Metodología KDSM para el estudio de dominios poco estructurados donde se presentan medidas seriadas muy cortas repetidas con factor de bloque.

Jorge Rodas jr@lsi.upc.es
Gabriela Alvarado geac@ee.ub.es
Karina Gibert karina@eio.upc.es
José Emilio Rojo jrojo@csub.scs.es
Ulises Cortés ia@lsi.upc.es

febrero de 2003

Resumen

Introducción del documento.- El presente documento representa 2 años de trabajo con técnicas de Inteligencia Artificial y Estadística para el análisis de medidas seriadas muy cortas repetidas.

Grandes cantidades de datos se almacenan diariamente debido a la dinámica transformación de la computación, en especial el Internet. Por tal motivo, cualquier organismo, dependencia, etc., tiene el interés de utilizar lo mejor posible sus datos, especialmente si esto les proporciona beneficios y más aun si éstos son económicos.

Objetivo del presente documento.- Presentar la metodología (KDSM)[1] como una aportación original, que permite el descubrimiento de conocimiento en datos que provienen de medidas seriadas muy cortas repetidas con factor de bloque y dichos datos pertenecen a *dominios poco estructurados* [Gib94].

Método para el desarrollo del documento.- De forma simple se expone la metodología KDSM mediante la presentación de los aspectos más relevantes de su aplicación a 2 ámbitos: el psiquiátrico y el laboral.

Conclusiones del documento.- Se presentan en 2 niveles: metodológico y aplicación. Referente al metodológico, se expone cómo la metodología KDSM enriquece a las técnicas de IA y Estadística, potenciándoles su capacidad de resolución de problemas al trabajar conjuntamente. En cuanto a la aplicación de KDSM a los ámbitos antes mencionados, se lograron con gran éxito los objetivos de cada estudio, lo cual motivó al autor(a) del presente trabajo para que diera a conocer, primeramente en su comunidad, su aportación por medio de ésta importante convocatoria.

[1] Del inglés Knowledge Discovery in Serial Measurement

Índice general

Glosario

X	Matriz que contiene las características de los individuos del dominio de estudio.
Y_T	Matriz que contiene el total de medidas seriadas—muy cortas y repetidas en el tiempo—de cierto atributo de interés.
Y_{T*}	Caso excepcional de la Matriz Y_T.
Y_0	Matriz de medidas seriadas basales muy cortas y repetidas en el tiempo.
Y_{0*}	Caso excepcional de la Matriz Y_0.
Y	Matriz de medidas seriadas—muy cortas y repetidas en el tiempo—tras cada evento.
Y_*	Caso excepcional de la Matriz Y.
D	Matriz que contiene la diferencia entre la matriz Y_{ij}^t y la anterior $Y_{i,j-1}^t$ para cada individuo.
Z	Matriz que contiene las características de los eventos.
$\mathcal{P}$	Partición de referencia.
$\mathcal{P}_\mathcal{R}$	Partición a partir de un sistema de reglas.
$Y_0^\mathcal{P}$	Indica la operación de cluster en Y_0
$D^{\mathcal{P}_\mathcal{R}}$	Indica la operación de Clasificación Basada en Reglas de la Matriz D.
$S^\mathcal{P}$	Sistema de Interpretación sobre la Partición de referencia $\mathcal{P}$.
$S^{\mathcal{P}_\mathcal{R}}$	Sistema de Interpretación sobre la Partición a partir de un sistema de reglas.
G	Módulo de interpretación gráfica de la herramienta COLUMBUS.
BC_0	Base de Conocimiento obtenida de los basales.
$BC_{\mathcal{P}_\mathcal{R}}$	Base de Conocimiento obtenida del análisis de los eventos.
$\mathcal{R}^0$	Sistema de reglas que conforma la Base de Conocimiento (BC_0).
$\mathcal{R}'$	Sistema de reglas que conforma la Base de Conocimiento ($BC_{\mathcal{P}_\mathcal{R}}$).
M	Método o Metodología utilizado(a) para obtener $\mathcal{R}^0$

$M_{\mathcal{R}}$	Método o Metodología utilizado(a) para obtener $\mathcal{R}'$
E	Conjunto de eventos que actúan sobre los objetos de estudio. Indica la ocurrencia de un evento, iniciando con la medición basal.
$c_{\#}^{\mathcal{P}}$	Indica la clase, donde el sub-índice corresponde al número o etiqueta de clase y el super-índice al número o etiqueta de la partición.
t	Indica el instante de tiempo en que se medirá Y tras la ocurrencia de E.
r	Indexa el instante de tiempo t.
VEC	Gráfico que muestra la variabilidad entre clases.
VIC	Gráfico que muestra la variabilidad interna de una clase.

Definiciones

$E =$	$\{E_{1,0} \dots E_{i,j} \dots E_{n,m}\}$; $i = 1 \dots n$ y $j = 0 \dots m$
$t \in$	$\{t_1 \dots t_r\}$
$r \in$	$\{1 \dots R\}$; donde R es pequeño
$X \stackrel{def}{=}$	$[x_{ik}]$; donde $i = 1 \dots n$ y $k = 1 \dots K$
$Y_T \stackrel{def}{=}$	Y_{ij}^t; $i = 1 \dots n$ y $j = 0 \dots m$; $Y_{i0}^t \vert Y_{ij}^t$; $j = 1 \dots m$
$Y_0 \stackrel{def}{=}$	Y_{i0}^t; $i = 1 \dots n$
$Y \stackrel{def}{=}$	Y_{ij}^t; $i = 1 \dots n$ y $j = 1 \dots m$
$Y_{T*} \stackrel{def}{=}$	Y_{ij}^t; $i = 1 \dots n$ y $j = 1 \dots m$; $Y_{i0}^t \vert Y_{ij}^t$; $j = 2 \dots m$
$Y_{0*} \stackrel{def}{=}$	Y_{i1}^t; $i = 1 \dots n$
$Y_* \stackrel{def}{=}$	Y_{ij}^t; $i = 1 \dots n$ y $j = 2 \dots m$
$D \stackrel{def}{=}$	$[d_{i,j-1}^t]$; donde $i = 1 \dots n$ y $j = 1 \dots m$
$Z \stackrel{def}{=}$	$[z_{vw}]$; donde $v = 1 \dots V$ y $w = 1 \dots W$
$\mathcal{P}:$	$Y_0 \mapsto Y_0^{\mathcal{P}}$
$\mathcal{P}_{\mathcal{R}}:$	$D \stackrel{\mathcal{R}}{\mapsto} D^{\mathcal{P}_{\mathcal{R}}}$
$Y_0^{\mathcal{P}} =$	$\{c_1^{\mathcal{P}} \dots c_\xi^{\mathcal{P}}\}$
$D^{\mathcal{P}_{\mathcal{R}}} =$	$\{c_1^{\mathcal{P}_{\mathcal{R}}} \dots c_\xi^{\mathcal{P}_{\mathcal{R}}}\}$
$S^{\mathcal{P}}:$	$X \mapsto X \vert Y_0^{\mathcal{P}}$
$S^{\mathcal{P}_{\mathcal{R}}}:$	$Z \mapsto Z \vert D^{\mathcal{P}_{\mathcal{R}}}$
$\mathcal{R} =$	$\{r_1 \dots r_{\mathcal{R}}\}$
$M:$	$X \vert Y_0^{\mathcal{P}} \mapsto \mathcal{R}$
$M_{\mathcal{R}}:$	$Z \vert D^{\mathcal{P}_{\mathcal{R}}} \mapsto \mathcal{R}'$

Capítulo 1

Introducción

En los últimos años la informática se ha convertido en parte integral del desarrollo de cualquier persona o institución. El incremento de información útil y su manipulación por medio de métodos clásicos estadísticos continúa siendo un problema difícil. Hoy, los métodos que emplean técnicas de análisis por computadora se han vuelto indispensables, siendo muy interesantes aquellos agrupados en la disciplina conocida como *Extracción de Conocimiento en Bases de Datos, KDD*[1] [AZ98].

En los métodos del tipo KDD intervienen varias disciplinas, como son: el reconocimiento de patrones, el aprendizaje automático, los sistemas expertos (todas ellas de IA), la tecnología de bases de datos, el datawarehouse, la estadística y las técnicas para visualización de información.

Según [BA96] un proceso de KDD presenta los siguientes pasos:

- *comprensión del dominio*,
- *conformación de las bases de datos a utilizar y su depuración*,
- *extracción de información valiosa oculta en los datos* (esta fase se conoce como minería de datos),
- *formulación, en forma de patrones o reglas, de la información valiosa extraída como conocimiento y*
- *el análisis de los resultados obtenidos*

[1] Del inglés Knowledge Discovery in Databases

En cuanto a disciplinas distintas de la informática, el proceso de KDD establece un puente entre la obtención de información y la toma de decisiones, propias a las mismas, por medio del cual se pretende mejorar y hacer efectivo el trabajo de cada especialista de las diferentes disciplinas.

Sin embargo, una buena parte de la información proviene de dominios que no presentan una estructura ordenada que permita distinguir de forma simple qué información es de interés, de la que no lo es. Dichos dominios, se sitúan en los denominados por [Gib94] como *dominios poco estructurados*.

Algunas de las características relevantes que presentan los *dominios poco estructurados* son:

Matrices con datos heterogéneos. Los atributos que describen los objetos pueden ser cuantitativos o cualitativos. Los cualitativos tienen muchas modalidades, tanto mayor cuanto mayor es la experiencia del usuario.

Existencia de información adicional sobre la estructura del dominio. Es común que se cuente con conocimiento declarativo sobre la estructura del dominio de estudio (relaciones entre atributos, objetivos de agrupación, etc).

Conocimiento parcial y no homogéneo. Los expertos suelen disponer de grandes cantidades de conocimiento implícito, además de manejar diversos grados de especificidad, lo que hace a este conocimiento no homogéneo.

La utilización de la informática para monitorizar un proceso, por ejemplo, el seguimiento a un paciente que esta sometido a una terapia; ofrece una gran cantidad de información tanto del proceso como de los *actores* (individuos y los eventos que intervienen sobre los individuos) en el proceso. Con frecuencia encontramos que muchos datos, que provienen de monitorizar un proceso, son resultado de medidas seriadas en el tiempo de duración de dicho proceso. Además éste proceso suele repetirse tantas veces sea necesario para lograr el objetivo esperado.

Los casos de estudio que serán tratados en los capítulos §3 y §4, cumplen con características propias de los *dominios poco estructurados*, además contienen datos representados por medidas seriadas muy cortas y repetidas de un atributo de interés. El objetivo general en dichos estudios es distinguir los distintos perfiles obtenibles de dichas medidas, por lo que se tiene la tarea de encontrar cómo manipularlas y extrarles información.

1.1 Motivación

El origen de este trabajo se sitúa en la colaboración entre el Dr. J. Emilio Rojo del Servicio de Psiquiatría de la Ciudad Sanitaria y Universitaria de Bellvitge en Barcelona, la Dra. Karina Gibert del Departamento de Estadística e Investigación Operativa y miembro del grupo de investigación KELM[2] del Departamento de Lenguajes y Sistemas Informáticos de la UPC.

El punto de partida del estudio que realiza el Dr. Rojo, trata sobre ciertos efectos que aparecen en pacientes a los que se les tuvo bajo tratamiento con la Terapia Electro-Convulsiva (TEC). Los datos del estudio consisten en medidas seriadas cortas y repetidas e información sobre cada paciente bajo este tipo de terapia. Los detalles sobre la terapia y sus efectos serán abordados en la sección §1.2 y en el capítulo §3.

Las medidas seriadas y repetidas consisten en observaciones llevadas a cabo sobre un mismo atributo característico en diversos tiempos [Lin99]. Lo que las distingue de otras observaciones dentro de los modelos estadísticos tradicionales de datos son:

- El mismo atributo se mide sobre la misma unidad de observación más de una vez: ello implica que los atributos no son independientes como sucede en un análisis de regresión común y

- se involucra más de una unidad de observación (individuo): los atributos no conforman una serie de tiempo simple.

Las bases de datos que contienen medidas seriadas suelen tener una o varias de ellas—una por cada atributo de interés—para cada unidad de observación, objeto o individuo incluido en el estudio. Sin embargo, no es correcto analizarlas todas juntas.

Una forma satisfactoria de analizar las medidas seriadas y repetidas de un atributo continuo en un conjunto de individuos es la propuesta de Matthews [Mat93]:

a) reducir las medidas seriadas a un conjunto pequeño de medidas independientes obteniendo alguna función apropiada para las observaciones en un objeto (promedio, área bajo la curva, etc) y

[2]Knowledge Engineering and Machine Learning Group

b) analizar esas síntesis, que ya son independientes entre sí, utilizando métodos estadísticos clásicos.

Aparentemente, estas dos características de las medidas de ser *seriadas* y *repetidas en el tiempo*, no ofrecen mucho problema para ser analizadas con técnicas clásicas de series de tiempo. Sin embargo, ¿qué sucede cuando la cantidad de medidas es sumamente pequeña?

Que en definitiva no permitiría un análisis clásico de series de tiempo.

Por otra parte, en situaciones así se cuenta con una gran cantidad de información adicional sobre los actores del proceso y el proceso mismo, que no son medidas seriadas, pero que sí guardan una estrecha relación con lo que sucede en el proceso. Además frecuentemente los actores conforman un factor de bloque sobre las medidas seriadas; entonces, ¿cómo aprovechar esta información adicional? ¿cómo se trabajaría dicha información, en relación a las medidas seriadas y repetidas, si ésta no se conforma de medidas sino de características de los actores en el proceso?

De hecho, los datos reales que actualmente estamos trabajando, provenientes de ámbitos como la Psiquiatría (capítulo §3) o la Capacitación Laboral (capítulo §4), presentan las características antes mencionadas. Concretamente, en esta clase especial de *dominios poco estructurados*, se presenta un conjunto de individuos y para cada uno se produce una cantidad variable de ocurrencias de un evento, en distintos instantes del tiempo. Justo después de la ocurrencia, interesa estudiar la evolución de un atributo de interés—relativo al individuo en cuestión—en un período de tiempo posterior al evento. Dicho atributo se mide un número fijo y pequeño de ocasiones, siendo las mismas en todos los casos, por lo que, conforman *medidas seriadas muy cortas y repetidas* (ver figura 1.1).

Sin embargo, se debe tener presente que este tipo de situaciones no son exclusivas de un determinado ámbito; de hecho, las medidas seriadas cortas y repetidas ocurren con frecuencia en ámbitos tales como:

Economía En una economía petrolera, el precio del crudo influye en las tasas de empleo. Por tanto, se hacen mediciones periódicas (ej. mensuales) de la tasa del desempleo, en las provincias que forman parte de dicha economía, tras la caída de su precio durante cierto tiempo, usualmente el año fiscal.

Vulcanología Con el fin de conocer más sobre el comportamiento de los volcanes y poner a punto los programas de prevención de desastres, éstos son puestos bajo un sistema especial de monitorización, durante períodos críticos, en los cuales se registra un evento que puede consistir de movimientos sísmicos o emisiones de gases y otras sustancias, o bien, ambas.

El período crítico, no ocurre a intervalos de tiempo homogéneo, ya que ni el movimiento sísmico ni las emisiones se presentan de forma continua durante todo ese período.

En muchos estudios el último evento podría significar el fin del período crítico y no la erupción de un volcán. Por ello, es de interés encontrar su patrón o patrones de comportamiento.

En este caso particular, la monitorización no es otra cosa que medidas repetidas de algunas variables (ej. nivel de contaminación, escala de richter, etc.), en diversos puntos del volcán y zonas geográficas cercanas al mismo, tras cada evento producido. El sistema especial de monitorización de volcanes está activo hasta que el período crítico cesa. Este tipo de estudios es de gran utilidad pues un volcán debe ser observado continuamente para obtener información relevante sobre el mismo, que permita salvar vidas, reducir daños materiales y los costos implicados.

Contaminación ambiental Otro estudio interesante, se presenta dentro de éste ámbito. Dicho estudio, analiza el proceso de una planta de tratamiento del agua, donde tras una tormenta eléctrica (evento) el agua a tratar contiene mayores cantidades de contaminantes.

En este caso, la monitorización consiste en medidas repetidas de varios atributos que dan idea del estado del agua (ej. nivel de sólidos en suspensión, nivel de cloro, etc.), en distintos puntos de la planta (cada fase del proceso de tratamiento), después de cada tormenta. La utilidad de este estudio se refleja directamente en una continua mejora del proceso de tratamiento del agua y en consecuencia de la calidad de la misma.

Medicina La medicina es uno de los campos científicos donde las medidas seriadas cortas y repetidas ocurren con gran frecuencia.

> Los tratamientos o terapias experimentales que se realizan en seres humanos tienen la finalidad de medir su eficacia ante un padecimiento o bien medir la evolución de una enfermedad. La mayoría de las veces, esta medida consiste en una serie de medidas para un atributo (ej. síntomas) o varios, a lo largo del período de tratamiento.
>
> Así, cuando una droga experimental (para cierta enfermedad) se utiliza, se realiza la medición—en un período fijo de tiempo—de los síntomas del enfermo, cada determinado número de horas tras haber sido suministrada y dicha monitorización se repite en cada toma.

Con respecto a la identificación de patrones de series de tiempo, existe una gran cantidad de métodos convencionales que se basan en una clasificación que usa análisis de frecuencia, como lo son: el análisis de Fourier [RJ93], el modelo de auto-regression [Wu94], o los modelos de Markov [RJ93, PAVL96, KP97]. Sin embargo, estos métodos se basan en modelos matemáticos que dependen de varios parámetros y trabajan de forma global con los patrones de series de tiempo conformados a partir de muchas observaciones y no tratan casos en los cuales existen pocas observaciones por serie. El problema que se origina es que es difícil realizar buenas estimaciones de los parámetros con un número pequeño de observaciones por serie, si se pretende utilizar el modelo matemático de series de tiempo. Además, si los modelos obtenidos no son muy simples resultan difíciles de interpretar para los seres humanos y las máquinas, pues no están acompañados de una explicación intuitiva sobre los resultados que generan.

Por otra parte, algunas de las aproximaciones desde la inteligencia artificial requieren de *conocimiento a priori* de las estructuras de las medidas seriadas o patrones temporales por descubrir [KP97]. Pero, dicho conocimiento con frecuencia no está disponible para casos de la vida real.

Por tales motivos, el manejo de las medidas seriadas muy cortas y repetidas en el tiempo con un factor de bloque (o empaquetadas) con alguno de los métodos citados anteriormente, no permiten realizar análisis adecuados. Además, en los métodos de IA, el no tener *conocimiento a priori* impedirá, indudablemente, el descubrimiento de conocimiento a partir de ese tipo de "mina"de datos.

Es así que, establecí una metodología que haga factible el descubrimiento de conocimiento

en dominios donde se presentan medidas seriadas muy cortas y repetidas con factor de bloque (conformado por los individuos) y donde se cuenta con información adicional de los actores en el proceso como lo son sus atributos característicos. Para ello, los sistemas híbridos, que utilizan técnicas de la estadística y la inteligencia artificial, parecen ser una aproximación prometedora para obtener buenos resultados en el descubrimiento de patrones en este tipo de medidas seriadas.

Esto motivó los trabajos para construir la metodología *KDSM*[3] (capítulo §2) la cual es una metodología híbrida para el descubrimiento de patrones en las situaciones donde se presenten este tipo de medidas; y el diseño e implementación del programa computacional *COLUMBUS*, que permite aplicar la metodología KDSM. Aunque dichos trabajos parten de un problema del ámbito de la Psiquiatría, de igual forma podría haber sido el ámbito de la Capacitación Laboral o cualquier otro de los mencionados en este capítulo o seguramente muchos de los que quedan por mencionar.

