51, Issue 9. August 2011.*ISSN: 0029-5515
http://dx.doi.org/10.1088/0029-5515/51/9/094022

[Sánchez et al., 2013] J. Sánchez, D. Alegre, A. Alonso, J. Alonso, P. Álvarez, J. Arévalo, E. Ascasíbar, A. Baciero, D. Baião, E. Blanco, M. Borchardt, J. Botija, A. Bustos, E. de la Cal, I. Calvo, A. Cappa, D. Carralero, R. Carrasco, F. Castejón, R. Castro, G. Catalán, A.A. Chmyga, M. Chamorro, L. Eliseev, T. Estrada, F. Fernández, J.M. Fontdecaba, L. García, R. García-Gómez, P. García-Sánchez, S. da Graça, J. Guasp, R. Hatzky, J. Hernández, J. Hernanz, J. Herranz, C. Hidalgo, J.A. Jiménez, A. Jiménez-Denche, I. Kirpitchev, R. Kleiber, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Razola, A. Martín, F. Martín-Díaz, F. Martín-Hernández, A.B. Martín-Rojo, J. Martínez-Fernández, K.J. McCarthy, F. Medina, M. Medrano, L. Melón, A.V. Melnikov, P. Méndez, B. van Milligen, P. Monreal, M. Navarro, I.S. Nedzelskiy, M.A. Ochando, J. Olivares, E. Oyarzábal, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, A. Petrov, S. Petrov, A.B. Portas, E. Rincón, L. Ríos, C. Rodríguez, B. Rojo, J.A. Romero, A. Ros, M. Sánchez, E. Sánchez, E. Sánchez-Sarabia, K. Sarksian, J.A. Sebastián, C. Silva, E.R. Solano, A. Soleto, B. Sun, F.L. Tabarés, D. Tafalla, M. Tereshchenko, A. Tolkachev, J. Vega, G. Velasco, J.L. Velasco, G. Wolfers and B. Zurro. **Dynamics of flows and confinement in the TJ-II stellarator.** *Nuclear Fusion. Vol. 53, Issue 10. October 2013.* ISSN: 0029-5515

http://dx.doi.org/10.1088/0029-5515/53/10/104016

[Saunders et al., 1998] C. Saunders, A. Gammerman, V. Vovk. **Ridge Regression Learning Algorithm in Dual Variables**. Proceeding ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning. Pages 515-521. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. 1998. ISBN:1-55860-556-8

[Saunders et al., 1999] C. Saunders, A. Gammerman, V. Vovk. **Transduction with Confidence and Credibility**. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Pages 722-726*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1999. ISBN:1-55860-613-0

[Schaffer et al., 1989] J.D.Schaffer, R.A. Caruna, L.J. Eshelman, R. Das. **A study of control parameters affecting online performance of genetic algorithms for function optimization**. Proceedings of the Third International Conference on Genetic Algorithms, Morgan Kaufmann, 51-60.

[Scholkopf et al., 2002] B. Scholkopf, A. J. Smola. **Learning from data**. The MIT press. 2002. ISBN 0-262-19475-9.

[Shafer y Vovk, 2008] G. Shafer, V. Vovk. **A Tutorial on Conformal Prediction**. *Journal of Machine Learning Research*. 371-421. 2008

[Shawe-Taylor et al., 2004] J. Shawe-Taylor, H. Cristianini. **Kernel methods for pattern analysis**. Cambridge University press. 2004. ISBN. 10 0-511-21237-2.

[Smith, 1997] Steven W. Smith. **The Scientist and Engineer's Guide to Digital Signal Processing**. *California Technical Publishing. San Diego, EEUU. 1997*. ISBN 0-9660176-3-3

[Theodoridis et al., 2003] S. Theodoridis, K. Koutrombas. **Pattern Recognition**. Second edition. Elsevier academic press. 2003. ISBN: 0-12-685875-6

[Thornton, 2011] A. Thornton. **The impact of transient mitigation schemes on the MAST edge plasma**. Dissertation. *University of York. Department of Physics. 2011*.

[Tong y Svetnik, 2002] C. Tong, V. Svetnik. **Novelty Detection in Mass Spectral Data using a Support Vector Machine Method**. *Proc. of Interface*. 2002

[Vapnik, 2000] V. Vapnik. **The nature of statistical learning theory**. Second edition. Springer. New York, 2000. ISBN: 0-387-98780-0

[Vega, 2007] J. Vega. **Data retrieval based on physical criteria: a structural approach for massive databases**. EFDA-JET Seminar. 26th April 2007. HOW Room. Culham Science Center. UK

[Vega et al., 2004] J. Veja. E. Sanchez, A. Portas, M. Ochando, A. Mollinedo, J. Munoz, M. Ruiz, E. Barrera, S. Lopez. **A distributed synchronization system for the TJ-II local area network**. *Fusion Eng. 71 (2004), pp. 117-221*.

[Vega et al., 2004b] J. Vega, E. Sánchez, A. Portas, M. Ruiz, E. Barrera, S. López. "**A multi-tier approach for data acquisition programming in the TJ-II remote participation system**". *Review of Scientific Instruments. 75, 10 (2004) 4251-4253*

[Vega et al., 2005] J. Vega, I. Pastor, J. L. Cereceda, A. Pereira, J. Herranz, D. Pérez, M. C. Rodríguez, G. Farias, S. Dormido-Canto, J. Sánchez, R. Dormido, N. Duro, S. Dormido, G. Pajares, M. Santos, J. M. de la Cruz. **Application of intelligent classification techniques to the TJ-II Thomson Scattering diagnostic.** 32nd EPS Conference on Plasma Physics and Controlled Fusion combined with the 8th International Workshop on Fast Ignition of Fusion Targets. Tarragona (Spain), 27 June - 1 July 2005. *Europhysics Conference Abstracts (ECA), Vol.29C, P-2.090 (2005)*
http://epsppd.epfl.ch/Tarragona/pdf/P2_090.pdf

[Vega et al., 2005b] J. Vega, E. Sánchez , A. Portas, A. Pereira , A. Mollinedo, J.A. Muñoz , M. Ruiz, E. Barrera, S. López, D. Machón, R. Castro, D. López. **Overview of the TJ-II remote participation system.** 5th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 12 - 15 July 2005 , Budapest , Hungary. *Fusion Engineering and Design, Vol 81, Issues 15-17, pp 2045-2050. July 2006.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2006.04.015

[Vega et al., 2006] J. Vega, E. Sánchez, A. Portas, A. Pereira, A. López, E. Ascasíbar, S. Balme, Y. Buravand, P. Lebourg, J. M. Theis, N. Utzel, M. Ruiz, E. Barrera, S. López, D. Machón, R. Castro, D. López, A. Mollinedo, J. A. Muñoz. **TJ-II Operation Tracking from Cadarache.** 15th International Stellarator Workshop. October 3 - 7, 2005, Madrid (Spain). *Fusion Science and Technology, Vol. 50, nº 3, pp 464-471. Octubre 2006.* ISSN: 1535-1055
http://www.ans.org/pubs/journals/fst/va-50-3-464-471

[Vega et al., 2007] J. Vega, G. Rattá, A. Murari, P. Castro, S. Dormido-Canto, R. Dormido, G. Farias, A. Pereira, A. Portas, E. de la Luna, I. Pastor, J. Sánchez, N. Duro, R. Castro, M. Santos, H. Vargas. **Recent results on structural pattern recognition for Fusion massive databases.** Proceedings of the IEEE International Symposium on Intelligent Signal Processing, WISP 3-5, Alcalá de Henares (Spain), Oct. 2007. *IEEE Transactions on Instrumentation and Measurement.. 3-5 October 2007.* E-ISBN: 978-1-4244-0830-6, Print ISBN: 978-1-4244-0829-0
http://dx.doi.org/10.1109/WISP.2007.4447569

[Vega et al., 2007b] J. Vega, M. Ruiz, E. Sánchez, A. Pereira, A. Portas, E. Barrera. **Real-time lossless data compression techniques for long-pulse operation.** Proceedings of the 24th Symposium on Fusion Technology - SOFT-24, Warsaw, Poland, 11-15 September 2006. *Fusion Engineering and Design. Vol . 82, Issues 5-14, pp 1301-1307. October 2007.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2007.06.014

[Vega et al., 2008] J. Vega, A. Pereira, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, M. Santos, E. Sánchez, G. Pajares. **Data mining technique for fast retrieval of similar waveforms in Fusion massive databases.** *Fusion Engineering and Design. Vol 83, Issue 1, pp 132-139. January 2008.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2007.09.011

[Vega et al., 2008b] J. Vega**,** A. Murari, A. Pereira, A. Portas, P. Castro and JET-EFDA Contributors. **Intelligent technique to search for patterns within images in massive databases**. Proceedings of the HTPD High Temperature Plasma Diagnostic 2008, Albuquerque, New Mexico. 11-15 May 2008. *Review of Scientific Instruments. Vol. 79. Issue 10. 10F327 (2008).* ISSN: 0034-6748, E-ISSN: 1089-7623
http://dx.doi.org/10.1063/1.2955863

[Vega et al., 2008c] J. Vega, A. Murari, G. A. Rattá, P. Castro, A. Pereira, A. Portas, and JET-EFDA Contributors. **Structural Pattern Recognition Techniques for Data Retrieval in Massive Fusion Databases**. Burning Plasma Diagnostics: An International Conference. 24–28 September 2007. Varenna (Italy). *AIP Conf. Proc. 988, December, 2008. pp. 481-484.* ISBN: 978-0-7354-0507-3
http://dx.doi.org/10.1063/1.2905118

[Vega et al., 2009] J. Vega, A. Murari, A. Pereira, A. Portas, R. Castro and JET-EFDA Contributors. **Overview of intelligent data retrieval methods for waveforms and images in massive fusion databases.** Proceedings of the 25th Symposium on Fusion Technology - SOFT-25. Rostock, Germany 15–19 September. *Fusion Engineering and Design. Volume 84, Issues 7-11, June 2009, Pages 1916-1919.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2008.11.097

[Vega et al., 2010] J. Vega, A. Murari, A. Pereira, S. González, I. Pastor. **Accurate and reliable image classification by using conformal predictors in the TJ-II Thomson Scattering**. Proceedings of the 18th Topical Conference on High Temperature Plasma Diagnostics, May 16-20, 2010, Wildwood, New Jersey. *Review of Scientific Instruments. Vol. 81 Issue 10. 10E118 (2010).* ISSN: 0034-6748, E-ISSN: 1089-7623
http://dx.doi.org/10.1063/1.3478689

[Vega et al., 2012] J. Vega, A.Murari, S. González, A. Pereira, I. Pastor, and JET-EFDA Contributors. **Overview of statistically hedged prediction methods: from off-line to real-time data analysis.** Proceedings of the 8th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Francisco, California, United States of America 20-24 June 2011. *Fusion Engineering and Design. Vol. 87, Issue 12, December 2012, Pages 2072–2075.* ISSN: 0920-3796
http://dx.doi.org/10.1016/j.fusengdes.2012.01.037

[Vega et al., 2013] Jesús Vega, Andrea Murari, Sergio González, Augusto Pereira, Ignacio Pastor. **Spatial location of local perturbations in plasma emissivity derived from projections using conformal predictors.** 2nd International Conference Frontiers in Diagnostic Technologies. November 28-30, 2011. Frascati, Italy. *Nuclear Instruments and Methods in Physics Research Section A:*

Accelerators, Spectrometers, Detectors and Associated Equipment. Vol. 720, 21. August 2013, Pages 14–19. ISSN: 0168-9002
http://dx.doi.org/10.1016/j.nima.2012.12.046

[Vega et al., 2013b] J. Vega, A. Murari, R. Moreno, S. González, A. Pereira, S. Dormido-Canto, J. M. Ramírez, J. M. López, D. Alves and JET-EFDA Contributors. **Advanced data analysis techniques for event identification and prediction in plasma experiments.** International Conference on Research and Applications of Plasmas. Warsaw, Poland, September 2-6, 2013. *Physica Scripta* (*Submitted for publication*).

[Vega et al., 2013c] J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, J. M. Ramírez, J. M. López, D. Alves, G. Rattá, A. Pereira and JET-EFDA Contributors. **Real-time prediction of disruptions: results in JET and research lines for ITER.** 8th Workshop on Fusion Data Processing, Validation and Analysis. November 4-6, 2013 Ghent, Belgium. *Plasma Physics and Controlled Fusion.* (*Submitted for publication*)

[Vega et al., 2013d] J. Vega, S. Dormido-Canto, J.M. López, A. Murari, J.M. Ramírez, R. Moreno, M. Ruiz, D. Alves, R. Felton and JET EFDA contributors. **Results of the JET real-time disruption predictor in the ITER-like wall campaigns**, *Fus. Eng. Des. (2013).*

[Vega et al., 2014] J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, A. Acero and JET-EFDA Contributors. **Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks.** *Nuclear Fusion. 54 123001*
http://dx.doi.org/10.1088/0029-5515/54/12/123001

[Vega et al., 2014b] J. Vega, A. Murari, S. Dormido-Canto, D. Alves, G. Farias, J. M. López, R. Moreno, A. Pereira, J. M. Ramírez, G. Rattá and JET-EFDA Contributors. **Overview of real-time disruption prediction in JET: applicability to ITER.** 41st EPS Conference on Plasma Physics. Berlin (Germany), 23-27 June 2014. *Plasma Physics and Controlled Fusion.* (*Submitted for publication*)

[Vico, 2010] J. Vico. **Análisis masivo de datos: Aplicación a imágenes en Fusión**. Proyecto Fin de Carrera Ingeniería en Informática. UNED 2010.

[Vovk et al., 2005] V. Vovk, A. Gammerman, G. Shafer. **Algorithmic learning in a random world**. Springer. New York, 2005. ISBN-13: 000-0387001522

[Walker, 1999] Walker J.S. **A primer on wavelets and their scientific applications**. *Chapman & hall/CRC. University of Wisconsin. 1999.* ISBN: 1584887451

[Whitley et al., 1988] D. Whitley, J. Kauth. **GENITOR: A different genetic algorithm**. Proceedings of the Rocky Mountain Conference on Artificial Intelligence, Denver, CO, 118-130.

Anexos

A. Publicaciones y congresos

A.1. Revistas científicas indexadas

1. **Feature selection for disruption prediction from scratch in JET by using genetic algorithms and probabilistic predictors.** [Pereira et al., 2014]
   A. Pereira, J. Vega, R. Moreno, S. Dormido-Canto, G. Rattá and JET/EFDA Contributors.
   Fusion Engineering and Design.

2. **Distributed open environment for data retrieval based on pattern recognition techniques.** [Pereira et al., 2010]
   A. Pereira, J. Vega, R. Castro, A. Portas.
   Fusion Engineering and Design. Volume 85, Issues 3-4, July 2010, Pages 595-597.

3. **Optimized search strategies to improve structural pattern recognition techniques.** [Pereira et al., 2010b]
   A. Pereira, J. Vega, A. Portas, R. Castro, A. Murari and JET/EFDA Contributors.
   World Scientific Proceedings Series on Computer Engineering and Information Science. Computational Intelligence in Decision and Control. Vol. 1. (pp 405-410). September 2008.
   International Journal of Nuclear Knowledge Management. Vol. 4, No.1 pp. 18 – 24, 2010.

4. **Synchronization resources in heterogeneous environments: Time-sharing, real-time and Java.** [Pereira et al., 2006]
   A. Pereira, J. Vega, L. Pacios, E. Sánchez, A. Portas.
   Fusion Engineering and Design, Volume 81, Issues 15-17, July 2006. Pages 1869-1872.

5. **Overview of real-time disruption prediction in JET: applicability to ITER.** [Vega et al., 2014b]
   J. Vega, A. Murari, S. Dormido-Canto, D. Alves, G. Farias, J. M. López, R. Moreno, A. Pereira, J. M. Ramírez, G. Rattá and JET-EFDA Contributors.
   Plasma Physics and Controlled Fusion.
6. **Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks.** [Vega et al., 2014]
   J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, A. Acero and JET-EFDA Contributors.
   Nuclear Fusion

7. **Real-time prediction of disruptions: results in JET and research lines for ITER.** [Vega et al., 2013c]
   J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, J. M. Ramírez, J. M. López, D. Alves, G. Rattá, A. Pereira and JET-EFDA Contributors.
   Plasma Physics and Controlled Fusion

8. **Advanced data analysis techniques for event identification and prediction in plasma experiments.** [Vega et al., 2013b]
   J. Vega, A. Murari, R. Moreno, S. González, A. Pereira, S. Dormido-Canto, J. M. Ramírez, J. M. López, D. Alves and JET-EFDA Contributors.
   Physica Scripta

9. **Development of an efficient real-time disruption predictor from scratch on jet and implications for ITER.** [Dormido-Canto et al., 2013]
   S. Dormido-Canto, J. Vega, J. M. Ramírez, A. Murari, R. Moreno, J. M. López, A. Pereira and JET-EFDA Contributors.
   Nuclear Fusion, 53, Issue 11, 8pp, November 2013.

10. **Dynamics of flows and confinement in the TJ-II stellarator.** [Sánchez et al., 2013]
    J. Sánchez, D. Alegre, A. Alonso, J. Alonso, P. Álvarez, J. Arévalo, E. Ascasíbar, A. Baciero, D. Baião, E. Blanco, M. Borchardt, J. Botija, A. Bustos, E. de la Cal, I. Calvo, A. Cappa, D. Carralero, R. Carrasco, F. Castejón, R. Castro, G. Catalán, A.A. Chmyga, M. Chamorro, L. Eliseev, T. Estrada, F. Fernández, J.M. Fontdecaba, L. García, R. García-Gómez, P. García-Sánchez, S. da Graça, J. Guasp, R. Hatzky, J. Hernández, J. Hernanz, J. Herranz, C. Hidalgo, J.A. Jiménez, A. Jiménez-Denche, I. Kirpitchev, R. Kleiber, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Razola, A. Martín, F. Martín-Díaz, F. Martín-Hernández, A.B. Martín-Rojo, J. Martínez-Fernández, K.J. McCarthy, F. Medina, M. Medrano, L. Melón, A.V. Melnikov, P. Méndez, B. van Milligen, P. Monreal, M. Navarro, I.S. Nedzelskiy, M.A. Ochando, J. Olivares, E. Oyarzábal, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, A. Petrov, S. Petrov, A.B. Portas, E. Rincón, L. Ríos, C. Rodríguez, B. Rojo, J.A. Romero, A. Ros, M. Sánchez, E. Sánchez, E. Sánchez-Sarabia, K. Sarksian, J.A. Sebastián, C. Silva, E.R. Solano, A. Soleto, B. Sun, F.L. Tabarés, D. Tafalla, M. Tereshchenko, A. Tolkachev, J. Vega, G. Velasco, J.L. Velasco, G. Wolfers and B. Zurro.
    Nuclear Fusion. Vol. 53, Issue 10. October 2013.

11. **Spatial location of local perturbations in plasma emissivity derived from projections using conformal predictors.** [Vega et al., 2013]
    Jesús Vega, Andrea Murari, Sergio González, Augusto Pereira, Ignacio Pastor.
    Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. Vol. 720, 21. August 2013, Pages 14–19.

12. **Automatic Determination of L/H Transition Times in DIII-D Through a Collaborative Distributed Environment.** [Farias et al., 2012]

G. Farias, J. Vega, S. González, A. Pereira, X. Lee, D. Schissel, P. Gohil.
Fusion Engineering and Design. Vol. 87, Issue 12, December 2012, Pages 2081–2083.

13. **Automatic location of L/H transition times for physical studies with a large statistical basis.** [González et al., 2012]
S. González, J. Vega, A. Murari, A. Pereira, S. Dormido-Canto, J.M. Ramírez and JET-EFDA contributors.
Plasma Physics and Controlled Fusion 54, 065009, Issue 6, 19 pp. June 2012.

14. **Region selection and image classification methodology using a non-conformity measure.** [González et al., 2012b]
S. González, J. Vega, A. Pereira, I. Pastor.
Progress in Artificial Intelligence. September 2012, Vol. 1, Issue 3, pp 215-222.

15. **Automated Analysis of Edge Pedestal Gradient Degradation During ELMs.** [González et al., 2012c]
S. González, J. Vega, A. Murari, A. Pereira and JET-EFDA contributors.
Fusion Science and Technology. Volume 62. Number 3. November 2012. Pages 403-408.

16. **H/L transition time estimation in JET using conformal predictors.** [González et al., 2012d]
S. González, J. Vega, A. Murari, A. Pereira, S. Dormido-Canto, J.M. Ramírez and JET EFDA contributors.
Fusion Engineering and Design. Vol. 87, Issue 12, December 2012, Pages 2084–2086.