En las siguientes subsecciones detallamos 2 situaciones reales, que se ajustan a las características aquí expuestas, de los dominios que constituyen nuestros casos de estudio detallados en los capítulos §3 y §4.

1.1.1 Terapia Electro-Convulsiva

La Terapia Electro-Convulsiva (terapia basada en aplicación de electroshocks, TEC) durante años fue marginada de la práctica común de la psiquiatría debido al mal uso de la misma, especialmente, durante la segunda guerra mundial. No fue sino hasta los años 90 en que vuelve el interés por estudiarla, pues a pesar de su mala fama, es considerada una terapia segura y que proporciona muy buenos resultados.

El equipo del Dr. Emilio Rojo del Hospital de Bellvitge, actualmente dirige un estudio muy interesante sobre la terapia TEC. Dicho estudio versa sobre los efectos psicofisiológicos que se presentan en los pacientes que se someten a dicha terapia.

Como se ha mencionado con anterioridad, registrar la eficacia de la aplicación de una terapia para su estudio, generalmente se hace por medio de medidas seriadas de un atributo de interés

[3]Del inglés Knowledge Discovery in Serial Measurement

que, en una primer etapa del estudio del Dr. Rojo, corresponde al *Tiempo de Reacción* (TR)[4], a lo largo del período de tratamiento.

Así, antes de iniciar la TEC, se realiza la medición basal del TR del paciente, cada determinado número de horas que establece un *protocolo especial* (ver §3). Una vez iniciada la TEC tras cada aplicación de un electroshock se repiten las medidas seriadas—en un período fijo de tiempo—hasta el término de la TEC.

1.1.2 Capacitación para el trabajo

El Servicio Estatal de Empleo surge a partir de la coordinación de la Secretaría del Trabajo y Previsión Social (Ejecutivo Federal) y los Gobiernos de los Estados (Ejecutivo del Estado), para la realización de programas de sentido social que tienen como objetivo promover el diseño y aplicación de políticas orientadas a la generación de empleo.

Las acciones del Departamento del Servicio Estatal de Empleo se dirigen a vincular a los demandantes de trabajo, con las necesidades de mano de obra del aparato productivo, promoviendo la inserción productiva de los trabajadores y la oportuna interacción entre ellos. Para lograr estos objetivos, entre otras actividades relevantes, se implementó el Programa de Becas de Capacitación para Trabajadores Desempleados (PROBECAT).

Tanto para el gobierno federal como para los gobiernos estatales es de gran importancia conocer en que medida la inversión económica en este rubro cumple el objetivo de colocación de personas desempleadas y qué otros factores intervienen para lograr este fin, de forma que los gobiernos en actuación conjunta actúen en beneficio de todos.

La realización de la capacitación consiste en la impartición de un determinado tipo de curso, que depende de las necesidades específicas de cada municipio en cada estado. El curso se impartirá tantas veces sea necesario. Para registrar la eficacia de dicho programa se toman medidas seriadas del atributo ID que indica cuantos participantes en el curso han sido contratados a lo largo de 3 meses.

Así, antes de iniciar la capacitación, se realiza la medición basal del municipio (correspon-

[4]TR: tiempo en que un paciente tarda en reaccionar ante un estímulo que puede ser visual, auditivo o una combinación de éstos (para más detalles ver §3)

diente a la tasa de desempleo), cada determinado número de días establecido por la STyPS. Una vez iniciada la capacitación tras cada impartición de un curso se repiten las medidas seriadas—en un período fijo de tiempo—hasta el término de la misma.

Como programa piloto, la metodología KDSM se aplicó para obtener conocimiento de la evolución del PROBECAT, así como de su efectividad a través del análisis de la información relacionada con: la colocación de egresados, características de los municipios del Estado de Chihuahua y necesidades del sector productivo. La información obtenida por KDSM permite a la Secretaría de Trabajo y Previsión Social conocer el efecto que cada curso impartido aporta al municipio donde actúa el PROBECAT en general para que a partir de las tendencias globales de los mercados laborales, dicha secretaría de estado, pueda incidir en el funcionamiento del programa de capacitación de manera más oportuna y eficaz.

A continuación se describe el problema que da origen a la metodología KDSM, donde se introducen una serie de conceptos importantes y para que sean más comprensibles se hace referencia a ejemplos relativos al ámbito psiquiátrico (capítulo §3). Esta ejemplificación se destaca con diferente sangría y en letra tipo typewriter, con el fin de destacarlos del resto del documento.

1.2 Formulación del Problema

En la figura 1.1, podemos observar la representación de una serie de *individuos* ($i_1..i_n$) sobre los que se producen m ocurrencias de un cierto *evento* E en distintos instantes del tiempo ($E_{i,1} \ldots E_{i,m}$). Ligado a la ocurrencia de dicho evento, existe un atributo (o un conjunto de atributos) de interés Y, que afecta al comportamiento del individuo en cuestión e interesa estudiar su evolución en un cierto período de tiempo $[t_1, t_r]$ posterior a cada ocurrencia de E.

Así, para cada individuo y cada ocurrencia de E se toma un cierto número fijo de medidas (r) de Y a intervalos de tiempo fijos también, a contar a partir de la ocurrencia de cada E.

Por ejemplo, con referencia a la aplicación real del capítulo §3 se tiene que:

$I = \{i_1, .., i_n\}$ es un conjunto de pacientes $\{p_1, .., p_n\}$,

E es la aplicación de un electroshock ES a un cierto paciente en un cierto instante de tiempo. Así para cada paciente se tiene una secuencia de $ES\,TEC_i = \{ES_{i,1} \ldots ES_{i,m}\}$.

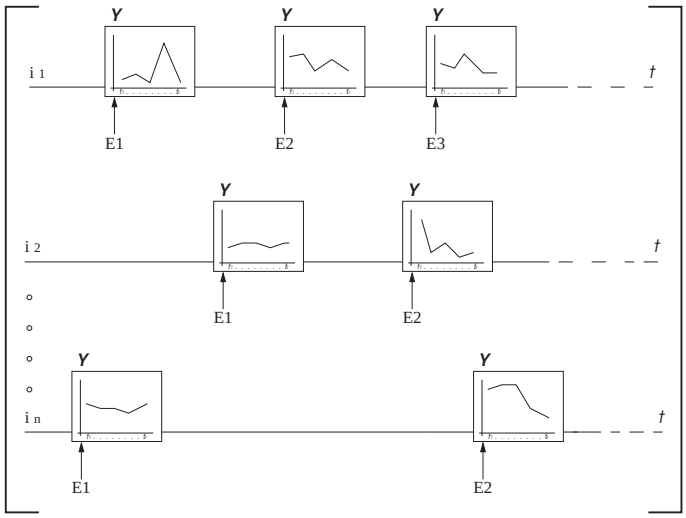


Figura 1.1: Medición de Y tras cada ocurrencia de E para cada i.

Y son los atributos de interés que corresponden, por ejemplo, al tiempo TR que tarda el paciente en reaccionar a un determinado *estímulo lumínico* tras cierto tiempo después de cada ES.

Las medidas seriadas de este atributo se llevan a cabo en las 24 horas que transcurren después de aplicar cada ES, específicamente a las 2h, 4h, 6h, 12h y 24h; así $Y = \{Y^2, Y^4, Y^6, Y^{12}, Y^{24}\}$.

La Fig. 1.2 es la representación gráfica de los ES aplicados a un conjunto de pacientes y las curvas que sigue el tiempo de reacción en las 24 horas que siguen a la aplicación de cada electroshock.

Un escenario como éste genera información estructurada del siguiente modo:

1. Para cada i se dispone de un conjunto de características cuantitativas y/o cualitativas $X_1 \dots X_K$. Esto da lugar a una matriz como la que se muestra en la tabla §1.1.

   En la matriz X, se tiene que x_{ik} $i = \{1 \dots n\}$ y $k = \{1 \dots K\}$, es el valor que toma X_K para un individuo n.

   Por ejemplo, si X_1 es la edad x_{11} sería la edad del paciente 1.

2. Para cada ocurrencia de E, se obtiene una secuencia de mediciones de Y en todos los instantes de tiempo fijos. Sea E_{ij} $i = \{1 \dots n\}$ y $j = \{1 \dots m\}$, la j-ésima ocurrencia del

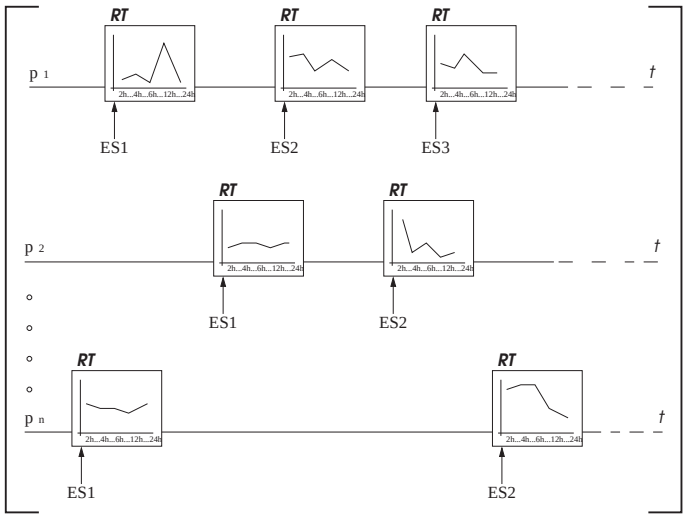


Figura 1.2: Electroshocks aplicados a ciertos pacientes.

	X_1	X_2	...	X_K
i_1	x_{11}	x_{12}	...	x_{1k}
i_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i_n	x_{n1}	x_{n2}	...	x_{nk}

X=

Tabla 1.1: *Matriz X.*

evento E sobre el individuo i. Así, para un cierto individuo i tenemos un número m de ocurrencias de E. Si en cada ocurrencia de E consideramos que el tiempo se inicializa a 0, será posible fijar $t_1 \dots t_r$ como los instantes de tiempo en que se medirá Y después de la ocurrencia de E.

Al final, las mediciones de Y dan lugar a una segunda matriz de datos como la que se muestra en la tabla §1.2.

3. Por otra parte, para cada E se dispone de un conjunto de características cuantitativas y/o cualitativas $Z_1 \dots Z_L$; dando lugar a otra matriz que se muestra en la tabla §1.3.

   En la matriz Z, se tiene que z_{ijl} $i = \{1 \dots n\}$, $j = \{1 \dots m\}$ y $l = \{1 \dots L\}$, es el valor que toma Z_L para un evento m relativo a un individuo n.

Las medidas del atributo de interés vienen dadas por Y_{ij}^r donde, $i = \{1 \dots n\}$ es el individuo,

Y=		t_1	t_2	...	t_r
	E_{11}	Y_{11}^1	Y_{11}^2	...	Y_{11}^r
	E_{12}	Y_{12}^1	Y_{12}^2	...	Y_{12}^r
	⋮	⋮	⋮	⋮	⋮
	E_{nm}	Y_{nm}^1	Y_{nm}^2	...	Y_{nm}^r

Tabla 1.2: *Matriz Y.*

Z=		Z_1	Z_2	...	Z_L
	E_{11}	z_{111}	z_{112}	...	z_{11l}
	E_{12}	z_{121}	z_{122}	...	z_{12l}
	⋮	⋮	⋮	⋮	⋮
	E_{nm}	z_{nm1}	z_{nm2}	...	z_{nml}

Tabla 1.3: *Matriz Z.*

$j = \{1 \dots m\}$ indica la j-ésima ocurrencia de E sucedida al individuo i y $r \in \{1 \dots R\}$ (donde R es pequeño) indexa el instante de tiempo, desde la ocurrencia de E_{ij}, en que se midió Y. Cabe precisar que *los tiempos de medición son los mismos* en relación a la ocurrencia de todos los eventos, para todos los individuos.

Fijado un par (i, j) las medidas de Y en el período de tiempo $t_1 \dots t_r$ se pueden representar por medio de *curvas muy cortas (donde r es pequeño)* y en apariencia independientes entre sí.

De hecho, cada individuo es independiente de los demás, por ello, *la cantidad de eventos y el instante del tiempo* en que se presentan pueden diferir de un individuo a otro sin ningún patrón subyacente.

Sin embargo, todos los eventos presentes en un mismo individuo tienen la influencia de sus características particulares. Esto hace que todas las medidas seriadas relativas al mismo individuo $\{Y_{ij}^1 \dots Y_{ij}^r\}$, $j = \{1 \dots m\}$ también reciban su influencia.

Por tanto, sobre la matriz Y, el individuo i se puede considerar como un *factor de bloque*,[5] que define *paquetes* de curvas que no son independientes entre sí (ver Tabla 1.4).

[5]Un factor (un atributo cualitativo) con efecto sobre el atributo respuesta, siendo o no de interés directo, pero que se debe tener en cuenta durante el experimento para obtener comparaciones homogéneas entre las observaciones donde el factor se mantiene constante [Pn89].

	t_1	t_2	...	t_R	
E_{11}	Y_{11}^1	Y_{11}^2	...	Y_{11}^r	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	bloque 1
E_{1m}	Y_{1m}^1	Y_{1m}^2	...	Y_{1m}^r	
E_{21}	Y_{21}^1	Y_{21}^2	...	Y_{21}^r	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	bloque 2
E_{2m}	Y_{2m}^1	Y_{2m}^2	...	Y_{2m}^r	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
E_{n1}	Y_{n1}^1	Y_{n1}^2	...	Y_{n1}^r	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	bloque n
E_{nm}	Y_{nm}^1	Y_{nm}^2	...	Y_{nm}^r	

Tabla 1.4: *Bloques de medidas seriadas.*

Así, tenemos que un bloque está conformado por todas las medidas seriadas $\{Y_{ij}^1 \dots Y_{ij}^r\}$, $j = \{1 \dots m\}$ que siguen a cualquier ocurrencia de E en el mismo individuo i, compuestas por un pequeño conjunto de medidas en un período de tiempo específico donde se presentan pocas observaciones, pero la misma cantidad tras cada evento y con la misma distribución en el tiempo en relación a la ocurrencia del evento. En concreto, se tendrá que analizar *un conjunto de medidas seriadas muy cortas en el tiempo con un factor de bloque.*

El propósito de este trabajo, fue por una parte, formalizar una metodología para encontrar qué patrón siguen las medidas seriadas $\{Y_{ij}^1 \dots Y_{ij}^r\}$ y, por otra, encontrar qué características del individuo $(X_1 \dots X_K)$ y del evento $(Z_1 \dots Z_L)$, están relacionadas con la evolución temporal de las medidas del atributo de interés Y. Sin embargo, para relacionar las Xs con las Ys y Zs hay que tener en cuenta lo siguiente:

- las características relativas a un individuo están representadas por una sola fila de la matriz X y
- que para cada paciente existen varias secuencias de mediciones de Y (m exactamente) que se sitúan libremente en la línea temporal y se representan en m filas de las matrices Y y Z que no son independientes entre sí.

Por tanto, será necesario encontrar cómo manipular la matriz X con las matrices Y y Z para realizar el análisis conjunto.

Si existiera un patrón fijo de ocurrencias de E en todos los individuos, respondiendo a la situación de la figura 3.3, se podría considerar una sola serie por individuo y se podría analizar con técnicas de series temporales usando una *política de intervenciones* [BJR94].

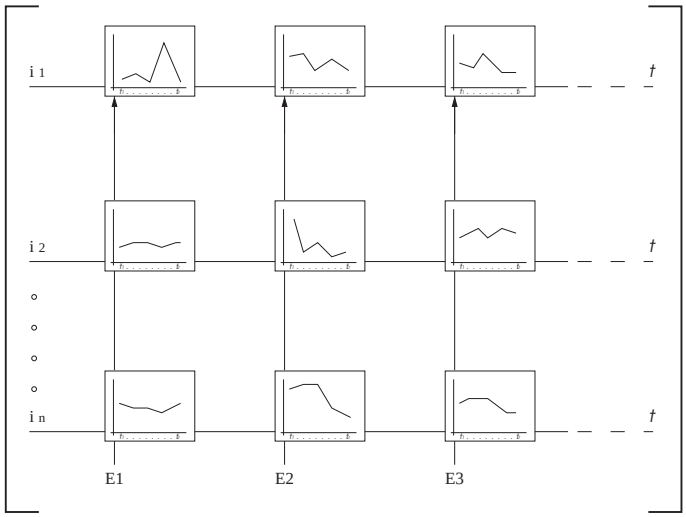


Figura 1.3: Patrón regular de eventos para todos los individuos.

Pero éste no es el caso que se está analizando ahora y por esta razón no es adecuado recurrir a un análisis temporal clásico. Una situación así supondría una hipótesis demasiado rígida para las numerosas situaciones reales que pretendemos cubrir. Sin ir más lejos, la Terapia Electro-Convulsiva (TEC) aplicada a un paciente consiste en un número *variable* de sesiones que dependen del estado de cada paciente, y su distribución en el tiempo se decide bajo criterios médicos e individualizada para cada caso (entre 6 a 12 ES por terapia) y además, la cadencia puede no ser constante durante todo el tratamiento (es frecuente espaciar las sesiones a medida que el paciente mejora, siendo usual 2 ó 3 ES por semana). Por ello, no es conveniente manejar el problema asumiendo esta hipótesis y en consecuencia tampoco es conveniente el análisis temporal clásico a este nivel.

Situaciones así no son, por supuesto, nuevas en los dominios antes mencionados (especialmente en el contexto médico), y hace tiempo que son objeto de estudio formal desde otros campos. En el contexto de las series temporales un método de análisis ampliamente utilizado en casos así parte de las ideas de Matthews [Mat93, Bin97] y consiste en la reducción de las

series de cada individuo a una sola serie que resuma a todas a través de la *media* en cada instante (*línea gruesa* de la figura 1.4(a)), *área media* por serie o *tendencia media* por serie. Esto permitiría reducir las mediciones de Y y Z a un sólo registro por individuo y las matrices X, Y y Z serían compatibles y permitirían un análisis clásico.

Sin embargo, si construimos la serie promedio de cada individuo—entendiendo *promedio* en el sentido amplio mencionado—estaremos con frecuencia perdiendo *demasiada* información relevante, puesto que la variabilidad depende de cada evento y también del *efecto individuo*, y las conclusiones a las que lleve un estudio sobre tal transformación pueden estar muy lejos de la realidad.

Las siguientes figuras permiten relacionar los conceptos, ya expuestos, con la aplicación real que se estudia.

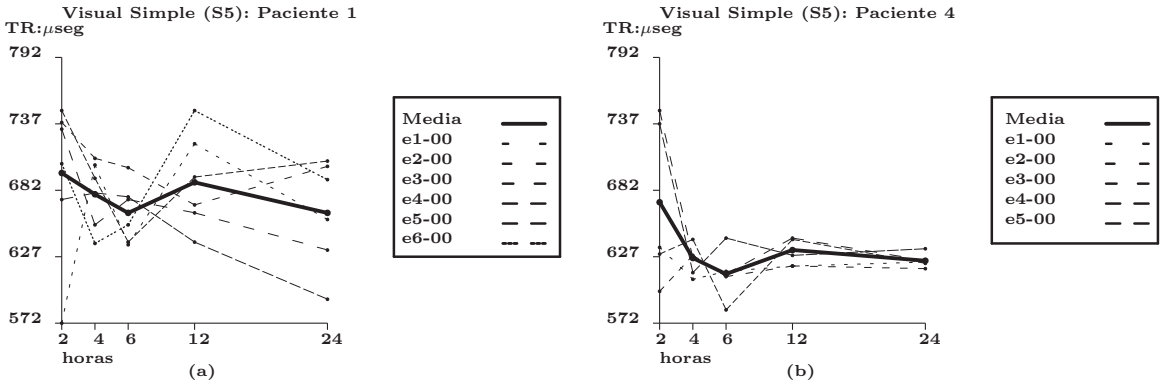


Figura 1.4: Curvas de la Prueba S5 del: (a) paciente 1; (b) paciente 4.

La Fig. 1.4(a) muestra líneas que unen los tiempos de reacción de una *prueba visual simple* (S5) medidos a las 2, 4, 6, 12 y 24 horas después de la aplicación de un ES al paciente 1. Este paciente recibió una TEC de 6 electroshocks y cada curva representa la evolución de su tiempo de reacción.

La Fig. 1.4(b) muestra la evolución de la TEC del paciente 4 (S5). El paciente recibió una TEC de 5 electroshocks.

Como podemos observar en estas figuras, la representación de la evolución del paciente por medio de una curva prototipo (línea gruesa) que corresponde a la media de los tiempos de reacción no es tan fiel a la situación real pues la variabilidad entre los

diversos ES es muy alta, y la curva media no recoge bien lo que en realidad ocurre en los distintos ES. Por lo que, perdemos información relevante. En efecto, considerando una sola línea prototipo por cada paciente se perderían las diferencias entre las reacciones de los pacientes en los diferentes ES. No obstante, lo que sí ofrece esta curva media es una idea de la tendencia general de la evolución del paciente.

De hecho, existe un cambio significativo en las curvas de paciente a paciente y de prueba a prueba, siendo difícil encontrar un patrón general para dichas curvas.

Además, los psiquiatras no han llegado a un acuerdo acerca de la cantidad de ES que deban ser aplicados a un paciente dado. Por lo que, la reducción de la información de los pacientes a un sólo registro en las matrices Y y Z no es el camino a seguir. Por tanto, estamos interesados en mantener todas las curvas de todos los pacientes en dichas matrices tomando en cuenta este *efecto paciente* para el análisis.

Concluyendo, el problema formal que se pretende resolver es el siguiente:

Dado un conjunto de individuos $I = \{i_1 \dots i_n\}$, *un conjunto de atributos (cuantitativos y/o cualitativos)* $X_1 \dots X_K$ *que definen* I, *una matriz* $X = [x_{ik}]_{nK}$, $i = \{1 \dots n\}$, $k = \{1 \dots K\}$, *siendo* x_{ik} *el valor de* X_K *para* i; *dado el atributo de interés que configura las series* $\{Y^t_{i,j}\}$, $t = \{0 \dots r\}$, *para* $i = \{1 \dots n\}$, $j = \{1 \dots m\}$, *la matriz que contiene todas las medidas seriadas en el tiempo* $[Y^t_{ij}]_{Nr}$, $N = \sum_1^i m$, $t = \{0..r\}$; *un conjunto de eventos* $E = \{E_{i,1} \dots E_{i,m}\}$, *un conjunto de atributos (cuantitativos y/o cualitativos)* $Z_1 \dots Z_K$ *que caracterizan a* E, *la matriz* $Z = [z_{ijl}]_{nmL}$, $i = \{1 \dots n\}$, $j = \{1 \dots m\}$, $l = \{1 \dots L\}$, *siendo* z_{ijl} *el valor de* Z_L *para* E; *y considerando,*

- *los individuos* $I = \{i_1 \dots i_n\}$ *actúan como un factor de bloque sobre la matriz* Y,
- *los puntos de medición* $t = \{0..r\}$ *representan una distribución fija en el tiempo para todas las medidas seriadas,*
- *el número de observaciones por medida seriada* r *es pequeño, y*

- *para cada i existe un número variable de medidas seriadas m*

se quiere encontrar un modelo de comportamiento de las medidas seriadas explicitando

- *el (o los) patrón(es) de comportamiento de las medidas seriadas $\{Y_{i,j}^t\}$ y*
- *la relación entre las medidas seriadas $\{Y_{i,j}^t\}$, la matriz X y la matriz Z.*

1.3 Objetivos del Trabajo

Los objetivos se presentan a varios niveles:

1. Relativos a los *dominios poco estructurados* con presencia de medidas seriadas muy cortas y repetidas.
   - Facilitar su estudio.
   - Delimitar una metodología para su análisis, partiendo de su tipo de estructuras (presentadas en §1.2).
   - Obtener resultados fácilmente interpretables por el experto.
2. De orden metodológico.
   - Establecer una nueva metodología híbrida para el descubrimiento de conocimiento (capítulo §2) que utilice herramientas de IA y estadística y permita resolver el problema planteado en la sección §1.2.
   - Obtener un modelo de conocimiento explícito que formalmente describa la estructura del dominio objeto de estudio.

1.4 Estructura del trabajo

La estructura del presente documento está conformado por una introducción a la clase de dominios que se analizan con la metodología KDSM, capítulo §1. Además se expone la motivación que llevó a realizar la investigación, y se introduce el problema de estudio y los objetivos que

se pretenden alcanzar. En el capítulo §2, se presentan una serie de conceptos que dan soporte a la metodología KDSM, se establecen sus pasos y se mencionan las aportaciones en diferentes áreas. En el capítulo §3, se detalla el caso de estudio, relativo a la psiquiatría, que dio origen a la metodología KDSM. En el capítulo §4, se detalla la aplicación de la metodología KDSM al caso de estudio del ámbito laboral. Finalmente, el capítulo §5, contiene las conclusiones del presente trabajo y diversas tareas identificadas como trabajo futuro.

En el apéndice, se encuentra una aclaración importante sobre lo que el(la) autor(a)—del presente trabajo—considera da soporte a su aportación innovadora, debido a que el presente trabajo se clasifica como científico.

Capítulo 2

Metodología para el descubrimiento de conocimiento en medidas seriadas cortas y repetidas (KDSM)

2.1 Introducción

En base a nuestra experiencia en la aplicación real §3, en este capítulo, se presentan en detalle los pasos de la metodología propuesta KDSM, cuya justificación se detalla en el reporte técnico [RGRC01].

La metodología KDSM realiza básicamente tres tareas:

1. Identificación de los diferentes perfiles iniciales de los individuos por medio del estudio de las medidas basales, $Y_{i_0}^t$, y su relación con la matriz X.

2. El conocimiento inducido de la tarea anterior se utiliza como entrada de esta tarea para el estudio del efecto de cierto evento E sobre el atributo de interés Y y el descubrimiento de distintos patrones según las ocurrencias del E en relación a los individuos donde se produce.

3. Los resultados de la segunda tarea se cruzan con la matriz Z para encontrar las relaciones entre ellos y determinar los atributos relevantes en la conformación de los patrones encontrados.

A continuación se presentan una serie de conceptos básicos a fin de dar mayor claridad a la exposición de la metodología KDSM.

2.2 Conceptos Básicos

2.2.1 Resumen de Diseño de Experimentos

El objetivo del diseño de experimentos es estudiar cómo realizar comparaciones, lo más homogéneas posibles, para aumentar la probabilidad de detectar cambios o identificar atributos influyentes sobre cierto fénomeno de interés [Pn89]. Comprobar si un evento mejora un proceso requiere comparar los resultados antes y después de la ocurrencia del mismo. Cuando existe una variabilidad alta entre los resultados—o, en otros términos, un gran error experimental—sólo se detectarán como influyentes, aquellos eventos que produzcan cambios muy grandes con relación al error experimental.

El objetivo de un experimento es estudiar el efecto que sobre un atributo de interés tienen un conjunto de otros atributos, factores o eventos.

En cualquier experimento en que se investiga el efecto de un evento, existen *a priori* un gran número de atributos que pueden influir sobre los resultados y presentan lo que se conoce como *confusión de los efectos*. Conceptualmente existen tres caminos para eliminar el efecto de un atributo:

1. mantenerlo fijo durante toda la realización del experimento;

2. reorganizar la estructura del experimento de manera que las comparaciones de interés se efectúen para valores fijos de este atributo, lo que supone eliminar estadísticamente su efecto y

3. evitar su influencia aleatorizando su aparición en eventos.

Una vez leído el capítulo §1, se puede comprender la importancia que tienen los conceptos del diseño de experimentos en la formulación de nuestro problema (sección §1.2) y el diseño de nuestra metodología. Dado el objetivo de la segunda tarea de la metodología, se debe reorganizar la estructura de datos, del caso de estudio, de forma que se elimine el factor de bloque que ejerce el individuo sobre los eventos y poder estudiar sólo el comportamiento de los mismos.

Para más información sobre el diseño de experimentos ver [Pn89, WMS98]).

2.2.2 Factor de bloque

Se denomina *factor de bloque* [Pn89] al factor (u objeto) que tiene un efecto sobre la respuesta, aunque no es directamente de interés, se debe considerar en el experimento para obtener comparaciones homogéneas en los grupos de observaciones donde dicho factor se mantiene constante.

Para dar un ejemplo haremos referencia al factor de bloque presente en el caso de estudio, dominio psiquiátrico, del capítulo §3. Así tenemos,

> curvas del tiempo de reacción a lo largo de las 24 horas tras la aplicación de un electroshock, el paciente al que se le aplicó el ES no es directamente de interés, pero influye en la representación y hay que tenerlo en cuenta porque determina *bloques* de curvas—las de un mismo paciente—con influencia.

Para evitar la influencia del factor de bloque que conforma el individuo sobre las medidas seriadas, se determinó realizar la diferencia entre las medidas seriadas del evento actual y las medidas seriadas del evento anterior (o viceversa). Así, se puede medir el efecto *per se* de un evento dado sobre el atributo de interés, independientemente de las características del individuo. Estos datos, sólo toman en cuenta el incremento o decremento Y debido a la ocurrencia del evento comparando lo sucedido antes y después de la ocurrencia del mismo.

2.2.3 Prueba de Kruskal Wallis

La mayoría de los métodos estadísticos (métodos paramétricos) para la prueba de *hipótesis estadística*[1] se basan en la suposición de que las muestras aleatorias se seleccionan de poblaciones normales. Sin embargo, a menudo en las situaciones reales nos encontramos ante escasez de conocimiento acerca de las distribuciones de las poblaciones fundamentales. Además en muchas aplicaciones, de la ciencia y la ingeniería, los datos se reportan como una escala ordinal tal que es bastante natural asignar rangos a los datos. Por ello, en este tipo de situaciones, los analistas de datos utilizan con frecuencia métodos estadísticos alternativos (métodos no paramétricos); pues son muy atractivos e intuitivos.

[1] Una hipótesis estadística es una aseveración o conjetura con respecto a una o más poblaciones, más detalles en [WMS98]

Uno de estos métodos no paramétricos es la prueba de Kruskal-Wallis, también conocida como prueba H de Kruskal-Wallis. Introducida en 1952 por W. H. Kruskal y W. A. Wallis, la prueba es un procedimiento no paramétrico para probar la igualdad de las medias en el análisis de la varianza de un factor cuando el experimentador desea evitar la suposición de que las muestras se seleccionaron de poblaciones normales.

> Para probar la hipótesis H_0 de que k muestras independientes son de poblaciones idénticas, calcular
>
> $$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1)$$
>
> Si h es mayor o igual al valor crítico χ^2_α, rechazar H_0 en el nivel de significancia α; en cualquier otro caso, aceptar H_0.

En otras palabras, la prueba de Kruskal-Wallis nos resultó interesante pues, a partir de ella, se puede determinar qué atributos tienen una gran relevancia, estadísticamente hablando, en la conformación de clases como las obtenidas en los casos de estudio de los capítulos §3 y §4. Gracias a esta prueba, se logra optimizar la interacción con el experto, ahorrando su tiempo de análisis pues así el sólo trabaja con los atributos relevantes descartando aquéllos que no son de interés para el objetivo que se pretende lograr con el análisis de los datos de estudio. Para más detalles sobre la prueba de Kruskal-Wallis recomendamos ver [WMS98].

2.2.4 CIADEC

El sistema CIADEC (Caracterización e Interpretación Automática de Descripciones Conceptuales en Dominios poco Estructurados usando atributos cuantitativos) es un sistema híbrido (inteligencia artificial y estadística) que surge de la necesidad de automatizar la caracterización e interpretación de clases en *dominios poco estructurados* previamente particionados. Mediante la automatización de la metodología formal, denominada *Generación Automática de Reglas Difusas en Dominios poco Estructurados con atributos cuantitativos* [VG01a], se pretende reducir el tiempo para la caracterización e interpretación de descripciones conceptuales usando atributos cuantitativos, dando agilidad tanto a las actividades asociadas al análisis de datos como a la obtención de información relevante que posteriormente sea útil en la gestión y toma de decisiones en esa clase de dominios.

Además, la automatización de esta metodología ofrece un conjunto de herramientas que permiten:

- Construir un sistema de reglas.
- Visualizar funciones de pertenencia de un atributo X_k a las distintas clases.
- Evaluar un conjunto de objetos nuevos de acuerdo a las reglas generadas.
- Validar la calidad de las clases resultantes.

En resumen, CIADEC nos permite determinar qué atributos son relevantes, estadísticamente hablando, en la conformación de clases. Gracias a CIADEC, en combinación con la prueba Kruskal-Wallis, se mejora la actuación del experto, pues le ahorra tiempo de análisis (sólo trabaja atributos relevantes). Para más detalles sobre CIADEC recomendamos ver [VG01a].

2.3 Metodología KDSM

Sea Y_T, una matriz de datos que contiene medidas seriadas cortas y repetidas en el tiempo, la podemos representar como: $Y_T = Y_0|Y$

Los pasos que conforman la *metodología para el descubrimiento de conocimiento en medidas seriadas* (KDSM) son los siguientes.

2.3.1 Obtención de la matriz basal Y_0

Basales: Conjunto de medidas seriadas que se toman para cada individuo antes de la ocurrencia del primer evento. Cabe la posibilidad de una excepción en situaciones donde siempre ocurren eventos; en estos casos el analista y el experto determinarán que conjunto de medidas seriadas serán las basales para poder aplicar la metodología KDSM.

Se extrae una matriz de basales Y_0 desde una base de datos que contiene medidas seriadas cortas y repetidas en el tiempo (Tomando $Y_0 = [Y_{i0}^t]$, $i = \{1 \dots n\}$, $t = \{0 \dots r\}$). Esta matriz contendrá los datos que representan las condiciones iniciales de los individuos (curvas *a priori* para cada individuo).

2.3.2 Cluster Jerárquico de Y_0

Se realiza el cluster jerárquico de la matriz de basales Y_0, para encontrar la estructura *a priori* de los individuos, antes de la ocurrencia del primer evento E.

Por ejemplo, con relación al caso que se presenta en el capítulo §3, se identifican grupos de pacientes que antes de iniciar la TEC tuvieran curvas parecidas relativas al atributo de interés Y.

$$Y_0^{\mathcal{P}} = \mathcal{P}(Y_0) = \left|\begin{array}{cccc|c} Y_{10}^1 & Y_{10}^2 & \dots & Y_{10}^r & c_k^{\mathcal{P}} \\ Y_{20}^1 & Y_{20}^2 & \dots & Y_{20}^r & c_l^{\mathcal{P}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{n0}^1 & Y_{n0}^2 & \dots & Y_{n0}^r & c_m^{\mathcal{P}} \end{array}\right| \sim [Y_0|\mathcal{P}]$$

2.3.3 Interpretación de $Y_0^{\mathcal{P}}$ a partir de la matriz X

La interpretación de las clases derivadas de Y_0 se lleva a cabo a partir de la proyección de los atributos característicos de los individuos (matriz X). Buscando aquéllos atributos que son relevantes y que determinan la estructura *a priori*.

Sea $S^{\mathcal{P}}$ un sistema de interpretación que trabajará con la partición obtenida en el paso 2. Mediante dicho sistema (CIADEC) podemos obtener una interpretación de la matriz X.

$$S^{\mathcal{P}}(X) = \left|\begin{array}{cccc|c} x_{11} & x_{12} & \dots & x_{1k} & c_k^{\mathcal{P}} \\ x_{21} & x_{22} & \dots & x_{2k} & c_l^{\mathcal{P}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & c_m^{\mathcal{P}} \end{array}\right|$$

2.3.4 Obtención del conjunto de reglas

La inducción de Reglas se realiza a partir de la comparación entre clases, basándose en los atributos X_k.

El análisis de la partición, del paso anterior, por medio de algún método o metodología (M) como CIADEC, permite obtener un conjunto de reglas $\mathcal{R}^0$, expresadas en lógica de primer orden.

Donde $M : X|Y_0^{\mathcal{P}} \mapsto \mathcal{R}'$, donde $\mathcal{R}'$ conforma la Base de Conocimiento obtenida de los basales,

BC_0.

En este paso, actualmente trabajamos 2 formas de obtención de reglas que reflejen la estructura de los datos:

1. Boxplots múltiples que se utilizan como primera alternativa para la comparación. Éstos se presentan al experto para que él determine qué atributos son de su interés, una vez hecho esto se obtienen reglas, de tipo *crisp*[2], que representan a dichos atributos y que constituyen la base de conocimiento inicial y parcial (BC_0) del dominio; y

2. Obtener los atributos relevantes, estadísticamente hablando, por medio de la prueba Kruskal-Wallis [SC88] y en combinación con la metodología CIADEC [VG01a, VG01b], que emplea teoría de lógica difusa, se obtienen reglas "difusas"que se ponen a consideración del experto para él valore la representación que hacen de la estructura y determine si cuáles de ellas le son de utilidad.

2.3.5 Obtención de la matriz D

Se construye la matriz $D = [Y_{ij}^t - Y_{i,j-1}^t]$ para medir el efecto *per se* de un evento dado sobre el atributo de interés, independientemente de las características del individuo. Estos datos, sólo toman en cuenta el incremento o decremento de Y debido a la ocurrencia del evento comparando lo sucedido antes y después de la ocurrencia del mismo.

Donde $D = [Y_{ij}^t - Y_{i,j-1}^t]; i = \{1 \dots n\}, j = \{1 \dots m\} \therefore \quad D = \begin{vmatrix} d_{11} & d_{12} & \dots & d_{1,m-1} \\ d_{21} & d_{22} & \dots & d_{2,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{n,m-1} \end{vmatrix}$

Según la aplicación concreta podría ser de interés trabajar con $D = [Y_{ij}^t - Y_{i,j-1}^t]$ ó $-D = [Y_{i,j-1}^t - Y_{ij}^t]$ de forma que la diferencia se mantuviera positiva y sea de ayuda a su interpretación.

2.3.6 Clasificación Basada en Reglas (ClBR) de la matriz D

Se lleva a cabo la *clasificación basada en reglas* de la matriz D con la base de conocimiento BC_0. La idea principal de este tipo de clasificación es poder compilar conocimiento adicional

[2]Reglas de lógica de predicados.

del dominio como un conjunto de reglas y usarlo para segmentar el conjunto de objetos en subconjuntos significativos que se mantendrán durante el proceso de cluster (más detalles en [GC97]). Los resultados representarán la estructura que sugiera la BC_0. En este caso, el objetivo es encontrar grupos de individuos con efectos similares del evento E, internos a los grupos de individuos similares al comienzo del proceso (BC_0).

$D^{\mathcal{P}_\mathcal{R}} = \mathcal{P}_\mathcal{R}(D)$, se aplica la metodología de ClBR haciendo uso de BC_0.

Entonces $D^{\mathcal{P}_\mathcal{R}} = [D|\mathcal{P}_\mathcal{R}]$ ó $D^{\mathcal{P}_\mathcal{R}} = \left|\begin{array}{cccc|c} d_{11} & d_{12} & \dots & d_{1,m-1} & c_r^{\mathcal{P}_\mathcal{R}} \\ d_{21} & d_{22} & \dots & d_{2,m-1} & c_s^{\mathcal{P}_\mathcal{R}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{n,m-1} & c_u^{\mathcal{P}_\mathcal{R}} \end{array}\right|$

2.3.7 Interpretación de las clases resultantes de la ClBR de D

Para la interpretación de las clases obtenidas a partir de la ClBR de D caracterizamos el patrón de curva típico de cada clase y estudiamos la variabilidad interna en cada clase y entre ellas, por medio de los gráficos VEC y VIC.