17. **Overview of statistically hedged prediction methods: from off-line to real-time data analysis.** [Vega et al., 2012]
J. Vega, A.Murari, S. González, A. Pereira, I. Pastor, and JET-EFDA Contributors.
Fusion Engineering and Design. Vol. 87, Issue 12, December 2012, Pages 2072–2075.

18. **Overview of TJ-II experiments.** [Sánchez et al., 2011]
J. Sánchez, M. Acedo, D. Alegre, A. Alonso, J. Alonso, P. Álvarez, J. Arévalo, E. Ascasíbar, A. Baciero, D. Baiao, L. Barrera, E. Blanco, J. Botija, A. Bustos, E. de la Cal, I. Calvo, A. Cappa, D. Carralero, R. Carrasco, B.A. Carreras, F. Castejón, R. Castro, G. Catalán, A.A. Chmyga, M. Chamorro, L. Eliseev, L. Esteban, T. Estrada, J.A. Ferreira, J.M. Fontdecaba, L. García, R. García-Gómez, J.M. García-Regaña, P. García-Sánchez, A. Gómez-Iglesias, S. González, J. Guasp, T. Happel, J. Hernanz, J. Herranz, C. Hidalgo, J.A. Jiménez, A. Jiménez-Denche, R. Jiménez-Gómez, I. Kirpitchev, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Razola, T. Madeira, F. Martín-Díaz, F. Martín-Hernández, A.B. Martín-Rojo, J. Martínez-Fernández, K.J. McCarthy, F. Medina, M. Medrano, L. Melón, A.V. Melnikov, P. Méndez, B. van Milligen, E. Mirones, A. Molinero, M. Navarro, I.S. Nedzelskiy, M. Ochando, J. Olivares, E. Oyarzábal, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, A. Petrov, S. Petrov, A. Portas, G. Rattá, J.M. Reynolds, E. Rincón, L. Ríos, C. Rodríguez, B. Rojo, J.A. Romero, A. Ros, M. Sánchez, E. Sánchez, G. Sánchez-Burillo, E. Sánchez-Sarabia, K. Sarksian, J.A. Sebastián, C. Silva, E.R. Solano, A. Soleto, F. Tabarés, D. Tafalla, J. Tera, A. Tolkachev, J. Vega, G. Velasco, J.L. Velasco, M. Weber, G. Wolfers and B. Zurro.
Nuclear Fusion. Vol. 51, Issue 9. August 2011.

19. **Securing MDSplus in a Multi-organization Environment.** [Castro et al., 2010]
R. Castro, J. Vega, T. Fredian, K. Purahoo, A. Pereira, A. Portas.
Fusion Engineering and Design. Volume 85, Issues 3–4, July 2010, Pages 614–617

20. **Real-time remote diagnostic monitoring test-bed in JET.** [Castro et al., 2010b]

R. Castro, K. Kneupner, J. Vega, G. De Arcas, J.M. López, K. Purahoo, A. Murari, A. Fonseca, A. Pereira, A. Portas and JET-EFDA Contributors.
Fusion Engineering and Design. Volume 85, Issues 3–4, July 2010, Pages 598–602

21. **Support vector machine based feature extractor for L/H transitions in JET**. [González et al., 2010]
S. González, J. Vega, A. Murari, A. Pereira, J.M. Ramírez, S. Dormido-Canto and JET-EFDA contributors.
Review of Scientific Instruments. Vol. 81 Issue 10. 10E123 (2010).

22. **Automatic ELM location in JET using a universal multi-event locator.** [González et al., 2010b]
S. González, J. Vega, A. Murari, A. Pereira, M. Beurskens and JET-EFDA contributors.
Fusion Science and Technology, Vol. 58, nº 3, pp 755-762. Nov 2010.

23. **Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: new image recognition classifier and fault condition detection.** [Makili et al., 2010]
L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, A. Portas, D. Pérez-Risco, M.C. Rodríguez-Fernández, P. Busch.
Fusion Engineering and Design. Volume 85, Issues 3-4, July 2010, Pages 415-418.

24. **Accurate and reliable image classification by using conformal predictors in the TJ-II Thomson Scattering**. [Vega et al., 2010]
J. Vega, A. Murari, A. Pereira, S. González, I. Pastor.
Review of Scientific Instruments. Vol. 81 Issue 10. 10E118 (2010).

25. **Confinement transitions in TJ-II under Li-coated wall conditions.** [Sánchez et al., 2009]
J. Sánchez, M. Acedo, A. Alonso, J. Alonso, P. Alvarez, E. Ascasíbar, A. Baciero, R. Balbín, L. Barrera, E. Blanco, J. Botija, A. de Bustos, E. de la Cal, I. Calvo, A. Cappa, J.M. Carmona, D. Carralero, R. Carrasco, B.A. Carreras, F. Castejón, R. Castro, G. Catalán, A.A. Chmyga, M. Chamorro, L. Eliseev, L. Esteban, T. Estrada, A. Fernández, R. Fernández-Gavilán, J.A. Ferreira, J.M. Fontdecaba, C. Fuentes, L. García, I. García-Cortés, R. García-Gómez, J.M. García-Regaña, J. Guasp, L. Guimarais, T. Happel, J. Hernanz, J. Herranz, C. Hidalgo, J.A. Jiménez, A. Jiménez-Denche, R. Jiménez-Gómez, D. Jiménez-Rey, I. Kirpitchev, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Rázola, A. López-Sánchez, S. Lysenko, G. Marcon, F. Martín, V. Maurin, K.J. McCarthy, F. Medina, M. Medrano, A.V. Melnikov, P. Méndez, B. van Milligen, E. Mirones, I.S. Nedzelskiy, M. Ochando, J. Olivares, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, G. Pérez, D. Pérez-Risco, A. Petrov, S. Petrov, A. Portas, D. Pretty, D. Rapisarda, G. Rattá, J.M. Reynolds, E. Rincón, L. Ríos, C. Rodríguez, J.A. Romero, A. Ros, A. Salas, M. Sánchez, E. Sánchez, E. Sánchez-Sarabia, K. Sarksian, J.A. Sebastián, C. Silva, S. Schchepetov, N. Skvortsova, E.R. Solano, A. Soleto, F. Tabarés, D. Tafalla, A. Tarancón, Yu. Taschev, J. Tera, A. Tolkachev, V. Tribaldos, V.I. Vargas, J. Vega, G. Velasco, J.L. Velasco, M. Weber, G. Wolfers, B. Zurro.
Nuclear Fusion, Vol. 49, Issue 10, 104018. October 2009.

26. **Data distribution architecture based on standard Real Time Protocol.** [Castro et al., 2009]
R. Castro, J. Vega, A. Pereira, A. Portas.
Fusion Engineering and Design. Volume 84, Issues 2-6, June 2009, Pages 565-568.

27. **Overview of intelligent data retrieval methods for waveforms and images in massive fusion databases.** [Vega et al., 2009]
J. Vega, A. Murari, A. Pereira, A. Portas, R. Castro and JET-EFDA Contributors.

Fusion Engineering and Design. Volume 84, Issues 7-11, June 2009, Pages 1916-1919.

28. EFDA-fed: European federation among fusion energy research laboratories. [Castro et al., 2008b]
R. Castro, J. Vega, A. Portas, A. Pereira, C. Rodriguez, S. Balme, J. M. Theis, J. Signoret, P. Lebourg, K. Purahoo, K.Thomsen, H. Fernandes, A. Neto, A. Duarte, F. Oliveira, F. Reis, J. Kadlecsik.
Campus-Wide Information Systems. Vol 25, Number 5, 2008, pp. 359-373, Emerald Group Publishing Limited.

29. Structural pattern recognition methods based on string comparison for fusion databases. [Dormido-Canto et al., 2008]
S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, N. Duro, H. Vargas, G. Rattá, A. Pereira, A. Portas.
Fusion Engineering and Design. Vol 83, Issues 2-3, pp 421-424. April 2008.

30. An event-oriented database for continuous data flows in the TJ-II environment. [Sánchez et al., 2008]
E. Sánchez, A. de la Peña, A. Portas, A. Pereira, J. Vega, A. Neto, H. Fernandes.
Fusion Engineering and Design. Vol 83, Issues 2-3, pp 413-416. April 2008.

31. Data mining technique for fast retrieval of similar waveforms in Fusion massive databases. [Vega et al., 2008]
J. Vega, A. Pereira, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, M. Santos, E. Sánchez, G. Pajares.
Fusion Engineering and Design. Vol 83, Issue 1, pp 132-139. January 2008.

32. Intelligent technique to search for patterns within images in massive databases. [Vega et al., 2008b]
J. Vega**,** A. Murari, A. Pereira, A. Portas, P. Castro and JET-EFDA Contributors.
Proceedings of the HTPD High Temperature Plasma Diagnostic 2008, Albuquerque, New Mexico. 11-15 May 2008.
Review of Scientific Instruments. Vol. 79. Issue 10. 10F327 (2008)

33. First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET. [Rattá et al., 2008]
G. A. Rattá, J. Vega, A. Pereira, A. Portas, E. de la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari.
Fusion Engineering and Design. Vol 83, Issues 2-3, pp 467-470. April 2008.

34. Structural Pattern Recognition Techniques for Data Retrieval in Massive Fusion Databases. [Vega et al., 2008c]
J. Vega, A. Murari, G. A. Rattá, P. Castro, A. Pereira, A. Portas, and JET-EFDA Contributors.
AIP Conf. Proc. 988, pp. 481-484.

35. Recent results on structural pattern recognition for Fusion massive databases. [Vega et al., 2007]
J. Vega, G. Rattá, A. Murari, P. Castro, S. Dormido-Canto, R. Dormido, G. Farias, A. Pereira, A. Portas, E. de la Luna, I. Pastor, J. Sánchez, N. Duro, R. Castro, M. Santos, H. Vargas.
IEEE Transactions on Instrumentation and Measurement.. 3-5 October 2007.

36. Remote control of data acquisition devices by means of message oriented middleware. [Sánchez et al., 2007b]

E. Sánchez, A. Portas, A. Pereira, J. Vega, I. Kirpitchev.
Fusion Engineering and Design. Vol . 82, Issues 5-14, pp 1365-1371. October 2007.