Sea $G^{\mathcal{P}_\mathcal{R}}$ la representación gráfica de $\mathcal{P}_\mathcal{R}$ (usando BC_0), donde los resultados son dos tipos de gráficos que ofrecen una idea general de: a) la variabilidad interna de cada clase (una curva por cada individuo de la clase y una curva media de la clase) y b) la tendencia general de las clases y la variabilidad entre ellas (la curva media de cada clase).

2.3.8 Caracterización e interpretación de la matriz Z

Sea $S^{\mathcal{P}_\mathcal{R}}$ la utilización de CIADEC usando la partición $\mathcal{P}_\mathcal{R}$.

$S^{\mathcal{P}_\mathcal{R}}(Z) = \left|\begin{array}{cccc|c} z_{111} & z_{112} & \dots & z_{11l} & c_r^{\mathcal{P}_\mathcal{R}} \\ z_{121} & z_{122} & \dots & z_{12l} & c_s^{\mathcal{P}_\mathcal{R}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{nm1} & z_{nm2} & \dots & z_{nml} & c_u^{\mathcal{P}_\mathcal{R}} \end{array}\right|$

2.3.9 Análisis de las clases de la ClBR de D

Una vez realizada la proyección de los atributos característicos de Z, paso anterior, sobre las clases generadas a partir de la ClBR en D y encontrados los atributos relevantes y que deter-

minan en algún sentido el comportamiento de los individuos; el experto procede a establecer el significado a los resultados encontrados.

Esto es, el análisis de $S^{\mathcal{P}_\mathcal{R}}$ por medio de CIADEC generará un conjunto de reglas $\mathcal{R}'^0$, expresadas en lógica de primer orden que describen a cada clase.
Donde $M_\mathcal{R} : Z|D^{\mathcal{P}_\mathcal{R}} \mapsto \mathcal{R}'^0$, donde $\mathcal{R}'^0$ conforma la Base de Conocimiento obtenida del análisis de los eventos, $BC_{\mathcal{P}_\mathcal{R}}$, que a su vez será en la que el experto se apoyará para otorgar el significado a las clases encontradas.

2.4 Aportaciones

Inteligencia Artificial Nueva forma de representación gráfica del conocimiento resultante (gráficos VEC y VIC). Generación de conocimiento nuevo que permite ampliar el dominio de conocimiento del experto.

Estadística Medio para abordar problemas donde se presentan medidas seriadas cortas y repetidas con factor de bloque.

Híbridos Nueva metodología para el descubrimiento de conocimiento en dominios poco estructurados donde se presenten medidas seriadas como las mencionadas anteriormente.

2.5 Resumen

Finalmente podemos decir que, la metodología KDSM para el descubrimiento de conocimiento en medidas seriadas muy cortas y repetidas con factor de bloque adapta conceptos de la minería de datos [FPSS96, FPSSU96], de métodos clásicos para el análisis de series de tiempo [PW83, BO93, BJR94], de la estadística [SC88], del diseño de experimentos [Pn89], de la IA de los híbridos. De la minería de datos viene el enfoque del descubrimiento de patrones. Del análisis de series de tiempo viene la teoría para analizar series lineales. Que son, finalmente, las limitaciones de los métodos tradicionales para el análisis de series de tiempo las que abren la posibilidad a nuevos métodos. De la estadística toma el método de Kruskal-Wallis para determinar qué atributos tienen una gran relevancia, estadísticamente hablando, en la conformación

de clases. Del diseño de experimentos proviene la teoría para realizar comparaciones, lo más homogéneas posibles, y aumentar la probabilidad de identificar atributos relevantes sobre cierto fénomeno de interés. Además esta teoría es útil para determinar cómo manejar la influencia de algún factor de bloque. De la IA y los híbridos se toman las bases para analizar patrones, caracterizar grupos, interpretación y representación de resultados, etc.

Capítulo 3

Caso de estudio: Terapia Electroconvulsiva

3.1 Dominio del caso de estudio

Una de las áreas de interés de la psiquiatría corresponde al estudio de la Terapia Electro-Convulsiva (TEC). Dicha terapia es ampliamente aceptada por su seguridad, efectividad y rapidez en su aplicación a trastornos depresivos graves y otros cuadros psiquiátricos [Fin01]. La TEC suele ser un proceso complejo que se basa en la aplicación de algunos electroshocks (ES), dos o tres veces a la semana. El electroshock es una descarga eléctrica que pasa a través del cerebro con una energía suficiente para provocar polarizaciones y despolarizaciones neuronales de forma general (convulsiones), necesarias para obtener una respuesta terapéutica. Sin embargo, los eventos de carácter biológico que suceden en el cerebro y que están relacionados con la efectividad de la terapia son desconocidos.

Los efectos neuropsicológicos y psicofisiológicos de la TEC son su principal inconveniente y consisten en cambios cognitivos que involucran la orientación, atención, capacidad de cálculo y la memoria del paciente (más detalles en [RV94]).

Mucho se ha estudiado sobre la respuesta fisiológica del ES a través de los resultados de pruebas, que se llevan a cabo en el paciente antes y después de la terapia, como: ritmo cardíaco, presión sanguínea, efectos electrocardiográficos, enzimas cardíacas, efectos electroencefalográficos y respuesta hormonal; siendo éstas importantes para la comprensión de los principales efectos que se presentan con la aplicación del ES. Sin embargo, no se ha logrado que converjan en la formalización de una técnica para el análisis de los efectos psicofisiológicos del ES;

además de que existen muy pocos trabajos sobre los efectos de la TEC sobre los parámetros psicofisiológicos como lo es el tiempo de reacción (TR), directamente relacionado con la pérdida de memoria. Desde el campo de la Psiquiatría existe un interés en el estudio de este efecto y una de las formas de abordarlo es estudiando el *tiempo de reacción*[1]. En psiquiatría se quiere conocer más sobre el efecto de una TEC sobre el TR de los pacientes y es sobre este objetivo que nace el presente estudio.

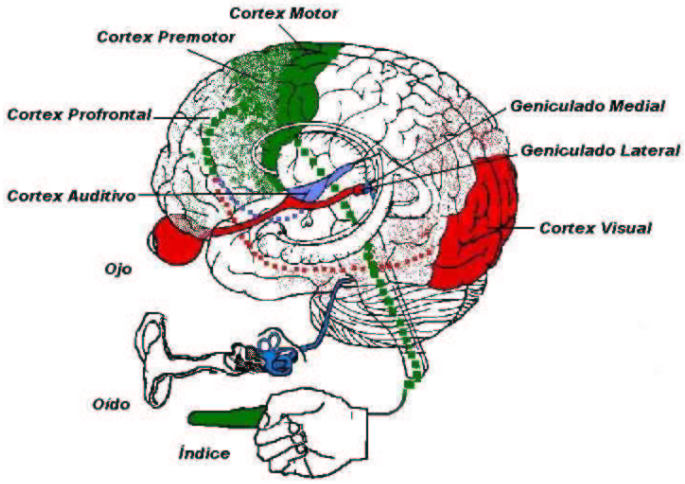


Figura 3.1: Vías nerviosas y sus áreas implicadas.

Existen diferentes tipos de tiempo de reacción los cuales son definidos por la complejidad y clase del estímulo. En la figura 3.1 se pueden observar las diferentes vías nerviosas relacionadas con dichos tipos. En el tiempo de reacción simple se debe responder ante la percepción de un estímulo visual (una luz roja) o auditivo (un tono). En el tiempo de reacción visual complejo se debe responder ante una determinado estímulo (luz roja) de entre varios estímulos visuales (luz roja, amarilla, amarilla + roja). En el tiempo visoauditivo complejo se debe responder ante una determinada señal (luz roja + tono) de entre varias señales (luz roja, amarilla, amarilla + roja, roja + tono, amarilla + tono y amarilla + roja + tono). La reacción a estímulos visuales sean simples (S5) o complejos (S7) utilizan la vía nerviosa visual (en color naranja en la figura 3.1). Los auditivos (S6) usan la vía nerviosa auditiva (color azul) y los visoauditivos (S8) usan ambas vías nerviosas (colores naranja y azul). Los estímulos simples utilizan las áreas cerebrales primarias (color sólido) y los estímulos más complejos deben utilizar, además las áreas cerebrales

[1]Tiempo entre un estímulo y alguna clase de reacción observada en el paciente.

secundarias y terciarias (puntos en color) necesarias para categorizar el estímulo o inhibir la reacción, etc.

La zona de asociación del córtex prefrontal—importante para la toma de decisiones—inicia la respuesta indicando al córtex pre-motor esa decisión, éste diseña el movimiento (respuesta motora) y el córtex motor lo ejecuta, enviando las señales nerviosas a través de la vía piramidal y la médula hasta el sistema muscular. Así, se supone que los tiempos en que reaccionan los pacientes a señales simples son menores a los de las complejas. ¿Qué ocurre con los TR cuando pasa una corriente eléctrica a través del cerebro?

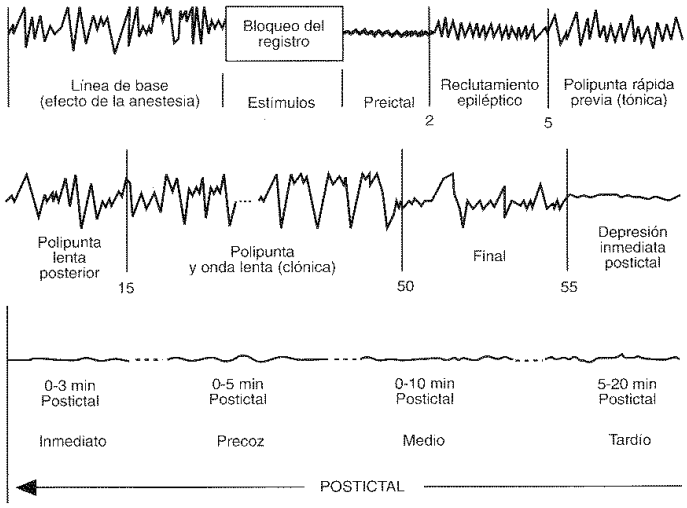


Figura 3.2: Registro electroencefalográfico de las etapas de la TEC.

La figura 3.2 da una idea del complejo proceso, implicado al llevar a cabo la terapia electroconvulsiva, tras la aplicación de un ES. Es así, que el estudio que se presenta se centra en el hallazgo de los perfiles de respuesta de los TR tras la aplicación del mismo. Indudablemente, esto genera información en forma de curvas de difícil interpretación para los Psiquiatras.

Es la primera ocasión, que se tenga noticia, que se estudia la caracterización de los efectos de la TEC sobre los tiempos de reacción tanto visuales como auditivos. De existir un perfil formal de los tiempos de reacción, es nuestro interés delimitar aquellos atributos que tienen una influencia directa en los mismos (y, en consecuencia, en el estado cognitivo del paciente), así como establecer una metodología que permita tratar este tipo de dominios.

3.2 Objetivos del estudio

Encontrar los perfiles de respuesta de los TR tras la aplicación de un ES, así como la identificación de aquéllos atributos que ejerzan una influencia directa en el estado cognitivo del paciente.

3.3 Descripción de los Datos

En este estudio, se da seguimiento a un conjunto 108 ES pertenecientes a 13 pacientes que presentan desórdenes depresivos o esquizofrenia y que se han tratado con TEC durante un cierto tiempo en el servicio de psiquiatría de la Ciudad Sanitaria y Universitaria de Bellvitge (CSUB). Para cada terapia, se optimiza el radio terapéutico seleccionando parámetros del estímulo eléctrico tales como: nivel de energía, duración del estímulo, amplitud y frecuencia de los pulsos. Además se monitorizan múltiples respuestas de los pacientes mediante ElectroEncefalGrama(EEG) y ElectroCardioGrama(ECG), oximetría y tensión arterial; y se lleva a cabo una evaluación clínica completa de los pacientes.

Tras cada sesión se evalúan una serie de variables psicofisiológicas usando la Unidad de Reacción Viena por medio del paradigma de Sternberg [Ste69] (más detalles en [Sch92]). El paradigma de Sternberg es un protocolo estándar para la medición, entre otros, de los *tiempos de reacción*[2] basados en estímulos visuales y auditivos: *Tiempo de reacción visual simple* (prueba S5), donde se responde a un estímulo visual (una luz de un color específico) que se presenta en intervalos irregulares. *Tiempo de reacción auditivo simple* (prueba S6), donde se responde a un estímulo auditivo (un tono audible) que se presenta también en intervalos irregulares. En las pruebas complejas, se requiere de un proceso de categorización después de la percepción del estímulo para poder decidir entre reaccionar o no. Por lo tanto, el tiempo de reacción ante un estímulo complejo incluye el tiempo en que se lleve a cabo dicha categorización. Así, en realidad son *tiempos de categorización y de reacción visual: Visual Complejo* (prueba S7). *Tiempos de categorización y de reacción visoauditivo: Visoauditivo Complejo* (prueba S8).

Para cada prueba se presentaron 32 estímulos, cada uno de ellos de un segundo de duración

[2]En este caso, el tiempo de reacción (TR) es el tiempo entre la emisión de un estímulo y el momento en que el paciente presiona un botón con su dedo índice.

con intervalos de 1 y 1.5 segundos. Como resultado de la prueba se registró la mediana de los tiempos de reacción (en microsegundos) debido a la variabilidad de las medias y la prueba fue realizada en varias ocasiones a cada paciente: el día previo a la TEC (basales) y después de cada sesión hasta el día siguiente de la última TEC a las 2, 4, 6, 12 y 24 horas, después de la aplicación de cada electroshock.

Los parámetros que conforman a la matriz Y se registraron para cada prueba de S5 a S8 (más detalles [RGR01c]) y son los siguientes:

- *Tiempos de Reacción.* Para cada ES aplicado a cada paciente—en microsegundos— $TRhhS$ representa el TR después de hh la(s) hora(s) de la sesión ES, en la prueba S. Donde $hh \in \{2, 4, 6, 12, 24\}$ y $S \in \{S5, S6, S7, S8\}$.

De hecho, para cada ES se ha registrado una gran cantidad de información de distinta índole:

- *Identificador del paciente.*
- *Características de los Electroshocks*: energía aplicada, impedancia, características del estímulo eléctrico (frecuencia, duración, localización).
- *Variables medidas después del Electroshock*: descenso de los niveles de oxígeno, tensión del oxígeno y tensión arterial.
- *Complicaciones médicas.* Complicaciones que presentó el paciente después de la aplicación de un ES, si las hubo (confusión, dolor de cabeza y arritmias).
- *Variables psicofisiológicas.* Resultados del protocolo VIENA a las 2, 4, 6, 12 y 24 horas después de aplicado el ES para todas las pruebas de S5 a S8. Además, se tienen los basales del protocolo VIENA (medidas antes de iniciar la TEC), los cuales constituyen la matriz de basales Y_0.

Además se dispone de información adicional sobre cada paciente, tal como edad, peso, escolaridad, rayos X de tórax, resultados de análisis de sangre y orina, electroencefalogramas y electrocardiogramas previos.

En [RGR01c] se podrá encontrar la descripción detallada de dichos atributos.

$Y_0 =$

patient	s5timer2	s5timer4	s5timer6	...	s8timer6	s8tmer12	s8tmer24
pac01	570.00	607.00	492.00	...	540.00	566.00	554.00
pac02	655.00	643.00	684.00	...	849.00	764.00	818.00
pac03	655.00	482.00	524.00	...	518.00	473.00	466.00
pac04	890.00	931.00	883.00	...	908.00	498.00	916.00
pac05	784.00	796.00	808.00	...	950.00	950.00	963.00
pac06	424.00	414.00	420.00	...	526.00	478.00	552.00
pac07	686.00	500.00	483.00	...	545.00	630.00	666.00
pac08	592.00	633.00	628.00	...	810.00	814.00	840.00
pac10	612.00	697.00	644.00	...	835.00	840.00	810.00
pac11	568.00	502.00	679.00	...	870.00	776.00	901.00
pac12	725.00	829.00	707.00	...	817.00	936.00	914.00
pac13	604.00	624.00	658.00	...	555.00	624.00	629.50

Tabla 3.1: *Muestra de la Matríz de TR basales.*

3.4 Análisis usando la metodología KDSM

Considerando lo expuesto en la sección §1.2, estamos interesados en determinar si existen diferentes patrones en los tiempos de reacción y sus posibles relaciones con las características de los pacientes.

Se ha justificado el motivo por el cual las matrices X y Y no pueden convertirse en una sola; por lo que, se analizarán por separado aplicando la metodología KDSM §2. Resultados de esta aplicación se encuentran en [RGR01a, RGR01b, RGRC01]

3.4.1 Extracción de Y_0 de la matriz Y

El primer paso, consiste en extraer la matriz de tiempos de reacción basales Y_0 de la matriz Y. Esta nueva matriz Y_0 (ver Tabla 3.1) contendrá los datos que determinan las condiciones iniciales de los pacientes antes de someterse a la TEC (patrones *a priori* de cada paciente).

Los datos de la Tabla 3.1, corresponden a los tiempos de reacción medidos para cada prueba psicofisiológica (S5-S8) y cada paciente. Cada fila contiene los tiempos de reacción de un paciente

antes de iniciar la TEC y las columnas se distinguen por cuatro grupos: s5timerXX, s6timerXX, s7timerXX y s8timerXX; donde XX indica la hora de la medición (2, 4, 6, 12 ó 24) a contar desde la aplicación del último ES.

3.4.2 Clasificación Jerárquica de los pacientes usando Y_0

La clasificación de la matriz Y_0 se llevo a cabo con un *método de clasificación jerárquica* (ver [RGR01c]). Una de las técnicas jerárquicas comunes es el *método de Ward* [War], el cual se basa en la minimización estadística de la expansión de la agrupación. Una característica importante por la cual se utiliza el método de Ward, es que preserva el máximo de homogeneidad dentro de los grupos formados, a la vez que el máximo de heterogeneidad entre las agrupaciones.

El resultado obtenido lo representa el árbol jerárquico (dendrograma) de la figura 3.3.

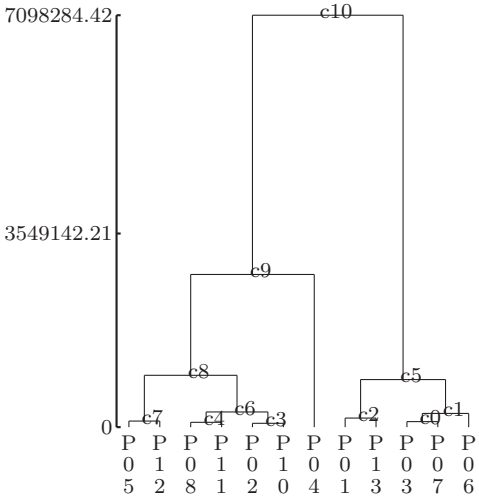


Figura 3.3: Árbol de la clasificación de Y_0.

Debido a la estructura obtenida, se realizó un corte en 3 clases (c5,c8 y P04) usando criterios heurísticos clásicos en el ámbito de la clasificación automática [Vol85].

Quedando las clases de la siguiente manera:

- *Clase c5*=$\{$pacientes: 1,3,6,7 y 13$\}$,
- *Clase c8*=$\{$pacientes: 2,5,8,10,11 y 12$\}$, y
- *Clase p04)*=$\{4^o$ paciente$\}$.

Con esta información procedimos a analizar las características de cada grupo formado.

Descripción de las Clases. A continuación, en la figura 3.4, se muestra para cada clase y prueba la evolución general de los tiempos de reacción en un período previo a la TEC de 24 horas. Cada curva en la figura está compuesta por la media de los tiempos de reacción de todos los pacientes de la misma clase y en una cierta prueba. Como se ha comentado, el uso de curvas promedio que sinteticen el comportamiento de todos los pacientes en una clase dan una idea de la tendencia general que siguen las clases durante la terapia y son útiles para evaluar alrededor de qué rango se sitúan las mediciones de cada clase. Sin embargo, es importante resaltar que la variabilidad dentro de una clase puede ser muy alta y este tipo de síntesis podrían no ser prototipos ideales, con lo cual no deberían usarse como representantes de clase.

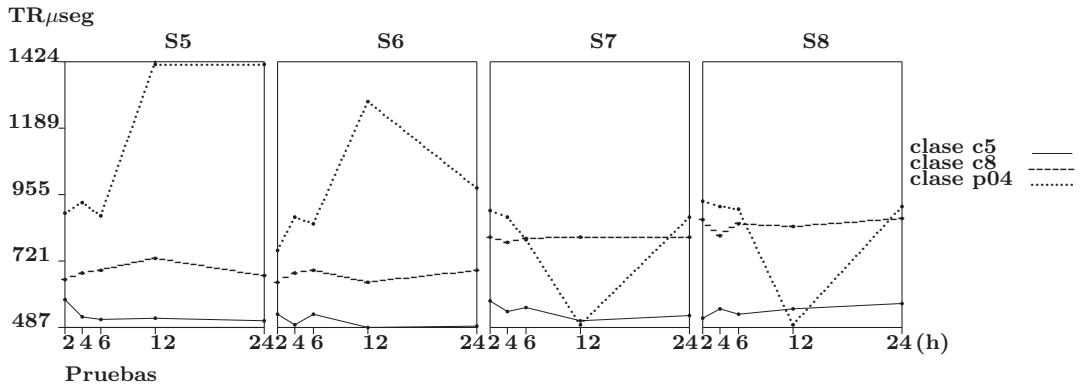


Figura 3.4: Curvas de las 3 clases para las 4 pruebas (S5-S8).

En la Fig. 3.4 observamos que una de las curvas (clase p04, línea a puntos) presenta una forma particular. Dicha curva pertenece a una clase de un sólo paciente, el paciente 4. Este paciente reaccionó a las diversas pruebas de forma muy singular debido a que se le había suministrado cierto fármaco y presentaba un cuadro orgánico con múltiples imágenes de infarto cerebrales, en el escáner, antes de efectuar las pruebas basales. Por otra parte, la tendencia general es que la clase c5 (línea continua) está formada por tiempos de reacción que se mantienen regulares para todas las pruebas. La clase c8 (línea a guiones) contiene tiempos de reacción en media más altos que la clase c5 y además para las pruebas complejas (S7 y S8) se produce un incremento mucho mayor respecto a los de las pruebas simples.

3.4.3 Interpretación de las clases utilizando la matriz X.

Como mencionamos antes, contamos con información adicional específica sobre cada paciente, matriz X, de la cual analizamos sus atributos en busca de aquéllos que identifiquen patrones que distingan claramente a las clases. Algunos de estos atributos se han usado con carácter ilustrativo para la interpretación de las clases. El estudio de su distribución en las distintas clases permite encontrar rasgos distintivos de los mismos. Se han realizado *boxplots múltiples* [Tuk77] de cada atributo *ilustrativo* versus las 3 clases.

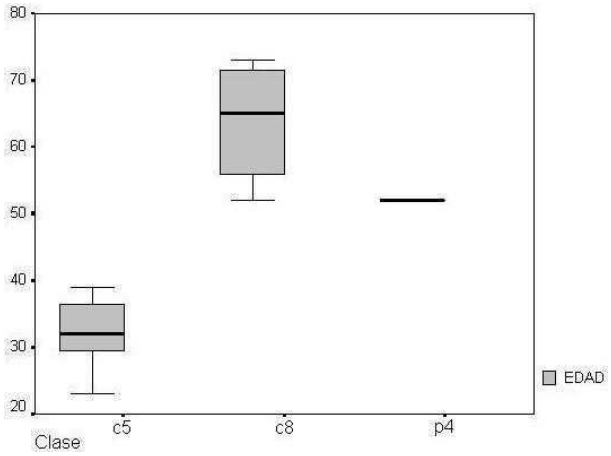


Figura 3.5: Boxplot múltiple de la edad.

El boxplot múltiple es una herramienta gráfica introducida por [Tuk77] que permite comparar la distribución de un atributo cuantitativo en las distintas clases. Para cada clase, el intervalo de valores que toma el atributo cuantitativo se visualiza desplegando una caja desde Q1 (primer cuartil) hasta Q3 (tercer cuartil) con una marca sobre la mediana. Cada caja incluye el 50 % central de los elementos de la clase, mientras que el resto está distribuído de la caja hacia las líneas ("bigotes") que salen de las cajas. Por otra parte, las observaciones extrañas (aisladas) se marcan con un "*".

La figura 3.5, visualiza la distribución de la *edad* en las 3 clases obtenidas de la clasificación anterior. De acuerdo con el experto y revisando la situación presentada en dicha figura, se concluye que el atributo *edad* tiene una influencia importante en el comportamiento de las curvas basales. Como se puede observar, la clase c5 agrupa *todos* los basales de los pacientes jóvenes (menores de 40 años), la clase c8 contiene los basales de los pacientes mayores a 50 años. La

clase p04 contiene solamente el basal del paciente 4. Tras confirmar su peculiar comportamiento y la influencia de la *edad* en las pruebas psicofisiológicas [Lez95] el experto decidió omitir la clase p04 del estudio. Del estudio de los tiempos de reacción basales se ratifica y se concluye que, en efecto, existen dos grupos de pacientes que parten de situaciones distintas:

1. los jóvenes, con tiempos de reacción basales menores y regulares en todas las pruebas (S5-S8) y

2. los de edad más avanzada, con tiempos de reacción basales mayores y presentando tiempos más lentos en las pruebas complejas (S7 y S8).

De aquí, el siguiente paso será derivar las reglas.

3.4.4 Inducción de Reglas

Se obtuvieron dos reglas sencillas, que describen a las clases c5 y c8 mencionadas en el paso anterior:

$$BC_0 = \begin{cases} \text{Si EDAD} \leq 40 \rightarrow \text{clase c5 (Basales de los pacientes jóvenes)} \\ \text{Si EDAD} > 50 \rightarrow \text{clase c8 (Basales de los pacientes mayores)} \end{cases} \tag{3.1}$$

Como se puede observar no tenemos un conjunto de reglas completo, puesto que la gente entre 40 y 50 años no se incluye en ninguna de las reglas. De hecho, no hay evidencias empíricas en el conjunto de datos que proporcionen alguna información que permita determinar si el comportamiento de dicho grupo es similar al de la clase c5 o la clase c8, es decir, en el conjunto de datos no existen pacientes cuya edad esté entre los 40 y los 50 años. Algunas veces, para este tipo de casos se utilizan Hipótesis de Mundo Cerrado [Nn90], que de forma más o menos arbitraria generalizan las reglas obtenidas hasta cubrir todos los casos y completar la Base de Conocimiento. Sin embargo, esto no es conveniente para nuestros propósitos, por lo que, mantendremos abierta esta brecha por el momento. Inclusive, es claro que esta BC_0 no es suficiente para afrontar cualquier método de BC en un sentido clásico, debido principalmente a que es incompleta. Sin embargo, en los siguientes pasos se verá cómo esta BC_0 parcial y simple será suficiente para hallar grupos de significado claro usando *clasificación basada en reglas* y que

a pesar de su incompletitud el resultado de esta clasificación podrá ser completo, sin necesidad de hipótesis artificiales.

3.4.5 Construcción de la matriz de diferencias D

Como se ha mencionado los tiempos de reacción se miden a las 2, 4, 6, 12 y 24 horas después de la aplicación de cada ES a cada paciente. Así cada grupo de estas medidas representa la serie de medidas en el tiempo y como el protocolo determina que sólo se hacen 5 observaciones hace que dicha serie sea muy corta. Estas series se repiten para cada paciente tras la aplicación de cada ES y por supuesto sin independencia entre ellas, puesto que hay grupos de series que pertenecen a un mismo paciente.

Entonces, el estudio de los efectos de la TEC se puede analizar comparando los tiempos de reacción de cada paciente antes y después de una determinada TEC. Por tanto, para clasificar los *efectos* de cada electroshock en los tiempos de reacción, se construyó una nueva matriz D (ver Tabla 3.2) que contiene las diferencias entre los tiempos de reacción medidos tras la aplicación de un ES y aquéllos que se registraron cuando el ES anterior fue aplicado. Se usó esta nueva matriz D que ya mide directamente el efecto del electroshock eliminando lo que sería el efecto de cada paciente (una forma de manejar el factor de bloque, explicada en §1.2) y mantiene la independencia entre las series en esta nueva matriz.

Esta matriz D (Tabla 3.2) contiene las diferencias de los tiempos de reacción antes y después de cada electroshock. Todas las diferencias están expresadas en microsegundos, la etiqueta pXX-YY-ZZ significa (XX: paciente, YY: ES actual ZZ: ES previo). Cada fila contiene la curva diferencia para un ES dado a un paciente, existiendo cuatro grupos de columnas—s5timerXX, s6timerXX, s7timerXX y s8timerXX; donde XX significa las horas transcurridas tras la aplicación de un ES en que se realizó la medición—que contienen las diferencias por prueba (S5-S8). La información de la matriz D sólo toma en cuenta la mejora o deterioro del tiempo de reacción debido a un electroshock y lo sucedido antes de su aplicación.

3.4.6 Clasificación Basada en Reglas de la matriz D

Antes de presentar los resultados de este paso, es necesario hacer una breve introducción a la *clasificación basada en reglas* y el programa informático que permite llevarla a cabo.

$D =$

id	s5timer2	s5timer4	s5timer6	...	s8timer6	s8tmer12	s8tmer24
p01-01-00	-2	-125	-152	...	-181	-150	-222
p01-02-01	208	8	86	...	21	98	55
p01-03-02	12	-22	-82	...	-14	-154	-67
p01-04-03	-20	-52	47	...	-62	-22	-64
p01-05-04	-78	36	3	...	-104	64.31	-114
p01-06-05	40	-56	-32	...	153	-62.31	94
p02-01-00	73	25	10	...	151	62	104
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
p13-03-02	-70	-60.14	-145.61	...	-161.89	-91.69	-32.3
p13-04-03	0	0	0	...	0	0	0
p13-05-04	-88	-99.86	-103.39	...	-59.11	-101.31	-37.7
p13-06-05	88	99.86	103.39	...	59.11	101.31	37.7

Tabla 3.2: *Extracto de la matriz de diferencias de los tiempos de reacción del ES actual y el anterior.*

Clasificación basada en reglas y $KLASS+$. Las características de los *dominios poco estructurados* [GC97], mencionadas en la introducción, hacen que su análisis, a través de alguna herramienta convencional, sea difícil; por lo que es necesario emplear un método o herramienta que se adecúe a dichas características. Así, la *clasificación basada en reglas* [Gib94] es una metodología, particularmente útil, para el trabajo con este tipo de dominios.

La principal ventaja que ofrece esta metodología es la posibilidad de manejar bases de conocimiento incompletas y poder llevar a cabo un proceso de cluster de forma conjunta; combinando conocimiento previo del experto con un método de cluster automático. Dicha metodología, finaliza obteniendo un conjunto de clases que presumiblemente son interpretables a partir de un conjunto simple de objetos: inicialmente, se da un proceso de adquisición del conocimiento de base disponible, aún y cuando el dominio no este completamente definido, seguido de un proceso de cluster *strictu sensu* que respeta la estructura inducida por la Base de Conocimiento. Algunos detalles y ventajas del uso de esta metodología de clasificación son descritos en [GC98].

La herramienta informática para llevar a cabo el análisis de la matriz D es $KLASS+$. Esta herramienta fue desarrollada en el departamento de EIO de la UPC e implementa la metodología de *clasificación basada en reglas* para encontrar la estructura de un conjunto de datos, usando un algoritmo de cluster conocido como vecinos recíprocos encadenados [Gib94]. En particular, $KLASS+$ permite trabajar con las siguientes métricas: euclídea, euclídea estándar, χ^2 (chi-cuadrada), métrica mixta (introducida por Gibert en [GC98]), Gower [Gow71], Ralambondrainy [Ral95], Diday-Gowda [DG92], la métrica general de Minkowski [IY94].

Aplicación a los datos de estudio. En este caso, $KLASS+$ llevó a cabo la *clasificación basada en reglas* de la matriz D (subsección §3.4.5), haciendo uso de BC_0 (expresión 3.1)—obtenida en la clasificación previa de basales—para clasificar los efectos de la TEC sobre los tiempos de reacción, utilizando la métrica euclídea por ser todas las variables numéricas y el criterio de Ward. Se realizaron dos procesos locales de cluster: uno para los pacientes jóvenes y otro sobre los maduros, integrando después ambas jerarquías y encontrado una partición global tal como se aprecia en el dendrograma obtenido (ver Fig. 3.6), donde la nomenclatura de las clases indica dónde se insertan las clasificaciones locales al dendrograma general. Conviene notar

aquí que la inserción se produce a distintos niveles del árbol de acuerdo a la heterogeneidad (o grado de generalidad) de las clases inducidas por las reglas. En esta aplicación concreta no hay gran diferencia y esto no tiene mayor importancia pero es la clave para que se pueda trabajar libremente con bases de conocimiento no homogéneas [GC93].

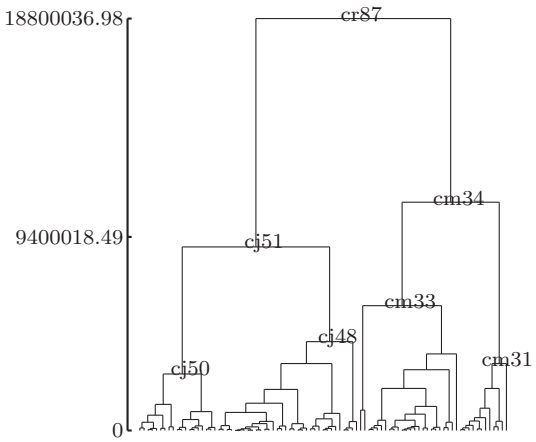


Figura 3.6: Árbol general de clasificación.

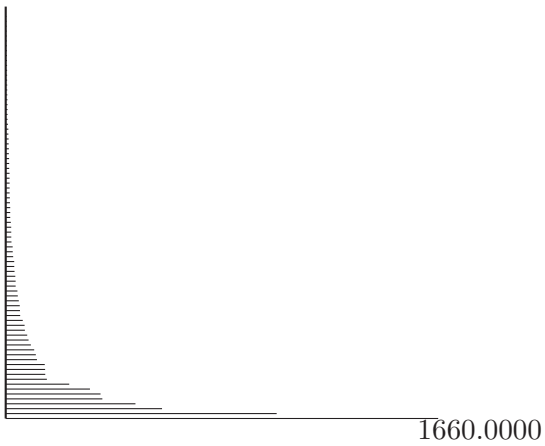


Figura 3.7: Inercia interna de las clases formadas

El análisis de la estructura del dendrograma y su histograma de barras correspondiente figura 3.7, sugieren 4 clases (un corte en 4): dos clases (cj50 y cj48) del grupo de los jóvenes y otras dos (cm33 y cm31) del grupo de los mayores. En este sentido, $KLASS+$ puede recomendar, usando un criterio heurístico, el corte más apropiado según la clasificación obtenida. En este

caso, la recomendación de $KLASS+$ resultó ser una propuesta apropiada.

Las 4 clases están conformadas de la siguiente forma:

- *Clase cj50*={diferencias: p13-03-02, p05-11-10, p13-05-04, p03-02-01, p10-02-01, p08-04-03, p05-10-09, p13-01-00, p01-01-00, p06-07-06, p01-03-02, p05-09-08, p08-08-07, p01-05-04, p06-05-04, p05-06-05, p10-03-02, p08-06-05 y p08-02-01},

- *Clase cj48*={diferencias: p10-05-04, p06-08-07, p10-01-00, p05-01-00, p13-04-03, p03-03-02, p06-09-08, p03-05-04, p03-04-03, p10-06-05, p10-04-03, p05-08-07, p05-02-01, p06-04-03, p05-03-02, p01-06-05, p05-05-04, p13-02-01, p01-02-01, p06-02-01, p06-01-00, p08-01-00, p08-07-06, p08-05-04, p03-01-00, p10-07-06, p01-04-03, p05-04-03, p06-06-05, p08-03-02, p13-06-05, p10-08-07, p06-03-02 y p05-07-06},

- *Clase cm33*={diferencias: p11-05-04, p02-02-01, p02-01-00, p11-03-02, p07-05-04, p12-08-07, p11-02-01, p02-03-02, p12-11-10, p11-04-03, p07-02-01, p02-05-04, p12-10-09, p12-03-02, p07-01-00, p11-06-05, p11-01-00, p02-04-03, p12-04-03, p07-04-03 y p12-06-05} y

- *Clase cm31*={diferencias: p12-12-11, p11-07-06, p12-02-01, p12-09-08, p07-03-02, p02-06-05, p07-06-05, p12-07-06, p12-05-04 y p12-01-00}.

3.4.7 Interpretación de las clases resultantes

A continuación, en la figura 3.8, se aprecia el efecto del ES en los tiempos de reacción de las pruebas S5-S8 para cada clase. Se representó la curva media por clase, obtenida promediando todas las curvas en esa clase, que da idea de la evolución general. Como se ha señalado con anterioridad, este tipo de representación no muestra la variabilidad dentro de las clases por lo que ésta no puede ser evaluada. Sin embargo, nos da idea de la tendencia general que los ES tienen en tales clases durante la terapia.

En la Fig. 3.8 se ve cómo las clases cj50 (línea continua) y cm31 (puntos) contienen los ES cuyas diferencias de TR van disminuyendo respecto del ES anterior, generando diferencias que se sitúan entorno a valores negativos, es decir los pacientes mejoran.

Por el contrario, las clases cj48 (guiones) y cm33 (asteriscos) sitúan sus diferencias de TR entorno a valores positivos lo que es claro indicativo de que el paciente se retarda más después

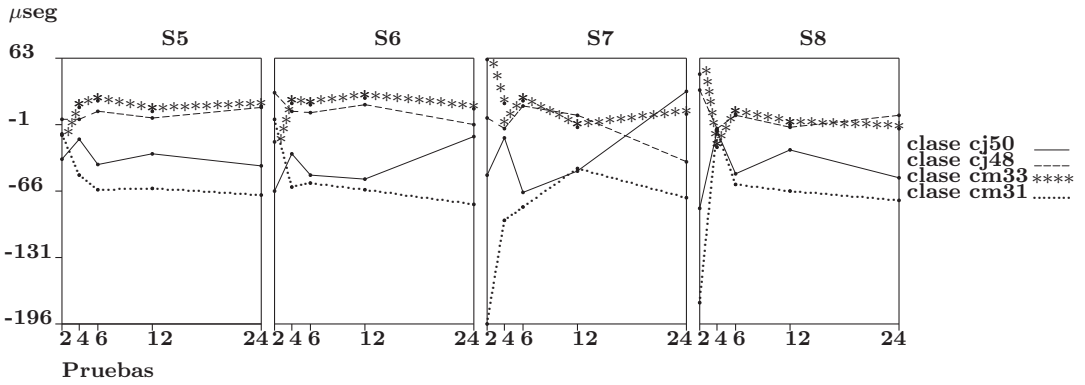


Figura 3.8: Curvas de las 4 clases en las 4 pruebas (S5 a S8).

de cada ES, siendo el tiempo del ES en curso mayor al del ES anterior, es decir hay deterioro en la tendencia de los pacientes. Podemos apreciar que en las pruebas complejas (S7 y S8) existen oscilaciones mayores que en las simples (evolución menos estable a lo largo del tiempo).