37. Real-time lossless data compression techniques for long-pulse operation. [Vega et al., 2007b]
J. Vega, M. Ruiz, E. Sánchez, A. Pereira, A. Portas, E. Barrera.
Fusion Engineering and Design. Vol . 82, Issues 5-14, pp 1301-1307. October 2007.

38. Overview of TJ-II experiments. [Sánchez et al., 2007]
J. Sánchez, M. Acedo, A. Alonso, J. Alonso, P. Alvarez, F. de Aragón, E. Ascasíbar, A. Baciero, R. Balbín, L. Barrera, E. Blanco, J. Botija, B. Brañas, E. de la Cal, E. Calderón, I. Calvo, A. Cappa, J.A. Carmona, B.A. Carreras, R. Carrasco, F. Castejón, G. Catalán, A.A. Chmyga, N.B. Dreval, M. Chamorro, S. Eguilior, J. Encabo, L. Eliseev, T. Estrada, A. Fernández, R. Fernández, J.A. Ferreira, J.M. Fontdecaba, C. Fuentes, J. de la Gama, A. García, L. García, I. García-Cortés, J.M. García-Regaña, B. Gonçalves, J. Guasp, J. Herranz, A. Hidalgo, C. Hidalgo, R. Jiménez-Gómez, J.A. Jiménez, D. Jiménez, I. Kirpitchev, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Rázola, A. López-Sánchez, E. de la Luna, G. Marcon, F. Martín, L. Martínez-Fresno, K.J. McCarthy, F. Medina, M. Medrano, A.V. Melnikov, P. Méndez, E. Mirones, B. van Milligen, I.S. Nedzelskiy, M. Ochando, J. Olivares, R. Orozco, P. Ortiz, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, D. Pérez-Risco, A. Petrov, S. Petrov, A. Portas, D. Rapisarda, L. Ríos, C. Rodríguez, L. Rodríguez-Rodrigo, E. Rodríguez-Solano, J. Romero, A. Ros, A. Salas, E. Sánchez, M. Sánchez, E. Sánchez-Sarabia, X. Sarasola, K. Sarksian, C. Silva, S. Schchepetov, N. Skvortsova, A. Soleto, F. Tabarés, D. Tafalla, J. Tera, A. Tolkachev, V. Tribaldos, V.I. Vargas, J. Vega, G. Velasco, M. Weber, G. Wolfers, S.J. Zweben, B. Zurro.
Nuclear Fusion, Vol. 47, nº 10, pp 677-685. October 2007.

39. Applying a message oriented middleware architecture to the TJ-II remote participation system. [Sánchez et al., 2006]
E. Sánchez , A. Portas, A. Pereira, J. Vega.
Fusion Engineering and Design, Vol 81, Issues 15-17, pp 2063–2067. July 2006.

40. TJ-II Operation Tracking from Cadarache. [Vega et al., 2006]
J. Vega, E. Sánchez, A. Portas, A. Pereira, A. López, E. Ascasíbar, S. Balme, Y. Buravand, P. Lebourg, J. M. Theis, N. Utzel, M. Ruiz, E. Barrera, S. López, D. Machón, R. Castro, D. López, A. Mollinedo, J. A. Muñoz.
Fusion Science and Technology, Vol. 50, nº 3, pp 464-471. Octubre 2006.

41. Overview of the TJ-II remote participation system. [Vega et al., 2005b]
J. Vega, E. Sánchez , A. Portas, A. Pereira , A. Mollinedo, J.A. Muñoz , M. Ruiz, E. Barrera, S. López, D. Machón, R. Castro, D. López.
Fusion Engineering and Design, Vol 81, Issues 15-17, pp 2045-2050. July 2006.

42. Overview of TJ-II experiments. [Hidalgo et al., 2005Hidalgo et al., 2005Hidalgo et al., 2005]
C. Hidalgo, C. Alejaldre, A. Alonso, J. Alonso, L. Almoguera, F. de Aragón, E. Ascasíbar, A. Baciero, R. Balbín, E. Blanco, J. Botija, B. Brañas, E. Calderón, A. Cappa, J.A. Carmona, R. Carrasco, F. Castejón, J.R. Cepero, A.A. Chmyga, J. Doncel, N.B. Dreval, S. Eguilior, L. Eliseev, T. Estrada, J.A. Ferreira, A. Fernández, J.M. Fontdecaba, C. Fuentes, A. García, I. García-Cortés, B. Gonçalves, J. Guasp, J. Herranz, A. Hidalgo, R. Jiménez, J.A. Jiménez, D. Jiménez-Rey, I. Kirpitchev, S.M. Khrebtov, A.D. Komarov, A.S. Kozachok, L. Krupnik, F. Lapayese, M. Liniers, D. López-Bruna, A. López-Fraguas, J. López-Rázola, A. López-Sánchez, E. de la Luna, G. Marcon, R. Martín, K.J. McCarthy, F. Medina, M. Medrano, A.V. Melnikov, P. Méndez, B. van Milligen, I.S.

Nedzelskiy, M. Ochando, O. Orozco, J.L. de Pablos, L. Pacios, I. Pastor, M.A. Pedrosa, A. de la Peña, A. Pereira, A. Petrov, S. Petrov, A. Portas, D. Rapisarda, L. Rodríguez-Rodrigo, E. Rodríguez-Solano, J. Romero, A. Salas, E. Sánchez, J. Sánchez, M. Sánchez, K. Sarksian, C. Silva, S. Schchepetov, N. Skvortsova, F. Tabarés, D. Tafalla, A. Tolkachev, V. Tribaldos, I. Vargas, J. Vega, G. Wolfers, B. Zurro.
Nuclear Fusion, Vol. 45, nº 10, pp 266-275. October 2005.

43. **Application of intelligent classification techniques to the TJ-II Thomson Scattering diagnostic.** [Vega et al., 2005]
J. Vega, I. Pastor, J. L. Cereceda, A. Pereira, J. Herranz, D. Pérez, M. C. Rodríguez, G. Farias, S. Dormido-Canto, J. Sánchez, R. Dormido, N. Duro, S. Dormido, G. Pajares, M. Santos, J. M. de la Cruz.
Europhysics Conference Abstracts (ECA), Vol.29C, P-2.090 (2005)

A.2. Revistas de divulgación e informes técnicos

1. **Desarrollo de entornos cliente para un sistema de sincronización basado en eventos**. [Pereira y Vega, 2005]
A. Pereira, J. Vega.
Informes Técnicos Ciemat, nº 1064, Octubre 2005, 172 pp.

2. **Control and data acquisition system for the Spanish Beamline (BM25) at the ESRF**. [Pereira et al., 2005]
A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.
Informes Técnicos Ciemat, nº 1056, Abril 2005, 108 pp.

3. **Control y adquisición de datos en la línea española del sincrotrón de Grenoble.** [Pereira et al., 2005b]
A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.
Revista Técnica Industrial, nº 260, Diciembre 2005, pp 40-43.

4. **A flexible VME-PowerPC-Linux Device Driver for MEN A201S carrier board and MEN M37 mezzanine module**. [Pereira et al., 2004d]
A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.
Linux Gazette, Issue 107, October 2004.

5. **Adding 16 bits-ADC extensions on VME bus with MEN M36 Linux Device Driver**. [Pereira et al., 2004e]
A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.
Linux Gazette, Issue 107, October 2004.

6. **Compiling the Linux kernel 2.2.12 on VME-PowerPC architecture**. [Pereira et al., 2004b]
A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.
Linux Gazette, Issue 103, June 2004.

7. **Configuring a VME-Linux Diskless System on Motorola MVME24xx**. [Pereira et al., 2004c]
A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.
Linux Gazette, Issue 103, June 2004.

8. **Seguridad y privacidad en los sistemas de pagos.** [Pereira, 2002]
A. Pereira.
Revista Técnica Industrial, nº 244, Marzo 2002, pp 70-75.

9. **Control de transductores de posición en tiempo real.** [Pereira, 2001]
A. Pereira.
Revista Técnica Industrial, nº 242, Septiembre 2001, pp 70-73.

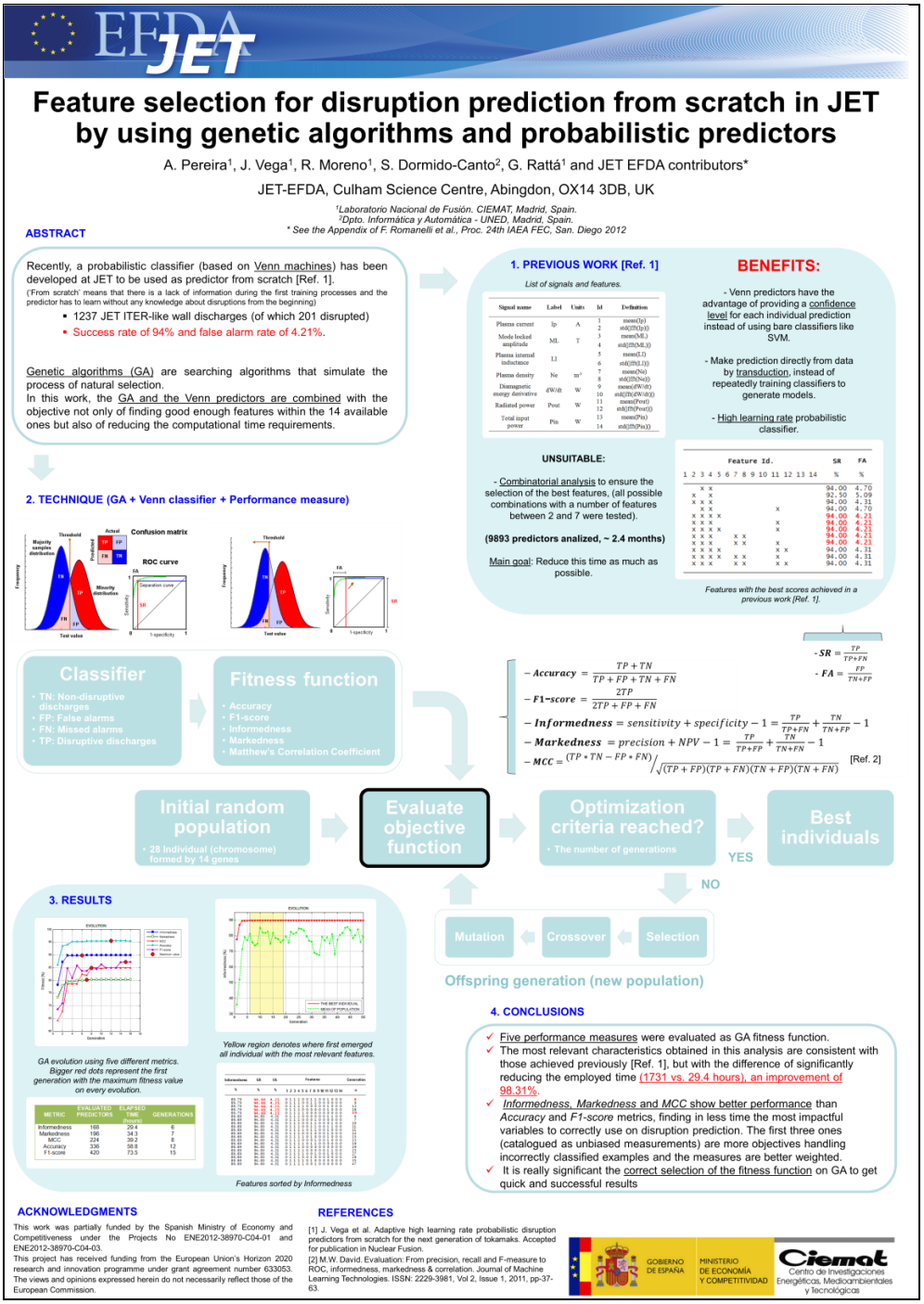
10. **Fundamentos de los criptosistemas de clave pública para internet.** [Pereira, 2000]
A. Pereira.
Revista Técnica Industrial, nº 238, Septiembre 2000, pp 40-44.