Si bien a partir de la figura 3.8 ya se intuye que hay 2 parejas de clases donde los efectos de los ES son parecidos (cj50, cm31 mejoran y cj48, cm33 empeoran), no se pueden reducir solamente a 2 grupos dicha clasificación. La razón es que mientras cj50 y cj48 son clases de ES aplicados a pacientes jóvenes, cm33 y cm31 lo son de ES aplicados a pacientes mayores.

Así, existe una descripción clara del contenido de estas 4 clases, que tienen una semántica muy evidente para un experto en la materia.

- *cj50.* ES aplicados a pacientes *jóvenes* que *reducen* el TR entre un ES y el siguiente.
- *cj48.* ES aplicados a pacientes *jóvenes* que *aumentan* el TR entre un ES y el siguiente.
- *cm33.* ES aplicados a pacientes *mayores* que *aumentan* el TR entre un ES y el siguiente.
- *cm31.* ES aplicados a pacientes *mayores* que *reducen* el TR entre un ES y el siguiente.

Proyección de Z sobre las clases En primer lugar se proyectó el atributo *Identificador del paciente* sobre la partición encontrada.

IDENTIFICADOR DEL PACIENTE En contra de lo esperado, al cruzar la matriz Z con la Y lo primero que sorprende es que los valores de éste atributo no se proyectan asociados a una sola clase.

Id. Pac.	Clase cj50	Clase cj48	Clase cm33	Clase cm31
p01	3	3		
p02			5	1
p03	1	4		
p05	4	9		
p06	2	7		
p07			4	2
p08	4	4		
p10	2	6		
p11			6	1
p12			6	6
p13	3	3		

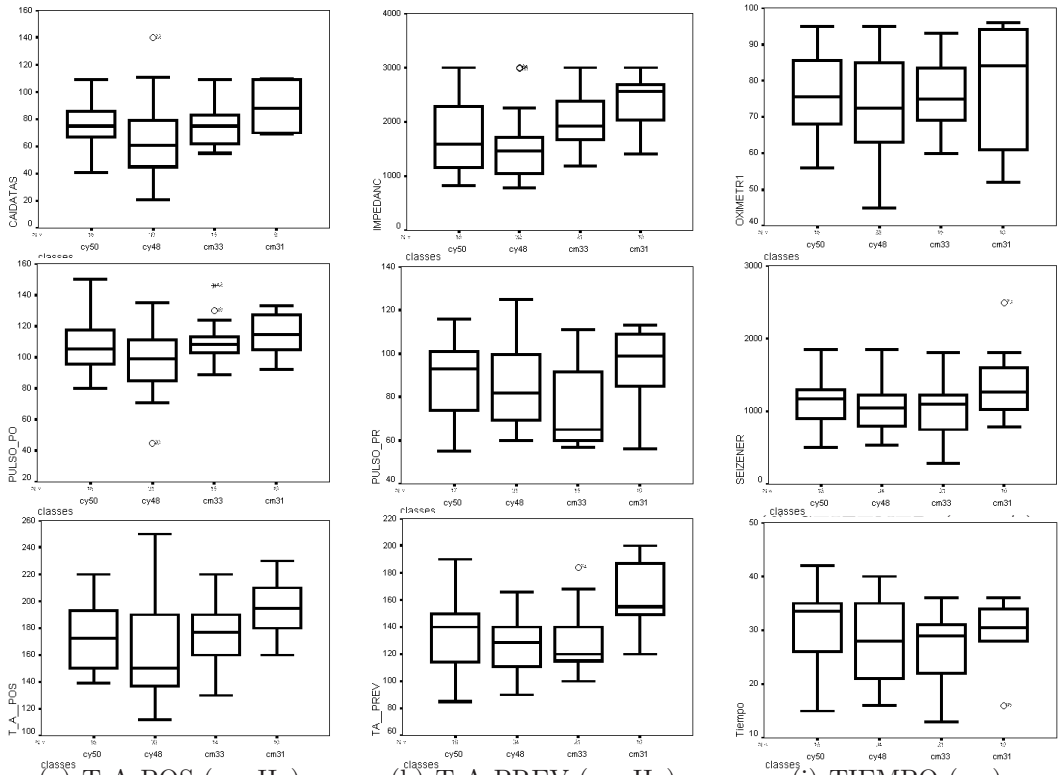
Tabla 3.3: *Relación paciente y cantidad de ES por clase.*

En la tabla 3.3 se puede ver cómo se distribuyen los valores de los ES relativos a cada paciente en las cuatro clases obtenidas, resaltando que se agrupan a las diferentes clases sin necesidad de que exista una asociación del paciente a una sola clase. Así, de los 13 ES aplicados al p05, que es joven, 9 se sitúan en cj48 (que es una clase de diferencias de TR positivas, es decir, el paciente empeora) mientras que los otros 4 se sitúan en cj50 (clase de diferencias de TR más negativas, es decir, el paciente mejora).

Este es un resultado especialmente relevante porque indica que lo que lleva al paciente a mejorar o empeorar tras una sesión de ES no sólo depende de él y el ES puede cambiar el efecto sobre un mismo paciente a lo largo de la TEC. Esta es la primera vez que se constata este resultado como se explicará más adelante en el capítulo §5.1.

Además, con la proyección de la totalidad de Z sobre las clases, se obtuvieron resultados relevantes adicionales de gran interés para el experto. El apéndice B contiene los boxplots múltiples que visualizan la distribución de dichos atributos condicionados a las clases.

De todos los atributos estudiados se presentan aquí los que resultaron relevantes a los objetivos del experto, aunque existen otros que marcan diferencias obvias entre jóvenes y mayores,



(g) T_A_POS (mmHg) (h) T_A_PREV (mmHg) (i) TIEMPO (sec)

Figura 3.9: Boxplots de los atributos relevantes y de interés para el experto.

como podría ser el atributo *Energía* (para más detalles sobre éste atributo ver [RGR01c]) y que justifican la existencia de 4 clases.

Además, se ha probado estadísticamente (usando el test no paramétrico de Kruscal-Wallis) que estas diferencias son realmente significativas.

CAIDATAS Diferencia de la tensión arterial. La tensión arterial, después del ES sube y, es mayor para las clases donde los pacientes mejoran (cj50 y cm31) que en las clases donde empeoran (cj48 y cm33). Ver gráfico 3.9 (a).

IMPEDANC Impedancia. Las clases donde los pacientes empeoran (cj48 y cm33), presentan menor impedancia. Ver gráfico 3.9 (b).

PULSO_PR y PULSO_PO Pulso previo y posterior al ES. Las clases donde los pacientes empeoran (cj48 y cm33), presentan un pulso más bajo de lo normal. Ver gráficos 3.9 (e) y (d).

T_A_PREV y T_A_POS Tensión distólica (previa al ES) y tensión sistólica (posterior al

ES). Las clases donde los pacientes empeoran (cj48 y cm33), presentan ambas tensiones más bajas del nivel esperado. Ver gráfico 3.9 (h) y (g).

Aunque resultan estadísticamente no significativos, el experto ha considerado también interesantes las leves tendencias que se observan en otros 3 atributos.

OXIMETR1 Oximetría (caída de oxígeno). Las clases donde los pacientes empeoran (cj48 y cm33), presentan caída del oxígeno después del ES. Ver gráfico 3.9 (c).

SEIZENER Energía de la convulsión. Las clases donde los pacientes empeoran (cj48 y cm33), tienen menor energía de la convulsión. Ver gráfico 3.9 (f).

TIEMPO Tiempo de convulsión. Las clases donde los pacientes empeoran (cj48 y cm33), tienen menor tiempo de convulsión. Ver gráfico 3.9 (i).

A pesar de que la TEC se utiliza desde 1936, desarrolló mala fama debido a un uso poco cuidadoso y demasiado extendido (usada como técnica de tortura a informadores durante la Segunda Guerra Mundial), lo que hizo que se abandonara su práctica en diferentes países y lugares, sin generar investigaciones a su alrededor. Desde los 80 y, sobretodo, los 90 vuelve el interés por saber más de ella ya que es considerada una terapia muy eficiente. Por tanto de confirmarse, lo que estos atributos indican, se podría hablar de un paso muy importante en la evolución de esta técnica terapeútica. Dado que, al estudiar cada electroshock en sí, se esta más cerca de determinar los factores que provocan alteraciones amnésicas y con ello, más cerca de aumentar su seguridad.

3.5 Resultados

Para la realización de este estudio, se ha partido de dos Bases de Datos clínicas. Una que contiene los atributos característicos del paciente (matriz X) y la otra con las medidas seriadas de los tiempos de reacción resultantes después de la aplicación de cada electroshock y los atributos característicos de los electroshocks (matrices Y y Z).

Se aplicó la metodología KDSM sobre estos datos, realizando un análisis en 3 fases:

1. *Caracterización de pacientes.* Se analizaron los tiempos de reacción basales debido a que representan las condiciones iniciales con que los pacientes se enfrentan a las pruebas. El objetivo de este paso fue encontrar información útil para descubrir los perfiles de pacientes que están en tratamiento con TEC y observar si existe una estructura *a priori* en el conjunto de pacientes que pueda determinar diferencias en los efectos del ES. Se derivaron dos reglas (expresión 3.1) que delimitan al grupo de pacientes jóvenes y al de los mayores.

2. *Análisis del efecto de cada electroshock aislado.* La clasificación basada en reglas se aplicó utilizando como base de conocimiento los resultados obtenidos en el análisis previo. Se utilizaron las diferencias de los tiempos de reacción para eliminar el efecto paciente y estudiar aisladamente el efecto ES. Este estudio nos revela directamente cómo es la evolución de los pacientes tras cada ES durante la terapia. En este caso en particular, se aprecia que el uso de la clasificación basada en reglas, a pesar de usar un conjunto pequeño de reglas simples, mejora la calidad de los resultados y la interpretabilidad de las clases a partir de la combinación de la base de conocimiento con procesos de cluster. Por otra parte, el uso aislado de una técnica de cluster no hubiera permitido jamás incorporar la *edad* como criterio de clasificación en el análisis de los tiempos de reacción, ya que ésta es un atributo que no está presente en la matriz Y ni en la matriz D. En cuanto a las técnicas de Inteligencia Artificial, ya se comentó que con una BC tan simple difícilmente se hubiera podido abordar el problema.

   Finalmente, se detectaron 4 clases que distinguen dos grupos de ES: los aplicados a jóvenes y a los mayores. En dos clases hay pacientes jóvenes y en otras dos mayores, que se subdividen según si su tendencia es a empeorar los tiempos de reacción, o bien a mejorarla.

3. *Proyección de Z sobre P.* Para finalizar el análisis, se realizó la proyección de todos los atributos en Z sobre las 4 clases obtenidas en la fase anterior. Por medio de esta proyección, se identificaron algunos atributos relevantes. Unos marcan diferencias entre pacientes jóvenes y mayores. Otros entre los que mejoran y los que no. Evidentemente es este último grupo el que resulta más relevante para el experto e identifica un conjunto

de factores asociados al hecho de que un ES produzca mejoras. Este resultado, pone de manifiesto la necesidad de llevar a cabo más estudios de la TEC; ya que de confirmarse las tendencias que estos atributos muestran, se podría mejorar—de forma importante—la eficiencia de dicha terapia.

Capítulo 4

Caso de Estudio: Chihuahua

4.1 Dominio del caso de estudio

El Servicio Estatal de Empleo, en un esfuerzo conjunto de la Secretaría del Trabajo y Previsión Social y los Gobiernos de los Estados, lleva a cabo programas de sentido social que tienen por objetivo promover el diseño y aplicación de políticas orientadas a la generación de empleo.

Para lograr este objetivo, se implementó el Programa de Becas de Capacitación para Trabajadores Desempleados (PROBECAT).

PROBECAT ofrece cursos de capacitación de corta duración (de 1 a 3 meses) en coordinación con los sectores productivo y educativo en diversas especialidades, para personas que se encuentran desempleadas y que no están estudiando en ningún nivel educativo.

El objetivo es proveer conocimientos actualizados a los participantes para poder ser puestos en práctica de manera inmediata mediante el apoyo de material de práctica, un instructor y una beca mensual correspondiente al salario mínimo como incentivo para lograr su permanencia en el curso.

El programa de becas tiene diferentes modalidades de capacitación siendo las siguientes:

Capacitación Mixta Que consiste en la realización de cursos a petición expresa y en coordinación con el sector empresarial, para satisfacer requerimientos específicos de personal calificado.

Capacitación Mixta en las Micro y Pequeñas Empresas Esta capacitación está orientada a capacitar y generar experiencia laboral a población joven buscadora de empleo, aprovechando la infraestructura productiva de las micro y pequeñas empresas.

Capacitación Escolarizada Consiste en impartir conocimientos teóricos y prácticos en las especialidades demandadas por los sectores productivos de la región en planteles educativos y de enseñanza.

Capacitación de Autoempleo Su propósito es promover alternativas de ocupación para personas desempleadas de zonas urbanas y rurales, para que adquieran conocimientos, habilidades y destrezas que les permitan iniciar una actividad por cuenta propia

Proyecto de modernización de la educación técnica y la capacitación Busca cubrir los requerimientos de calificación de los trabajadores para mejorar los niveles de productividad y competitividad de las empresas, así como ampliar las posibilidades de incorporación, desarrollo y permanencia de las personas en el empleo.

Iniciativas locales de empleo Opera con productores de escasos recursos agrupados en torno a un proyecto productivo que ya tengan desarrollado, y tiene como objetivo mejorar los conocimientos y habilidades de los integrantes del grupo para ejecutar el proyecto.

La República Mexicana es una entidad federativa de 31 Estados y un Distrito Federal compuesta de 2427 municipios con una extensión territorial en total de 1 964 375 km^2. Lo anterior implica que exista una gran diversidad de cursos con el fin de atender todas las necesidades particulares del sector productivo de cada uno de los municipios, así como abatir el índice de desempleo en los mismos.

Con el fin de que PROBECAT responda a las expectativas de las tres entidades gubernamentales (Gobierno Federal, Estados y Municipios) en México, se debe monitorizar de forma adecuada los progresos obtenidos tras la impartición de cada curso. El análisis de esta información, compuesta por: las características de los Municipios (matriz X), de la cantidad de colocados (medidas seriadas, matriz Y) en un período fijo de tiempo (3 meses), en seis ocasiones (una cada 15 días) y, las características de los cursos (matriz Z) no es trivial; pues tenemos conjuntos de medidas por cada uno de los tipos de curso existentes (180 especialidades de cursos aproximadamente). Donde cada uno de los municipios actúa como un factor de bloque sobre las medidas seriadas y las características de los cursos. Es decir, por cada municipio existe un paquete de medidas seriadas y otro de características de los cursos que se impartieron en él.

En resumén, tenemos que por cada uno de los 2427 municipios existen conjuntos de medidas relativas al atributo de interés (que en este caso de estudio corresponde al índice de contratación) para cada una de las 180 especialidades de cursos, además de las características tanto de los municipios como de los cursos.

4.2 Objetivos del estudio

Distinguir los atributos que ejercen una influencia directa (información verdaderamente importante) para medir la efectividad de PROBECAT.

Por otra parte, como análisis piloto del presente estudio se eligió trabajar con la información del Estado de Chihuahua, debido a que la experta, MCE. Gabriela Alvarado, ha colaborado directamente con el Departamento del Servicio Estatal de Empleo del gobierno Chihuahuense. Concretamente el estudio de Chihuahua se inicio analizando un curso de la rama textil ya que es uno de los más solicitados y del cual el Departamento del Servicio Estatal de Empleo tiene como objetivo adicional obtener mayor conocimiento sobre su comportamiento para poder incidir en su funcionamiento de manera más oportuna y eficaz.

4.3 Descripción de los datos

El conjunto de matrices de datos CUU contiene información sobre los atributos característicos de 68 municipios del Estado de Chihuahua en México (matriz X), medidas seriadas relativas a la cantidad de personas colocadas (ocupadas) en un puesto de trabajo (matriz Y) y los atributos característicos de los cursos de capacitación (matriz Z).

4.3.1 Descripción de los atributos en la matriz X

Atr.	Etiqueta	Descripción	Tipo
1	ni	Número de identificación de cada municipio	Numérico
2	municipio	Territorio que pertenece al Estado[1]	Cadena
3	p.total	Población total del municipio	Numérico
4	hombres	Total de hombres que habitan en el municipio	Numérico
5	mujeres	Total de mujeres que habitan en el municipio	Numérico
6	p12+	Total de la población a partir de 12 años	Numérico
7	pea	Población económicamente activa	Numérico
8	pea ocupada	Población económicamente activa ocupada[2]	Numérico
9	pea desocupada	Población económicamente activa desocupada	Numérico
10	alfabetas masculinos	Alfabetas masculinos (mayores de 15 años)	Numérico
11	alfabetas femeninos	Alfabetas femeninos (mayores de 15 años)	Numérico
12	analfabetas masculinos	Analfabetas masculinos (mayores de 15 años)	Numérico
13	analfabetas femeninos	Analfabetas femeninos (mayores de 15 años)	Numérico
14	ue sector manufactura	Establecimientos manufactureros en el mpio.	Numérico
15	ue sector comercio	Establecimientos comerciales en el municipio	Numérico
16	ue sector servicios	Establecimientos de servicios en el municipio	Numérico
17	ue sector minero	Establecimientos mineros en el municipio	Numérico
18	total de escuelas	Total de escuelas en el municipio	Numérico
19	tot. de viviendas hab.	Total de viviendas habitadas	Numérico
20	ocu. en viviendas part.	Ocupantes en viviendas particulares	Numérico
21	prom. de ocu. por vivienda	Promedio de ocupantes por vivienda	Numérico
22	t. de red carretera (km)	Total km de red carretera en el municipio	Numérico
23	ext. territorial (km^2)	Área en km^2 que comprende el municipio	Numérico
24	tmax	Temperatura media máxima (°C)	Numérico
25	tmin	Temperatura media mínima (°C)	Numérico
26	tmed	Temperatura media anual (°C)	Numérico
27	precip. med. anual (ml.)	Precipitación pluvial media anual (mililitros)	Numérico

4.3.2 Descripción de los atributos en la matriz Y

Atr.	Etiqueta	Descripción	Tipo
1	ec	Etiqueta del curso	Numérico
2	ic1	Medida seriada basal correspondiente al no. de personas colocadas	Numérico
3	ic2	1^{er} Medida seriada correspondiente al no. de personas colocadas	Numérico
4	ic3	2^{da} Medida seriada correspondiente al no. de personas colocadas	Numérico
5	ic4	3^{ra} Medida seriada correspondiente al no. de personas colocadas	Numérico
6	ic5	4^{ta} Medida seriada correspondiente al no. de personas colocadas	Numérico
7	ic6	5^{ta} Medida seriada correspondiente al no. de personas colocadas	Numérico

[1]Dicho territorio lo delimita la división política del Estado.

[2]Se consideran ocupadas aquéllas personas que trabajan al menos una hora a la semana, y las que realizan trabajos en el campo, aunque sean parcelas propias.

4.3.3 Descripción de los atributos en la matriz Z

Atr.	Etq.	Descripción	Tipo
1	ec	Etiqueta del curso[1].	Cadena
2	no.aut.	Número de autorización de cada curso.	Numérico
3	c.por rama econ.	Especificación del ramo económico de cada curso.	Cadena
4	especialidad	Actividad específica en que se desarrolla el curso.	Cadena
5	centro de capacitación	Lugar en donde se imparte el curso.	Cadena
6	localidad	Poblado en donde se desarrollan los cursos.	Cadena
7	municipio	Territorio que pertenece al Estado	Cadena
8	u.op	Oficina coordinadora de los cursos	Numérico
9	prog.	Personas programadas presupuestálmente en cada curso.	Numérico
10	t.i.	Total de personas inscritas en cada curso	Numérico
11	i.h.	Hombres inscritos	Numérico
12	i.m.	Mujeres inscritas	Numérico
13	t.e.	Total de personas egresadas de cada curso.	Numérico
14	e.h.	Hombres egresados	Numérico
15	e.m.	Mujeres egresadas	Numérico
16	t.c.	Total de personas colocadas[2].	Numérico
17	c.h.	Hombres colocados	Numérico
18	c.m.	Mujeres colocadas	Numérico
19	f.inicio	Fecha en que inicia el curso	Cadena
20	f.termino	Fecha en que termina el curso	Cadena
21	duración	Tiempo de duración de cada curso[3].	Cadena
22	horario	Turno en que se lleva a cabo el curso[4].	Cadena
23	inversión	Cantidad monetaria invertida en los cursos[5].	Numérico
24	modalidad	Modo de impartición del curso[6].	Cadena

4.4 Análisis usando la metodología KDSM

Para el análisis de los datos del Gobierno de Chihuahua, se aplicó la metodología KDSM lográndose sus 3 tareas principales:

1. la *caracterización* de la estructura conformada por los municipios sobre el primer índice de contratación, es decir, el establecimiento de las condiciones iniciales;

2. el *análisis del efecto de cada curso aislado eliminando el factor de bloque* que conforman los municipios; y

[1]Define curso, municipio y orden consecutivo.
[2]Cantidad de personas que obtuvieron un puesto de trabajo dentro de una empresa o que se encuentran trabajando por su cuenta.
[3]Mínimo un mes, máximo tres meses.
[4]Matutino, Vespertino o Mixto.
[5]Pago de las becas, instructores y material dependiendo de la modalidad de cada curso.
[6]Es el tipo de curso pudiendo ser: Mixta, Mixta en las Micro y Pequeñas Empresas, Escolarizada, Autoempleo, Proyecto de modernización de la educación técnica y la capacitación, e iniciativas locales de empleo.

3. la *identificación de las características relevantes de los cursos*, la descripción de su estructura y su interpretación.

4.4.1 Caracterización de los Municipios

Al realizar los pasos 1–4 de la metodología KDSM §2, obtuvimos como resultado una base de conocimiento conformada por reglas que describen la estructura de los municipios en relación al primer Índice de Contratación (IC).

A continuación, figura 4.1, podemos ver el árbol jerárquico obtenido al realizar la clasificación de la matriz de basales Y_0. Es decir la clasificación de las primeras medidas seriadas del IC.

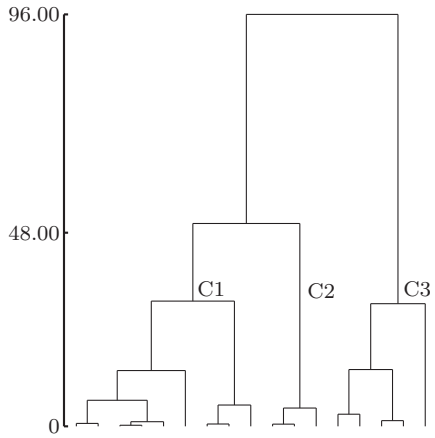


Figura 4.1: Estructura de los Municipios.

La figura 4.1 le sugirió a la experta que el corte más conveniente era en 3 clases:

Clase C1 Municipios de Meoqui, Ocampo, Camargo, Rosario, Delicias, Guadalupe y Calvo, Aquiles Serdan, Valle de Zaragoza y Parral.

Clase C2 Municipios de Madera, Santa Isabel y Ojinaga.

Clase C3 Municipios de Cuahtémoc, San Francisco del Oro, Balleza, Jimenez y Allende.

Para iniciar la interpretación de dichas clases se analizó la figura 4.2 donde se puede visualizar la caracterización del patrón de curva típico de cada clase (curva media de cada clase)

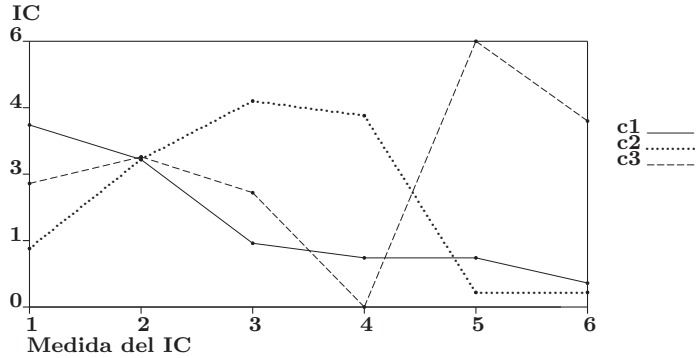


Figura 4.2: Curvas medias de las 3 clases.

además de la tendencia general de las clases y la variabilidad entre ellas.

Podemos ver que la clase C1 (de 9 municipios) presenta atributos que le distinguen especialmente porque se encuentran dentro de rangos donde los límites mínimos y máximos son los más pequeños y mayores de las 3 clases. Como ejemplo se citan algunos de los atributos que son de especial interés para la experta: extensión territorial oscila desde $335km^2$ a $16066km^2$ (ver figura 4.3), población económicamente activa (PEA) varia entre 702 hasta 44416 personas, PEA desocupada va desde 5 a 473 personas y las unidades de empresas manufactureras de 0 a 560 empresas. Además se puede ver en la figura 4.2 (línea continua) que la contratación más elevada se realiza en la primera y segunda mediciones.

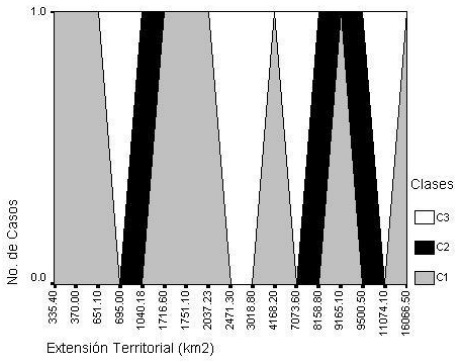


Figura 4.3: Extensión territorial por clase.

En la clase C2 (3 municipios) se presentan valores que en general se sitúan más cerca del punto medio de los rangos citados anteriormente. Donde la extensión territorial varia desde $1040km^2$ a $9500km^2$ (figura 4.3), población económicamente activa (PEA) varia entre 1375 hasta 9935 personas, PEA desocupada va desde 18 a 235 personas y las unidades de empresas

manufactureras de 9 a 110 empresas. Además se puede ver en la figura 4.2 (línea de puntos) que la contratación se realiza de forma más equilibrada distribuyendo su grosor entre la segunda y la cuarta mediciones.

Finalmente en la clase C3 (5 municipios) se tienen atributos con valores dispersos en rangos menores a la clase C1. La extensión territorial varia desde $695km^2$ a $11074km^2$ (figura 4.3), población económicamente activa (PEA) va desde 1909 hasta 42671 personas, PEA desocupada de 18 a 532 personas y las unidades de empresas manufactureras de 16 a 452 empresas. Además se puede ver en la figura 4.2 (línea discontinua) que el comportamiento muestra que en la primera y segunda mediciones se colocan en promedio de 2 a 3 personas pero se aprecia que en la quinta medición hay un incremento considerable.

4.4.2 Proceso de Caracterización (usando CIADEC)

En lo habitual cuando los individuos (en nuestro caso municipios) de un dominio complejo y real son descritos por atributos cuantitativos, no es común encontrar aquellas que caractericen a las distintas clases de una partición dada (o de referencia); así para realizar su análisis se debe relajar este concepto al de atributos parcialmente caracterizadoras [VG01a], esto es, atributos que son compartidos por otras clases y que en un sistema de reglas, se representan con un grado de pertenencia difuso a cada una de las clases de la partición de referencia, es decir, reglas difusas (en el sentido de certeza a una clase). Esto plantea un problema serio en la determinación de los atributos caracterizadores y en consecuencia en la caracterización y calidad de las clases resultantes de una clasificación de referencia, considerando la calidad de una clasificación, desde un punto de vista subjetivo como la utilidad o significado que las clases resultantes puedan tener para el experto, ya que no existe un criterio objetivo que determine esta calidad.

Como una aproximación al proceso de caracterización y en consecuencia a la obtención de una clasificación "útil" o de "calidad"para los propósitos del estudio se propone realizar los siguientes pasos:

1. una estadística descriptiva que nos proporcione información preliminar sobre la variabilidad de las mediciones, descripción gráfica de los boxplot para identificar los atributos caracterizadores y algunos otros parámetros útiles si los hubiera,

2. la inclusión del conocimiento a priori del experto para obtener las restricciones semánticas (reglas) de la clases resultantes de la partición que faciliten la significación de las clases,

3. la aplicación de CIADEC para la obtención del sistema de reglas que proporcionen las características relevantes de éstas para

4. determinar la calidad de las clases en términos del "significado" o "utilidad"de éstas,

5. éstas clases a la vista y análisis de la experta decidir si la estructura descubierta es útil, sino repetir el proceso; considerando otra clasificación donde se puede o no incluir nuevas restricciones semánticas, nuevo conocimiento del experto o bien combinando atributos en forma de reglas difusas que permitan obtener una nueva estructura de forma que ésta tenga significado para el objetivo del estudio. Si la clasificación es útil entonces continuamos con

6a el proceso de interpretación de resultados y la estructura descubierta en los datos puede usarse como *nuevo conocimiento* para la toma de decisiones o para continuar con

6b la segunda tarea de la metodología KDSM.

En la figura 4.4 muestra este proceso de caracterización de la matriz de basales Y_0 a partir de la matriz X.

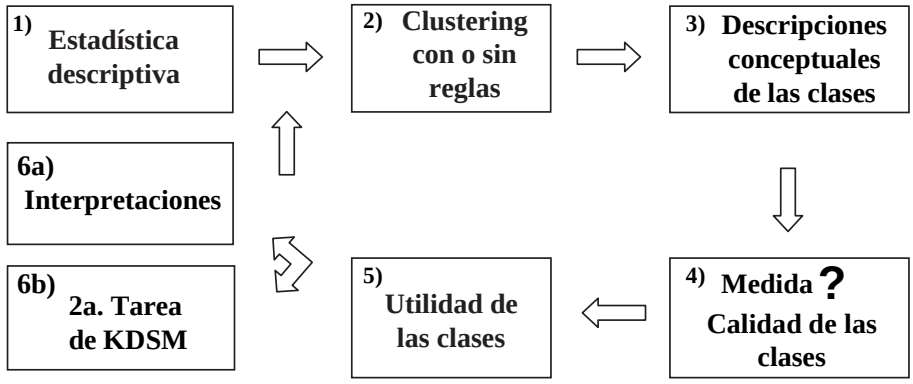


Figura 4.4: Diagrama del Proceso de Caracterización.

Después de haber realizado la estadística descriptiva sobre las medidas del IC para determinar la variabilidad de éstas se realizó la prueba de Kruskal-Wallis [SC88] y el boxplot

[VG01b, V02] sobre todos los atributos de la matriz X, para identificar aquéllos relevantes y caracterizadores, estadísticamente hablando no se obtuvo ninguna información al respecto.

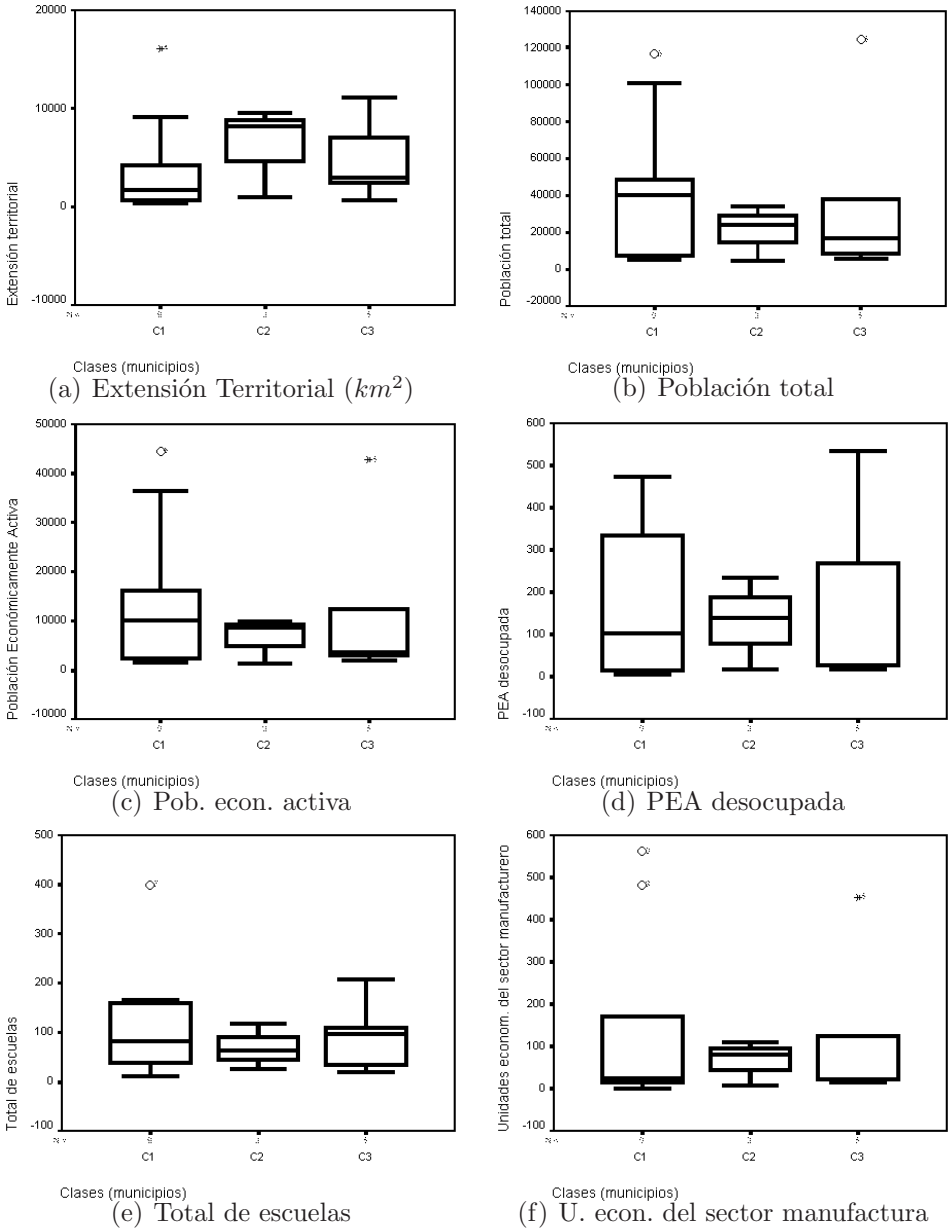
Sin embargo, la experta determinó que el atributo de extensión territorial (ET) es importante porque según la infraestructura es una posibilidad abierta para el sector empresarial de establecer nuevas empresas o de mantener solamente el mínimo ya existente. Por lo que, es un indicador para la Secretaría de Trabajo y Previsión Social de cuántos y qué tipo de cursos es conveniente impartir. Los atributos de población económicamente activa (PEA) y, principalmente, la PEA desocupada, son determinantes porque los cursos de capacitación de PROBECAT van dirigidos a este tipo de población, concretamente la que esta sin empleo. La población total es de interés porque es un indicador del costo social. El total de escuelas porque refleja la infraestructura para el soporte a los cursos y finalmente, el atributo de las unidades económicas del sector manufacturero es de suma importancia pues, en este programa piloto, tiene relación directa con el tipo de curso que se esta estudiando.

En resumen los atributos de ET, población total, PEA y PEA desocupada, total de escuelas y unidades económicas del sector manufactura son importantes, de acuerdo a la experiencia de la experta, debido a que la impartición de cursos y su contenido están sumamente relacionados con la población sin empleo y el territorio que éste ocupa. En la figura 4.5 se pueden apreciar los boxplots de los atributos de interés para lograr el objetivo de la experta.

A partir de estos atributos se aplicó CIADEC para obtener el sistema de reglas que permitiera la caracterización de cada una de las clases. Sin embargo, no se obtuvieron atributos caracterizadores y por lo tanto la clasificación obtenida no fue útil para los propósitos del estudio. Como una aproximación al problema de encontrar una clasificación "útil"se decidió combinar los atributos más significativos para la experta: ET (extensión territorial) y PEAD (Pob. Econ. Act. Desoc.) e introducir conocimiento, *a priori* de la experta, a través de reglas difusas que se utilizan como "sesgo"en la siguiente tarea de la metodología KDSM. Para la "fuzificación"de estos atributos se tomo la experiencia de la experta definiendo los rangos de valores de dichos atributos en tres etiquetas lingüísticas: pequeño (p), mediano (m) y grande (g), así se tiene:

Para el atributo ET:

- hasta 1,000 m^2 la etiqueta pequeña (p)



(a) Extensión Territorial (km^2)
(b) Población total
(c) Pob. econ. activa
(d) PEA desocupada
(e) Total de escuelas
(f) U. econ. del sector manufactura

Figura 4.5: Boxplots de los atributos de interés para la experta.

- entre 1,000 y 10,000 m^2 la etiqueta mediana (m)
- más de 10,000 m^2 la etiqueta grande (g)

Para el atributo PEAD:

- hasta 30 personas la etiqueta pequeña (p)
- entre 30 y hasta 270 personas la etiqueta mediana (m)
- más de 270 personas la etiqueta grande (g)

Así, considerando A el conjunto de atributos, se tiene que $A = \{NI, \ldots, ET, \ldots, PEAD, \ldots, PMA\}$ [1] y el conjunto de reglas obtenidas son:

$$BC_0 = \begin{cases} \text{Si } (x_{iET} \leftarrow \text{p} \ \wedge \ x_{iPEAD} \leftarrow \text{p}) \longrightarrow CR \\ \text{Si } (x_{iET} \leftarrow \text{p} \ \wedge \ x_{iPEAD} \leftarrow \text{m}) \longrightarrow CU \\ \text{Si } (x_{iET} \leftarrow \text{p} \ \wedge \ x_{iPEAD} \leftarrow \text{g}) \longrightarrow CU \\ \text{Si } (x_{iET} \leftarrow \text{m} \ \wedge \ x_{iPEAD} \leftarrow \text{p}) \longrightarrow CR \\ \text{Si } (x_{iET} \leftarrow \text{m} \ \wedge \ x_{iPEAD} \leftarrow \text{m}) \longrightarrow CD \\ \text{Si } (x_{iET} \leftarrow \text{m} \ \wedge \ x_{iPEAD} \leftarrow \text{g}) \longrightarrow CU \\ \text{Si } (x_{iET} \leftarrow \text{g} \ \wedge \ x_{iPEAD} \leftarrow \text{p}) \longrightarrow CR \\ \text{Si } (x_{iET} \leftarrow \text{g} \ \wedge \ x_{iPEAD} \leftarrow \text{m}) \longrightarrow CT \\ \text{Si } (x_{iET} \leftarrow \text{g} \ \wedge \ x_{iPEAD} \leftarrow \text{g}) \longrightarrow CU \end{cases} \tag{4.1}$$

donde: x_{iET} es el valor del atributo ET (extensión territorial) para el i-ésimo municipio y x_{iPEAD} es el valor del atributo PEAD (población econ. activa desocupada) para el i-ésimo municipio.

Una vez obtenidas las reglas se ponen a consideración de la experta para que ella valore la representación que hacen de la estructura y determine cuáles de ellas le son de utilidad.

4.4.3 Análisis del efecto de cada curso

Una vez que se llevaron a cabo los pasos de la metodología KDSM correspondientes a la primera tarea y que se obtuvieron las reglas que representan el conocimiento que es de utilidad para el objetivo del análisis, se procede a realizar la segunda tarea de KDSM que comprende los pasos 5-7 correspondientes al análisis del efecto de cada curso aislado, a partir de las *diferencias*[2] de los índices de contratación IC, eliminando el factor de bloque que conforman los municipios sobre los cursos.