11. **EFDA-Fed: Una federación internacional para investigación en fusión basada en PAPI**. [Castro et al., 2008]
Rodrigo Castro, Jesús Vega, Ana Portas, Augusto Pereira, Diego R. López.
Publicación de la red nacional de I+D. Boletín RedIris, nº 82-83, pp 17-23,Abril 2008

12. **Visual data analysis in the TJ-II remote participation system**. [Sánchez et al., 2006b]
E. Sánchez, A. Portas, A. Pereira, J. Vega.
Informes Técnicos Ciemat, nº 1094, Noviembre 2006, 20 pp,

A.3. Actas en conferencias y aportaciones a congresos

1. **Feature selection for disruption prediction from scratch in JET by using genetic algorithms and probabilistic predictors.** [Pereira et al., 2014]
   A. Pereira, J. Vega, R. Moreno, S. Dormido-Canto, G. Rattá and JET/EFDA Contributors.
   Proceedings of the 28th Symposium on Fusion Technology. San Sebastián, Spain. September 29th – October 3rd. SOFT 2014.



2. **Distributed open environment for data retrieval based on pattern recognition techniques.** [Pereira et al., 2010]
   A. Pereira, J. Vega, R. Castro, A. Portas.
   Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research, 15-19 June 2009. Aix-In-Provence (France).

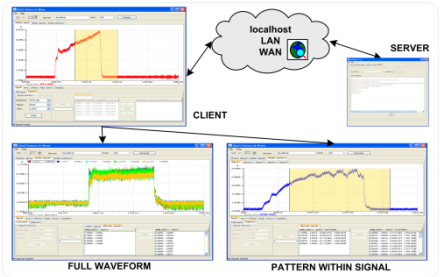
Distributed open environment for data retrieval based on pattern recognition techniques

A. Pereira, J. Vega, R. Castro, A. Portas
Asociación EURATOM/CIEMAT, Avda. Complutense 22, 28040 Madrid, Spain

Abstract

Pattern recognition methods for data retrieval have been applied to fusion databases for the localization and extraction of similar waveforms within temporal evolution signals. In order to standardize the use of these methods, a distributed open environment has been designed. It is based on a client/server architecture that supports distribution, interoperability and portability between heterogeneous platforms. The server part is a single desktop application based on J2EE, which provides a mature standard framework and a modular architecture. It can handle transactions and concurrency of components that are deployed on JETTY, an embedded web container within the Java server application for providing HTTP services. The data management is based on Apache DERBY, a relational database engine also embedded on the same Java based solution. This encapsulation allows hiding of unnecessary details about the installation, distribution, and configuration of all these components but with the flexibility to create and allocate many databases on different servers. The DERBY network module increases the scope of the installed database engine by providing traditional Java database network connections (JDBC-TCP/IP). This avoids scattering several database engines (a unique embedded engine defines the rules for accessing the distributed data). Java thin clients (Java 5 or above is the unique requirement) can be executed in the same computer than the server program (for example a desktop computer) but also server and client software can be distributed in a remote participation environment (wide area networks). The thin client provides graphic user interface to look for patterns (entire waveforms or specific structural forms) and display the most similar ones. This is obtained with HTTP requests and by generating dynamic content (servlets) in response to these client requests.

Pattern recognition tool



Search and retrieval of similar patterns and full waveforms.

Distributed architecture

Software embedded frameworks for remote computing environments

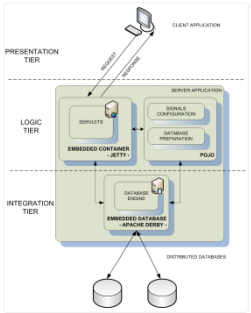
Embedded frameworks:

- Apache-Derby database: http://db.apache.org/derby/

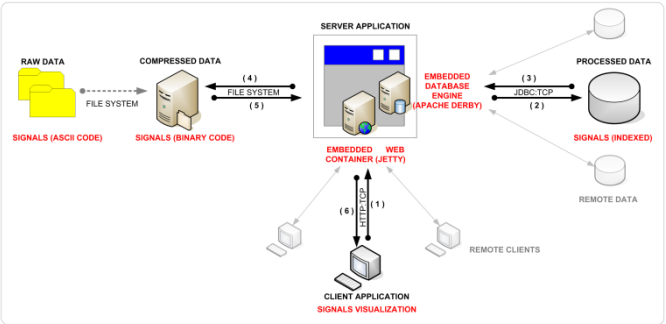
Apache Derby

- Jetty servlet container: http://www.mortbay.org/jetty/

Powered by jetty://



Operation environment



(1) CLIENT REQUEST
URL statement (Uniform Resource Locator) that specifies where the resource (database) is available and the mechanism (signal information) for retrieving it.

(2) DATABASE ENGINE REQUEST
Query over the remote database.

(3) REMOTE DATABASE RESPONSE
JDBC row result set.

(4, 5) RETRIEVAL OF RAW SIGNALS
Raw data for each signal are recovered from the locations of files repository .

(6) SERVLET REPLY
Each signal is sent back to the client with dynamic content.

Conclusions

- It avoids complex software installations (DB, http server), encapsulating everything in a single desktop application.
- It allows hiding these modules as well as other details related to software configuration and software distribution (prevent access to unauthorized setup options, prevent entering of incorrect arguments about software setup, exclude installation options that are not desired, less time wasted fixing external software).

Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA E INNOVACIÓN · Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

3. **Optimized search strategies to improve structural pattern recognition techniques.**
[Pereira et al., 2010b]
A. Pereira, J. Vega, A. Portas, R. Castro, A. Murari and JET/EFDA Contributors.
Proceedings of the 8th International FLINS Conference. September 21-24, 2008, Madrid (Spain).

Optimized search strategies to improve structural pattern recognition techniques

A. PEREIRA[1], J. VEGA[1], A. PORTAS[1], R. CASTRO[1], A. MURARI[2] AND JET-EFDA CONTRIBUTORS[3]

JET_EFDA, Culham Science Center, OX14 3DB, Abingdon, UK
[1]Asociación EURATOM/CIEMAT, Avda. Complutense 22, 28040 Madrid, Spain
[2]Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padua, Italy
[3]See the Appendix of M.L.Watkins et al., Fusion Energy 2006 (Proc. 21st Int. Conf. Chengdu, 2006) IAEA, (2006)

Abstract

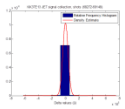
This paper provides a new solution to the localization and extraction of similar patterns in time-series data. Alternative searches are proposed to objectively increase the recognition of similar patterns so as to achieve better results on the data retrieval. Patterns are represented by string of characters. Looking for patterns means looking for characters. Thinner search strategies have been studied with excellent results in the detection of long subpatterns. Long subpatterns are not so easy to identify since even a single mismatch in one character can compromise similarity between two patterns. Identifying long patterns in a fast, fault tolerant and intelligent way is the aim of the analyzed strategies, formally based on statistical criteria and some aspects of probability theory.

1. Optimizing the pre-process

a) Data-Base modification.

- **Just 2 primitives** (labelled according to the sign value of the delta transformation) is the essential difference with regard to previous developments [1]. The probability that the longest pattern will happen is $\Omega = 1/2^{64}$ vs $\Omega = 1/4^{128}$ in a preceding approach [2].

b) Statistical analysis. (all set of deltas)



-To determine the range of the δ variation where the primitives can take any value (number of cases that fall into the central category, indistinctly 'a' and 'b', near to zero).

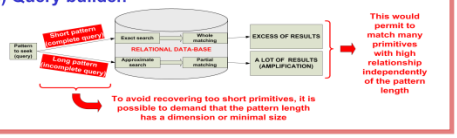
-The choice **[-σ/32 , σ/32] as central boundaries** is suitable to any kind of signal.

2. Defining similarity queries

a) Alternative searches.

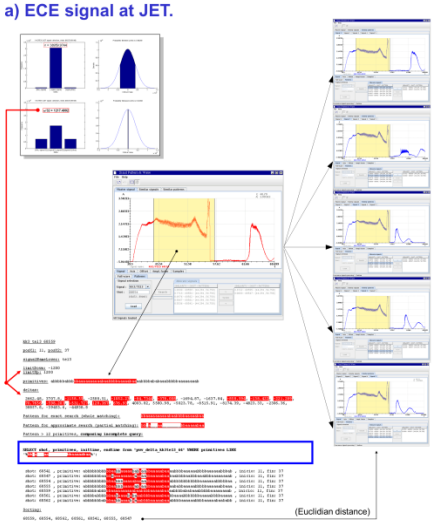
- The recognition problem is translated into a character-matching problem.

b) Query builder.

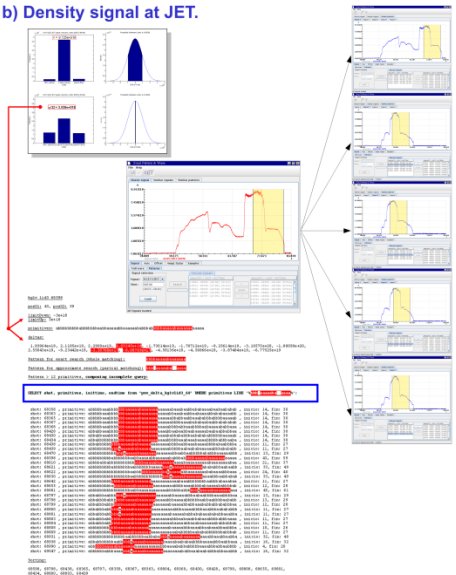


3. Improving the data retrieval

a) ECE signal at JET.



b) Density signal at JET.



Conclusions:

- The possibility to locate a lot of similar subsequences is less dependent on the pattern length.

- Stable decision on the primitive assignment (suitable to any kind of signal).

References:

- [1] J. Vega, A. Murari, et al., "Structural Pattern Recognition Techniques for Data Retrieval in Fusion Massive Databases", International Workshop on Burning Plasma Diagnostics. Villa Monastero, Varenna, Italy. 24-28 September (2007).
- [2] S. Dormido-Canto, G. Farias, et al., "Search and retrieval of plasma waveforms: structural pattern recognition approach". Rev. Sci. Ins. 77 (2006). 10F514.

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA E INNOVACIÓN · Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

4. **Synchronization resources in heterogeneous environments: Time-sharing, real-time and Java.** [Pereira et al., 2006]
   A. Pereira, J. Vega, L. Pacios, E. Sánchez, A. Portas.
   Proceedings of the 5th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 12 - 15 July 2005; Budapest (Hungary).

Synchronization resources in heterogeneous environments: time-sharing, real-time and JAVA

A. Pereira, J. Vega, L. Pacios, E. Sánchez, A. Portas
Asociación EURATOM/CIEMAT para Fusión. CIEMAT, Edificio 66, Avda. Complutense, 22, 28040 Madrid, Spain

Abstract

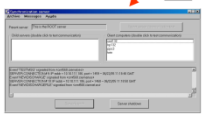
The Asynchronous Event Distribution System (AEDS) was built to provide synchronization resources within the TJ-II local area network. It is a software system developed to add "soft synchronization" capabilities to the TJ-II data acquisition, control and analysis environments. "Soft synchronization" signifies that AEDS is not a real-time system. In fact, AEDS is based on TCP/IP over ETHERNET networks. However, its response time is adequate for practical purposes when synchronization requirements can support some delay between event dispatch and message reception. Event broadcasters (or synchronization servers in AEDS terminology) are Windows computers. Destination computers (or synchronization clients) were also Windows machines in the first version of AEDS. However, this fact imposed a very important limitation on synchronization capabilities. To overcome this situation, synchronization clients for different environments have been added to AEDS: time-sharing operating systems (UNIX and LINUX), real-time operating systems (OS9 and VxWorks) and Java applications. The synchronization primitives that operate in these systems are very different between them and, therefore, several approaches were chosen in order to provide the same functionality to the various environments. POSIX thread library with its basic synchronization primitives (mutex and condition variables) was used to accomplish this task on UNIX/LINUX systems, IPC mechanisms for concurrent processes on OS9 and VxWorks real time operating systems, and 'synchronized - wait/notify' primitives on Java virtual machines.

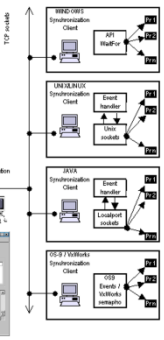
Porting to new environments

New client platforms were added to the distributed synchronization system [1] of the TJ-II experimental local area network.

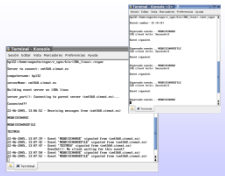
TJ-II Stellerator

Event source





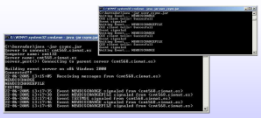
Unix/Linux systems



The local processes layer has been based on a previous development [2] that provides only inter-thread communication capabilities within a single process in Solaris environments. We have added inter-process communications to be able to synchronize any local process. The development is based on the POSIX threads API which allows us to reuse the programs in multiple platforms. The synchronization functions emulate the WaitFor methods of the Windows API. The clients have been tested in Unix HP-tru64 and Linux kernel 2.6 (Suse 9.1 and Red Hat Enterprise WS).

Java environments

Java Synchronized mechanism has been used to control concurrent access to objects and variables. Inter-thread communication has been solved implementing the safe and efficient wait/notify methods. Localport sockets were used to get inter-process local communication and Berkeley TCP sockets to communicate with the server event dispatch program



Real-time systems

OS-9

The emulation of win32 WaitFor methods in OS-9 systems has been obtained by means of OS9-events that is a 32-byte system global variable maintained by the system. OS9 events are multiple-value semaphores and are wonderfully versatile. They synchronize concurrent processes that are accesing shared resources.

VxWorks

The best model for event processing in VxWorks real-time applications is the semLib semaphores library. The routines semTake() and semFlush() were used to implement the VxWorks synchronization client.

Conclusions

AEDS synchronization clients have been developed for multiple platforms:

- UNIX (Sun-Solaris and hp-tru64)
- LINUX kernel 2.6.4 (SuSe and Red Hat)
- JAVA (J2 sdk 1.4.2_07)
- OS-9 (M68k-VME)
- VxWorks (PowerPC-VME diskless, MVME5500 card)

AEDS can be used in heterogeneous environments.

The synchronization capabilities are provided through a set of functions grouped in a software library (one library for each operating system).

References:

[1] J. Vega, et al., A distributed synchronization system for the TJ-II local area network, Fusion Eng. 71 (2004) 117-221.

[2] N. Nagarajayya, A. Gupta., Porting of Win32 API WaitFor to Solaris. http://developers.sun.com/solaris/articles/waitfor_api.html (2000).

Acknowledgements:

This work is partially funded by the Spanish Ministry of Education and Science under the Project No. ENE2004-07335. Special thanks are addressed to Fernando Lapayese, Angel De la Peña and Ricardo Carrasco during the development of OS-9 and VxWorks environments.

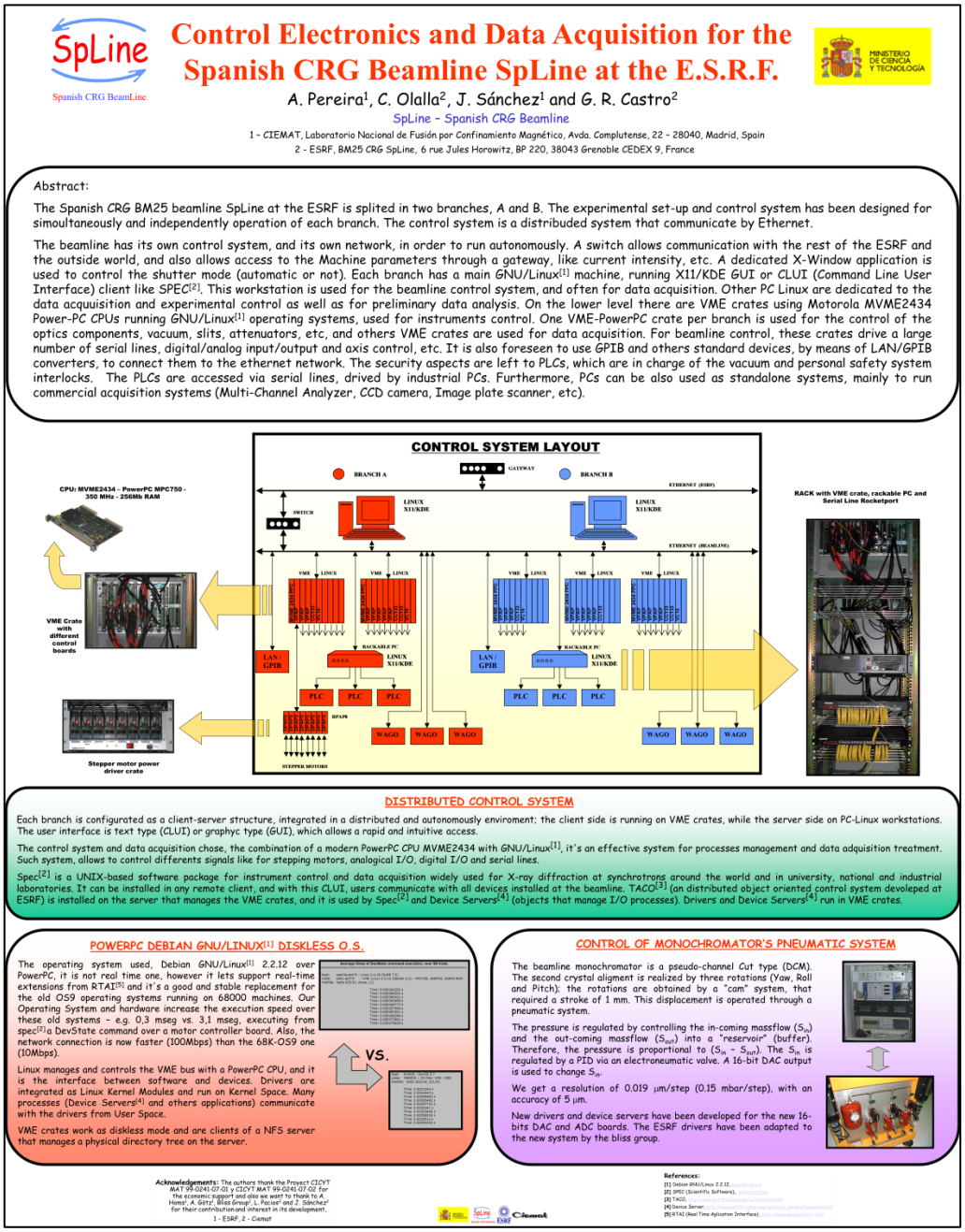
MINISTERIO DE EDUCACIÓN Y CIENCIA

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

5. Control Electronics and Data Acquisition for the Spanish CRG Beamline SpLine at the E.S.R.F. [Pereira et al., 2004]

A. Pereira, C. Olalla, J. Sánchez, G.R. Castro.

Proceedings 1ª Reunión Nacional de Usuarios de Radiación Sincrotrón, Torremolinos (Málaga), 5-6 Febrero 2004.



SpLine
Spanish CRG BeamLine

Control Electronics and Data Acquisition for the Spanish CRG Beamline SpLine at the E.S.R.F.

MINISTERIO DE CIENCIA Y TECNOLOGÍA

A. Pereira[1], C. Olalla[2], J. Sánchez[1] and G. R. Castro[2]

SpLine - Spanish CRG Beamline

1 - CIEMAT, Laboratorio Nacional de Fusión por Confinamiento Magnético, Avda. Complutense, 22 - 28040, Madrid, Spain

2 - ESRF, BM25 CRG SpLine, 6 rue Jules Horowitz, BP 220, 38043 Grenoble CEDEX 9, France

Abstract:

The Spanish CRG BM25 beamline SpLine at the ESRF is splited in two branches, A and B. The experimental set-up and control system has been designed for simoultaneously and independently operation of each branch. The control system is a distribuded system that communicate by Ethernet.

The beamline has its own control system, and its own network, in order to run autonomously. A switch allows communication with the rest of the ESRF and the outside world, and also allows access to the Machine parameters through a gateway, like current intensity, etc. A dedicated X-Window application is used to control the shutter mode (automatic or not). Each branch has a main GNU/Linux[1] machine, running X11/KDE GUI or CLUI (Command Line User Interface) client like SPEC[2]. This workstation is used for the beamline control system, and often for data acquisition. Other PC Linux are dedicated to the data acquisition and experimental control as well as for preliminary data analysis. On the lower level there are VME crates using Motorola MVME2434 Power-PC CPUs running GNU/Linux[1] operating systems, used for instruments control. One VME-PowerPC crate per branch is used for the control of the optics components, vacuum, slits, attenuators, etc, and others VME crates are used for data acquisition. For beamline control, these crates drive a large number of serial lines, digital/analog input/output and axis control, etc. It is also foreseen to use GPIB and others standard devices, by means of LAN/GPIB converters, to connect them to the ethernet network. The security aspects are left to PLCs, which are in charge of the vacuum and personal safety system interlocks. The PLCs are accessed via serial lines, drived by industrial PCs. Furthermore, PCs can be also used as standalone systems, mainly to run commercial acquisition systems (Multi-Channel Analyzer, CCD camera, Image plate scanner, etc).

DISTRIBUTED CONTROL SYSTEM

Each branch is configurated as a client-server structure, integrated in a distributed and autonomously enviroment; the client side is running on VME crates, while the server side on PC-Linux workstations. The user interface is text type (CLUI) or graphyc type (GUI), which allows a rapid and intuitive access.

The control system and data acquisition chose, the combination of a modern PowerPC CPU MVME2434 with GNU/Linux[1], it's an effective system for processes management and data adquisition treatment. Such system, allows to control differents signals like for stepping motors, analogical I/O, digital I/O and serial lines.

Spec[2] is a UNIX-based software package for instrument control and data acquisition widely used for X-ray diffraction at synchrotrons around the world and in university, national and industrial laboratories. It can be installed in any remote client, and with this CLUI, users communicate with all devices installed at the beamline. TACO[3] (an distributed object oriented control system devoleped at ESRF) is installed on the server that manages the VME crates, and it is used by Spec[2] and Device Servers[4] (objects that manage I/O processes). Drivers and Device Servers[4] run in VME crates.

POWERPC DEBIAN GNU/LINUX[1] DISKLESS O.S.

The operating system used, Debian GNU/Linux[1] 2.2.12 over PowerPC, it is not real time one, however it lets support real-time extensions from RTAI[5] and it's a good and stable replacement for the old OS9 operating systems running on 68000 machines. Our Operating System and hardware increase the execution speed over these old systems - e.g. 0,3 mseg vs. 3,1 mseg, executing from spec[2] a DevState command over a motor controller board. Also, the network connection is now faster (100Mbps) than the 68K-OS9 one (10Mbps).

VS.

Linux manages and controls the VME bus with a PowerPC CPU, and it is the interface between software and devices. Drivers are integrated as Linux Kernel Modules and run on Kernel Space. Many processes (Device Servers[4] and others applications) communicate with the drivers from User Space.

VME crates work as diskless mode and are clients of a NFS server that manages a physical directory tree on the server.

CONTROL OF MONOCHROMATOR'S PNEUMATIC SYSTEM

The beamline monochromator is a pseudo-channel Cut type (DCM). The second crystal aligment is realized by three rotations (Yaw, Roll and Pitch): the rotations are obtained by a "cam" system, that required a stroke of 1 mm. This displacement is operated through a pneumatic system.

The pressure is regulated by controlling the in-coming massflow (S_{in}) and the out-coming massflow (S_{out}) into a "reservoir" (buffer). Therefore, the pressure is proportional to (S_{in} - S_{out}). The S_{in} is regulated by a PID via an electroneumatic valve. A 16-bit DAC output is used to change S_{in}.

We get a resolution of 0.019 µm/step (0.15 mbar/step), with an accuracy of 5 µm.

New drivers and device servers have been developed for the new 16-bits DAC and ADC boards. The ESRF drivers have been adapted to the new system by the bliss group.

Acknowledgements: The authors thank the Proyect CICYT MAT 99-0241-07-01 y CICYT MAT 99-0241-07-02 for the economic support and also we want to thank to A. Homs', A. Götz', Bliss Group', L. Pacios" and J. Sánchez' for their contribution and interest in its development.

1 - ESRF, 2 - Ciemat

References:

[1] Debian GNU/Linux 2.2.12, http://www.debian.org

[2] SPEC (Scientific Software), http://www.certif.com

[3] TACO, http://www.esrf.fr/computing/cs/taco/taco.html

[4] Device Server, http://www.esrf.fr/computing/cs/taco/tutorial_device/device.html

[5] RTAI (Real Time Aplication Interface), http://www.aero.polimi.it/rtai/

6. **Overview of real-time disruption prediction in JET: applicability to ITER.** [Vega et al., 2014b]
   J. Vega, A. Murari, S. Dormido-Canto, D. Alves, G. Farias, J. M. López, R. Moreno, A. Pereira, J. M. Ramírez, G. Rattá and JET-EFDA Contributors.
   41st EPS Conference on Plasma Physics. Berlin (Germany), 23-27 June 2014.

7. **Real-time prediction of disruptions: results in JET and research lines for ITER.** [Vega et al., 2013c]
   J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, J. M. Ramírez, J. M. López, D. Alves, G. Rattá, A. Pereira and JET-EFDA Contributors.
   8th Workshop on Fusion Data Processing, Validation and Analysis. November 4-6, 2013 Ghent, Belgium.

8. **Advanced data analysis techniques for event identification and prediction in plasma experiments.** [Vega et al., 2013b]
   J. Vega, A. Murari, R. Moreno, S. González, A. Pereira, S. Dormido-Canto, J. M. Ramírez, J. M. López, D. Alves and JET-EFDA Contributors.
   International Conference on Research and Applications of Plasmas. Warsaw, Poland, September 2-6, 2013.

9. **Automated Analysis of Edge Pedestal Gradient Degradation During ELMs.** [González et al., 2012c]
   S. González, J. Vega, A. Murari, A. Pereira and JET-EFDA contributors.
   7th Workshop on Fusion, Data Processing, Validation and Analysis. March 26-28, 2012,ENEA, Frascati, Italy.

10. **Spatial location of local perturbations in plasma emissivity derived from projections using conformal predictors.** [Vega et al., 2013]
    Jesús Vega, Andrea Murari, Sergio González, Augusto Pereira, Ignacio Pastor.
    2nd International Conference Frontiers in Diagnostic Technologies. November 28-30, 2011. Frascati, Italy.

11. **H/L transition time estimation in JET using conformal predictors.** [González et al., 2012d]
    S. González, J. Vega, A. Murari, A. Pereira, S. Dormido-Canto, J.M. Ramírez and JET EFDA contributors.
    Proceedings of the 8th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Francisco, California, United States of America 20-24 June 2011.

12. **Automatic Determination of L/H Transition Times in DIII-D Through a Collaborative Distributed Environment.** [Farias et al., 2012]
    G. Farias, J. Vega, S. González, A. Pereira, X. Lee, D. Schissel, P. Gohil.
    Proceedings of the 8th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Francisco, California, United States of America 20-24 June 2011.

13. **Overview of statistically hedged prediction methods: from off-line to real-time data analysis.** [Vega et al., 2012]
    J. Vega, A.Murari, S. González, A. Pereira, I. Pastor, and JET-EFDA Contributors.
    Proceedings of the 8th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Francisco, California, United States of America 20-24 June 2011.

14. **Accurate and reliable image classification by using conformal predictors in the TJ-II Thomson Scattering**. [Vega et al., 2010]
J. Vega, A. Murari, A. Pereira, S. González, I. Pastor.
Proceedings of the 18th Topical Conference on High Temperature Plasma Diagnostics, May 16-20, 2010, Wildwood, New Jersey

15. **Support vector machine based feature extractor for L/H transitions in JET**. [González et al., 2010]
S. González, J. Vega, A. Murari, A. Pereira, J.M. Ramírez, S. Dormido-Canto and JET-EFDA contributors.
Proceedings of the 18th High Temperature Plasma Diagnostics (HTPD) conference, May 16-20, 2010, Wildwood, New Jersey, USA

16. **Automatic ELM location in JET using a universal multi-event locator.** [González et al., 2010b]
S. González, J. Vega, A. Murari, A. Pereira, M. Beurskens and JET-EFDA contributors.
6th Fusion Data Validation Workshop 2010. January 25-27, 2010, Madrid, Spain

17. **Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: new image recognition classifier and fault condition detection.** [Makili et al., 2010]
L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, A. Portas, D. Pérez-Risco, M.C. Rodríguez-Fernández, P. Busch.
Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 15 - 19 June 2009, Aix-en-Provence, France.

18. **Real-time remote diagnostic monitoring test-bed in JET.** [Castro et al., 2010b]
R. Castro, K. Kneupner, J. Vega, G. De Arcas, J.M. López, K. Purahoo, A. Murari, A. Fonseca, A. Pereira, A. Portas and JET-EFDA Contributors.
Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 15 - 19 June 2009, Aix-en-Provence, France.

19. **Securing MDSplus in a Multi-organization Environment.** [Castro et al., 2010]
R. Castro, J. Vega, T. Fredian, K. Purahoo, A. Pereira, A. Portas.
Proceedings of the 7th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 15 - 19 June 2009, Aix-en-Provence, France.

20. **Data distribution architecture based on standard Real Time Protocol.** [Castro et al., 2009]
R. Castro, J. Vega, A. Pereira, A. Portas.
Proceedings of the 25th Symposium on Fusion Technology - SOFT-25. Rostock, Germany 15–19 September.2008

21. **Overview of intelligent data retrieval methods for waveforms and images in massive fusion databases.** [Vega et al., 2009]
J. Vega, A. Murari, A. Pereira, A. Portas, R. Castro and JET-EFDA Contributors.
Proceedings of the 25th Symposium on Fusion Technology - SOFT-25. Rostock, Germany 15–19 September.2008

22. **EFDA-fed: European federation among fusion energy research laboratories.** [Castro et al., 2008b]
R. Castro, J. Vega, A. Portas, A. Pereira, C. Rodriguez, S. Balme, J. M. Theis, J. Signoret, P. Lebourg, K. Purahoo, K.Thomsen, H. Fernandes, A. Neto, A. Duarte, F. Oliveira, F. Reis, J. Kadlecsik.
Terena networking conference. Bruges (Belgium) 19-22 May 2008.

23. Intelligent technique to search for patterns within images in massive databases. [Vega et al., 2008b]
J. Vega**,** A. Murari, A. Pereira, A. Portas, P. Castro and JET-EFDA Contributors.
Proceedings of the HTPD High Temperature Plasma Diagnostic 2008, Albuquerque, New Mexico. 11-15 May 2008.

24. EFDA-Fed: Una federación internacional para investigación en fusión basada en PAPI. [Castro et al., 2008]
Rodrigo Castro, Jesús Vega, Ana Portas, Augusto Pereira, Diego R. López.
Jornadas Técnicas RedIRIS 2007, Mieres (Asturias), 19-23 de noviembre de 2007.

25. Recent results on structural pattern recognition for Fusion massive databases. [Vega et al., 2007b]
J. Vega, G. Rattá, A. Murari, P. Castro, S. Dormido-Canto, R. Dormido, G. Farias, A. Pereira, A. Portas, E. de la Luna, I. Pastor, J. Sánchez, N. Duro, R. Castro, M. Santos, H. Vargas.
Proceedings of the IEEE International Symposium on Intelligent Signal Processing, WISP 3-5, Alcalá de Henares (Spain), Oct. 2007.

26. Structural Pattern Recognition Techniques for Data Retrieval in Massive Fusion Databases. [Vega et al., 2008c]
J. Vega, A. Murari, G. A. Rattá, P. Castro, A. Pereira, A. Portas, and JET-EFDA Contributors.
Burning Plasma Diagnostics: An International Conference. 24–28 September 2007. Varenna (Italy).

27. First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET. [Rattá et al., 2008]
G. A. Rattá, J. Vega, A. Pereira, A. Portas, E. de la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari.
Proceedings of the 6th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. Inuyama, Japan 4–8 June 2007

28. Structural pattern recognition methods based on string comparison for fusion databases. [Dormido-Canto et al., 2008]
S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, N. Duro, H. Vargas, G. Rattá, A. Pereira, A. Portas.
Proceedings of the 6th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. Inuyama, Japan 4–8 June 2007.

29. An event-oriented database for continuous data flows in the TJ-II environment. [Sánchez et al., 2008]
E. Sánchez, A. de la Peña, A. Portas, A. Pereira, J. Vega, A. Neto, H. Fernandes.
Proceedings of the 6th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. Inuyama, Japan 4–8 June 2007.

30. Remote control of data acquisition devices by means of message oriented middleware. [Sánchez et al., 2007b]
E. Sánchez, A. Portas, A. Pereira, J. Vega, I. Kirpitchev.
Proceedings of the 24th Symposium on Fusion Technology - SOFT-24, Warsaw, Poland, 11-15 September 2006.

31. Real-time lossless data compression techniques for long-pulse operation. [Vega et al., 2008b]
J. Vega, M. Ruiz, E. Sánchez, A. Pereira, A. Portas, E. Barrera.

Proceedings of the 24th Symposium on Fusion Technology - SOFT-24, Warsaw, Poland, 11-15 September 2006

32. TJ-II Operation Tracking from Cadarache. [Vega et al., 2006]
J. Vega, E. Sánchez, A. Portas, A. Pereira, A. López, E. Ascasíbar, S. Balme, Y. Buravand, P. Lebourg, J. M. Theis, N. Utzel, M. Ruiz, E. Barrera, S. López, D. Machón, R. Castro, D. López, A. Mollinedo, J. A. Muñoz.
15th International Stellarator Workshop. October 3 - 7, 2005, Madrid (Spain).

33. Overview of the TJ-II remote participation system. [Vega et al., 2005b]
J. Vega, E. Sánchez , A. Portas, A. Pereira , A. Mollinedo, J.A. Muñoz , M. Ruiz, E. Barrera, S. López, D. Machón, R. Castro, D. López.
5th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 12 - 15 July 2005 , Budapest , Hungary.

34. Applying a message oriented middleware architecture to the TJ-II remote participation system. [Sánchez et al., 2006]
E. Sánchez , A. Portas, A. Pereira, J. Vega.
5th IAEA TM on Control, Data Acquisition, and Remote Participation for Fusion Research. 12 - 15 July 2005 , Budapest , Hungary.

35. Application of intelligent classification techniques to the TJ-II Thomson Scattering diagnostic. [Vega et al., 2005]
J. Vega, I. Pastor, J. L. Cereceda, A. Pereira, J. Herranz, D. Pérez, M. C. Rodríguez, G. Farias, S. Dormido-Canto, J. Sánchez, R. Dormido, N. Duro, S. Dormido, G. Pajares, M. Santos, J. M. de la Cruz.
32nd EPS Conference on Plasma Physics and Controlled Fusion combined with the 8th International Workshop on Fast Ignition of Fusion Targets. Tarragona (Spain), 27 June - 1 July 2005.

B. Algoritmos para obtener los coeficientes del hiperplano lineal libSVM

B.1. getHyperplane.c

```
/*****************************************************************************/
/* To compile:                                                               */
/*               gcc -g  -lm getHyperplane.c -o getHyperplane                */
/*                                                                           */
/* To execute:                                                               */
/*               getHyperplane <number of dimensions> < model                */
/*                                                                           */
/*****************************************************************************/

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <math.h>

main(int argc, char **argv)
{
    int dimension = atoi(argv[1]);
    char words[256];
    char linea[256];
    int total_sv;
    double rho;
    int label[2];
    int nr_sv[2];

    struct valores
    {
        double alpha;
        double value[dimension];
    };

    char caracteristica_valor[dimension][75];
    int k;
    int i;
    char *t;

    /* LEEMOS LA CABECERA */

    for(i=0;i<7;i++)
    {
        strcpy(words,"");
        scanf("%s",&words);
    }
    scanf("%d",&total_sv);
    strcpy(words,"");
    scanf("%s",&words);
    scanf("%lf",&rho);
    strcpy(words,"");
    scanf("%s",&words);
    scanf("%d %d", label, label+1);
    strcpy(words,"");
    scanf("%s",&words);
    scanf("%d %d", nr_sv, nr_sv+1);
    strcpy(words,"");
    scanf("%s",&words);

    struct valores *valor=(struct valores *) malloc(total_sv*sizeof(struct valores));

    /* LEEMOS LOS VECTORES SOPORTE */

    for(k=0;k<total_sv;k++)
    {
        strcpy(linea,"");
        scanf("%s", &linea);
        valor[k].alpha = atof(linea);
        for(i=0;i<dimension;i++)
        {
            strcpy(caracteristica_valor[i],"");
```

```
            scanf("%s",caracteristica_valor[i]);
            for(t = strtok(caracteristica_valor[i],":"); t!= NULL; t=strtok(NULL, ":"))
            {
                valor[k].value[i]=atof(t);
            }
        }
    }

    /* CALCULAMOS LOS COEFICIENTES PARA CADA CARACTERISTICA */

    int j;
    double sumaAlphaPorValueA[dimension];
    double sumaAlphaPorValueB[dimension];
    double coeficientes[dimension];
    printf("\n");
    double c2 = 0.0;
    for(i=0;i<dimension;i++)
    {
        sumaAlphaPorValueA[i]=0;
        for (j=0;j<nr_sv[0];j++)
        {
            sumaAlphaPorValueA[i] += valor[j].alpha*valor[j].value[i];
        }
        sumaAlphaPorValueB[i]=0;
        for (j=0;j<nr_sv[1];j++)
        {
            sumaAlphaPorValueB[i] += valor[nr_sv[0]+j].alpha*valor[nr_sv[0]+j].value[i];
        }
        coeficientes[i]=fabs(sumaAlphaPorValueB[i])-sumaAlphaPorValueA[i];
        c2 += coeficientes[i]*coeficientes[i];
        printf("x%d: %f\n", i+1, coeficientes[i]);
    }
    printf("rho: %f\n", rho);
    free(valor);
    printf("\n\nEnd of program.\n");
}
```

B.2. getHyperplane_deNormalized.c

```
/*****************************************************************************/
/* To compile:                                                               */
/*       gcc -g  getHyperplane_deNormalized.c -o getHyperplane_deNormalized  */
/*                                                                           */
/* To execute:                                                               */
/*       getHyperplane_deNormalized <ncoefici> < NormalizedCoefs             */
/*                                                                           */
/*****************************************************************************/

#include <stdio.h>
#include <stdlib.h>
#include <string.h>

main(int argc, char **argv)
{
  double coefs[200];
  double max[200];
  double min[200];
  double sumaK = 0.0;
  double k[200];
  int i,j;

    /* Input file */
  i=0;
  for(i=0;i<atoi(argv[1]);i++)
    {
          scanf("%lf%lf%lf", &coefs[i], &max[i], &min[i]);
          printf("%f %f %f\n", coefs[i], max[i], min[i]);
    }
  for(j=0;j<i-1;j++)
    {
        k[j] = coefs[j]/(max[j]-min[j]);
        sumaK -= k[j]*min[j];
    }
  k[i-1] = coefs[i-1];
  sumaK += k[i-1];
  k[i-1] = sumaK;
  printf("\n");
  for(j=0;j<i-1;j++)
    {
        printf("K%d = %e\n",j+1,k[j]);
    }
  printf("\nK0 = %e\n",k[i-1]);
  printf("\n\nEnd of program.\n");
}
```