El total de reglas fueron nueve, las cuales fueron evaluadas en el conjunto de objetos e indujeron una nueva clasificación conformada por las clases: U, D, T y R. Donde U, D y T reflejan el conocimiento de la experta y R conforma la clase residual o de los objetos que no se

[1] Por razones de simplicidad reducimos la notación de los atributos a etiquetas lingüísticas.

[2] Diferencias debido a que eliminan el factor de bloque establecido por los municipios

contemplan en dicho conocimiento.

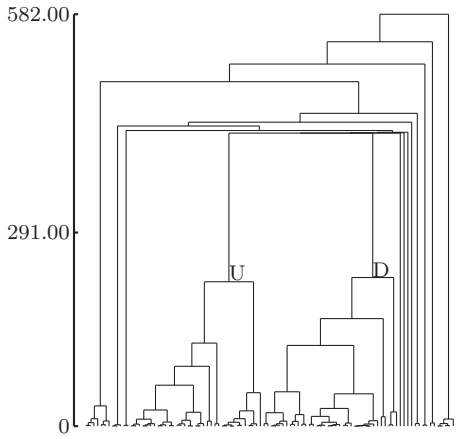


Figura 4.6: Estructura de los cursos.

En la figura 4.6, podemos ver el árbol jerárquico obtenido al realizar la ClBR de la matriz de las diferencias entre los índices de contratación. Donde la experta determinó que el corte más conveniente era en 13 clases debido a que 2 de ellas reflejan una situación de gran interés y utilidad para los objetivos de su estudio de donde logró resultados importantes y novedosos que se verán más adelante.

Clase U: COC-03-02, CCA-15-14, CCU-01-00, CMA-01-00, CD-05-04, CCA-10-09, CCA-04-03, CMA-04-03, CCA-12-11, CMA-02-01, CCA-07-06, CCA-02-01, CCU-03-02, CCA-17-16, CCA-14-13, COC-04-03, CCA-08-07, COC-02-01, CCA-05-04, CD-03-02, CD-04-03, CCU-02-01, CD-01-00, CCU-04-03, CCA-11-10, CCA-09-08, CD-06-05, CCA-03-02, CCA-06-05, CCA-01-00, COC-01-00, CCA-16-15, CCA-13-12, CMA-03-02 y CD-02-01.

5 cursos se impartieron en el municipio de Ocampo. Su especialidad fue corte y confección y su modalidad de autoempleo.

18 cursos en el municipio de Camargo, con especialidad de costura industrial y en modalidad mixta.

5 cursos en el municipio de Madera, con especialidad de corte y confección en modalidad de autoempleo.

5 cursos en Cuahtémoc de corte y confección y en modalidad de autoempleo. 7 cursos en Delicias de corte y confección y de modalidad autoempleo.

Clase D: CPA-19-18, CGC-02-01, CRO-02-01, CPA-13-12, CJI-02-01, CPA-16-15, CGC-01-00, CPA-05-04, CME-01-00, CPA-20-19, CPA-02-01, CME-04-03, CPA-06-05, CPA-01-00, CPA-17-16, CPA-10-09, CPA-04-03, CPA-11-10, CPA-08-07, CPA-14-13, CME-02-01, CRO-01-00, CPA-15-14, CPA-03-02, CPA-07-06, CJI-04-03, CPA-09-08, CPA-12-11, CJI-03-02, CPA-18-17, CME-03-02, CPA-21-20, CJI-05-04, CJI-06-05, CGC-03-02 y CJI-01-00.

22 cursos en el municipio de Parral, con especialidad en costura industrial y modalidad de capacitación mixta.

4 cursos en Guadalupe y Calvo de corte y confección y de modalidad autoempleo.

5 cursos en Meoqui de corte y confección y de modalidad autoempleo.

3 cursos en Rosales de corte y confección y de modalidad autoempleo.

7 cursos en Jimenez de costura industrial y de modalidad capacitación mixta.

Residuales: CSF-04-03, CB-04-03, CSF-01-00, COJ-01-00, COJ-04-03, CA-03-02, CSI-02-01, CVZ-03-02, CB-02-01, CA-02-01, CVZ-02-01, CA-04-03, CAS-01-00, CVZ-04-03, CAS-03-02, CVZ-01-00, COJ-03-02, CSF-02-01, CA-01-00, CAS-02-01, CSI-03-02, CB-01-00, CSF-03-02, CAS-04-03, CSI-01-00, CB-03-02 y COJ-02-01.

La experta decidió omitir las clases residuales de su análisis debido a que no representan información de su interés. Por otra parte, representan sólo el 20 % de los datos que para este caso de estudio en particular son irrelevantes.

Para iniciar la interpretación de éstas clases se caracterizó el patrón de curva típico de cada una de las 2 clases de interés para la experta (curva media de cada clase: Clase U y Clase D) figura 4.7 para visualizar la tendencia general de las clases así como la variabilidad entre ellas. Se descartaron las clases residuales debido a que no tienen ningún significado para la experta. Se puede observar como ambas clases presentan una tendencia inversa entre ellas, donde la Clase U (línea continua) se encuentra entorno a valores negativos, con excepción de la 4^{ta} medición, es decir que el grosor de la contratación se presenta en esta medida, y la Clase D (línea a puntos), entorno a valores positivos, que aunque discretos, indican que la contratación en general es estable a lo largo del tiempo de monitorización del PROBECAT.

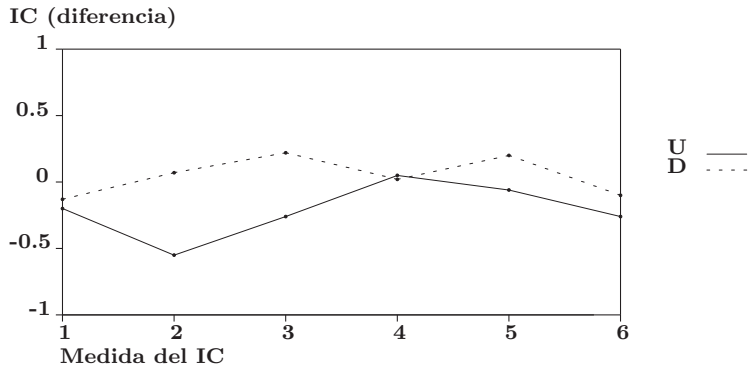


Figura 4.7: Curvas medias de las 2 clases.

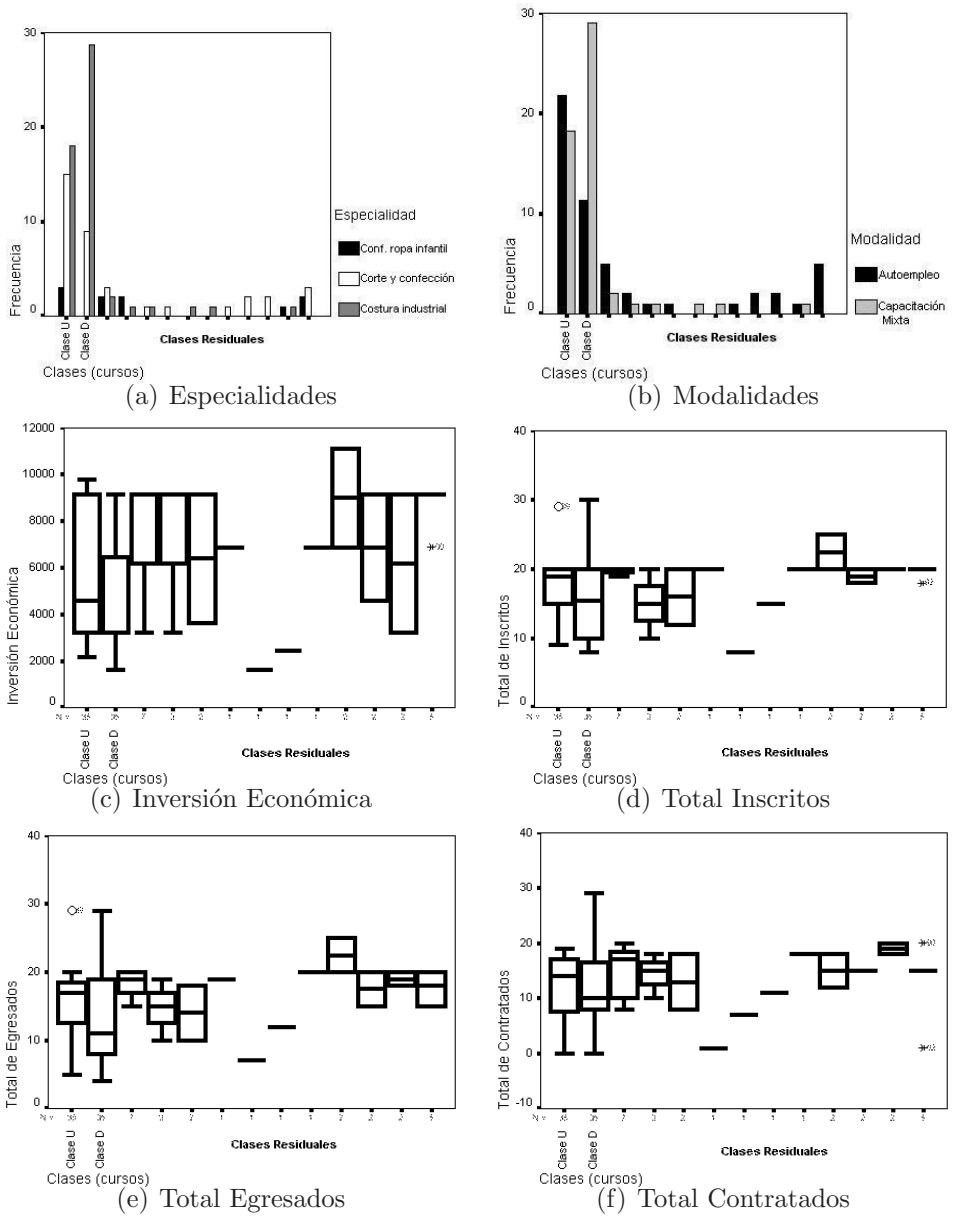
4.4.4 Identificación de las características relevantes de los cursos

Para la identificación de las características relevantes de los cursos, la descripción de su estructura y su interpretación se llevo a cabo la tercer y última tarea de la metodología KDSM correspondiente a los pasos 8 y 9.

Para ello se identificaron los atributos relevantes y caracterizadores de la proyección de las características de los cursos en las clases obtenidas de la ClBR en las diferencias de ICs, estadísticamente hablando, por medio de la prueba Kruskal-Wallis [SC88] y la metodología CIADEC [VG01a, VG01b].

Una vez identificados los atributos relevantes de la proyección de la matriz de características de los cursos en las clases obtenidas con la ClBR de las diferencias—que determinan en algún sentido el comportamiento de los municipios—la experta procedió a dar significado a los mismos.

Se encontró que la clase U se compone de 40 cursos de los cuales 22 corresponden a la modalidad de *autoempleo* capacitando aproximadamente a 440 personas, en donde la inversión económica osciló entre $4000 USD hasta $9500 USD por curso figura 4.8(c). La mayoría de los cursos iniciaron al 100 % de su capacidad (20 personas) logrando que tiempo después, al término del curso, se ocupara aproximadamente un 70 % de los participantes figura 4.8(f); ya sea trabajando por cuenta propia o bien, uniéndose en microempresas. Por otra parte, tenemos 18 cursos de modalidad *mixta* capacitando aproximadamente a 290 personas, en donde la inversión económica oscila entre $2200 USD y $6500 USD por curso figura 4.8(c). Donde se



(a) Especialidades (b) Modalidades

(c) Inversión Económica (d) Total Inscritos

(e) Total Egresados (f) Total Contratados

Figura 4.8: Boxplots de los atributos relevantes y de interés para la experta.

logró una contratación para un puesto de trabajo del 85 % de participantes aproximadamente figura 4.8(f).

La tendencia de esta clase, figura 4.7, se ve reflejada en la curva U (línea continua) donde se aprecia un repunte tardío hacia la cuarta medición de la diferencia del Índice de Contratación, que indica la evolución de dicho índice eliminando el efecto que el municipio ejerce en el curso. La tendencia de esta clase se encuentra estrechamente ligada a la modalidad de los cursos figura 4.8(b), ya que la mayoría de los mismos son de *autoempleo*. Por esta razón, los participantes

al egresar requieren de más tiempo para encontrar un puesto de trabajo o bien, establecer su propio negocio.

Se puede ver que la clase D se compone de 41 cursos de los cuales 29 cursos corresponden a la modalidad *mixta* capacitando aproximadamente a 429 personas, en donde la inversión económica osciló entre $1600 USD hasta $6800 USD por curso figura 4.8(c). Se logró una contratación para un puesto de trabajo del 90 % de participantes aproximadamente figura 4.8(f). Por último, tenemos 12 cursos de modalidad de *autoempleo* capacitando aproximadamente a 250 personas, en donde la inversión económica oscila entre $4500 USD y $9200 USD por curso figura 4.8(c), lográndose se ocupara aproximadamente el 85 % de participantes figura 4.8(f).

La tendencia de ésta clase, figura 4.7, se ve reflejada en la curva D (línea discontinua) donde se aprecia que la mayoría de las diferencias son positivas debido a que el IC esta directamente relacionado con la modalidad de los cursos figura 4.8(b), ya que la mayoría de los cursos son de *capacitación mixta* por lo que la mayoría de los participantes al egresar ya cuentan con un puesto de trabajo en las empresas que participan en el proceso de capacitación.

Para finalizar, se aprecia claramente que entre ambas clases se presenta una tendencia muy diferente marcada por la modalidad, donde cada curso tiene su efecto particular invitando a analizar aquellos atributos que influyen en el comportamiento encontrado y que esta información pueda ser la pauta para que la Secretaría de trabajo y Previsión Social (STPS) realice la planificación futura del PROBECAT.

4.5 Resultados

Para la realización de este estudio, se ha partido de tres Bases de Datos relativas al ámbito laboral. Una que contiene los atributos característicos de los municipios (matriz X), otra con las medidas seriadas del índice de contratación después de la impartición de cada curso (matriz Y) y finalmente la que contiene los atributos característicos de los cursos (matriz Z).

Se aplicó la metodología KDSM sobre estos datos, realizando un análisis en 3 fases:

1. *Caracterización de los municipios.* Se analizaron los primeros índices de contratación debido a que representan las condiciones iniciales que presentan los municipios antes de que el programa de capacitación inicie. El objetivo de este paso fue encontrar la información

útil para descubrir los perfiles de municipios que están participando del programa de capacitación (PROBECAT) y observar si existe una estructura *a priori* en el conjunto de los mismos que pueda determinar diferencias en el efecto que el PROBECAT ejerce en ellos. Se derivaron nueve reglas "difusas" (expresión 4.1) que delimitan municipios por extensión territorial y por población económicamente activa desocupada.

2. *Análisis del efecto de cada curso aislado.* La clasificación basada en reglas se aplicó utilizando la base de conocimiento obtenida en el análisis previo. Se utilizaron las diferencias de los índices de contratación para eliminar el efecto municipio y estudiar aisladamente el efecto del curso de capacitación. Este estudio nos revela directamente cómo es la evolución de la situación de desempleo en el municipio tras cada curso durante el tiempo de duración del PROBECAT en el mismo. En este caso en particular, se aprecia que el uso de la clasificación basada en reglas, mejora la calidad de los resultados y la interpretabilidad de las clases a partir de la combinación de la base de conocimiento con procesos de cluster. Por otra parte, el uso aislado de una técnica de cluster no hubiera permitido jamás incorporar reglas "difusas" como criterio de clasificación en el análisis de los índices de contratación, ya que los atributos que aparecen en estas reglas no están presentes en la matriz Y ni en la matriz D.

   Finalmente, se detectaron 2 clases: la clase U (línea), figura 4.7, donde la mayoría de los cursos son de modalidad de autoempleo y la clase D (línea a puntos) donde la mayoría de las diferencias son positivas y muestran que la mayoría de los cursos son de capacitación mixta.

3. *Identificación de las características relevantes de los cursos.* Para finalizar el análisis, se realizó la proyección de todos los atributos de la matriz Z sobre las clases obtenidas en la fase anterior. Por medio de esta proyección, se identificaron algunos atributos relevantes que muestra una tendencia claramente diferente entre ambas clases, dada principalmente por el atributo de modalidad de los cursos. La experta encontró que cada curso tiene un efecto particular y a su vez este conocimiento novedoso le permitió identificar aquéllos atributos que, a parte del ya mencionado, tienen una influencia especial en el comportamiento encontrado. Así la Secretaría de Trabajo y Previsión Social y los dife-

rentes Servicios Estatales de Empleo, podrán realizar su planeación siendo más eficientes y ahorrando—de forma importante—grandes cantidades de dinero en la inversión que la mismas realizan al PROBECAT.

En resumen, algunas conclusiones directas de la aplicación de la metodología KDSM al ámbito *laboral* son las siguientes:

KDSM como soporte en la toma de decisiones. Retroalimenta con conocimiento del PROBECAT a la STPS y a los SEE para que éstos actúen en consecuencia.

KDSM como auxiliar a la optimización y planeación. Fundamenta las decisiones en cuanto al tipo de curso y demás características de los mismos que le permitirán mejorar de forma continua y permanente el funcionamiento del PROBECAT.

KDSM como proveedor de nuevo conocimiento. Otorga conocimiento proveniente de la monitorización a los cursos que permitirá que la STPS actúe ante un desequilibrio en la relación costo/beneficio.

Capítulo 5

Conclusiones y trabajo futuro

5.1 Conclusiones

En este trabajo nos enfrentamos al estudio de medidas seriadas muy cortas repetidas de un atributo de interés, relativo a unos individuos u objetos, presentadas en puntos específicos de tiempo. Además estas medidas seriadas conforman bloques relativos a cada uno de los individuos.

A partir de una aplicación real, descrita en el capítulo §3, se diseño la metodología KDSM que permite descubrir conocimiento nuevo sobre medidas seriadas muy cortas repetidas periódicamente en un conjunto de individuos.

Así en el capítulo §2 e incluso en la sección §1.2 se explica el porqué estas medidas seriadas no pueden ser tratadas con algún método clásico de series de tiempo y cómo se trató la presencia del factor de bloque en los datos a estudiar. Esto llevó al autor(a) del presente trabajo a formular su metodología KDSM.

Por otra parte, basándose en los casos de estudio, capítulos §3 y §4 se concluye, en cuanto a la metodología se refiere, que:

1. La metodología KDSM integra herramientas y técnicas de la estadística, la inteligencia artificial y la lógica difusa en particular para ofrecer una posible solución cuando se presenta el problema de no encontrar atributos relevantes y es necesario caracterizar la matriz Y_0 en relación a los datos contenidos por la matriz X.

2. El uso de reglas difusas, constituidas por una combinación de atributos, permiten obtener una partición de utilidad para los objetivos de estudio del experto.

3. Cuando dominios como los tratados aquí son analizados por metodologías híbridas, como KDSM, una gran cantidad de información importante es recuperada, que en caso contrario la misma sería resumida enormemente que muchas características importantes nunca podrían ser encontradas.

Para finalizar este capítulo, las conclusiones en cuanto a las aplicaciones de la metodología KDSM.

De la aplicación de KDSM, descrita en el capítulo §3, a las mediciones de la terapia electroconvulsiva (TEC) se produjeron resultados muy satisfactorios, desde el punto de vista de la Psiquiatría. Se ha visto que las curvas de los tiempos de reacción (TR) de cada paciente no son inherentes al paciente, ni a la observación global de toda la terapia, sino que un mismo paciente puede reaccionar de modo diferente en cada sesión de electroshock (ES). Así, un sólo paciente tiene, por ejemplo, ESs en la clase de los jóvenes con tendencia a mejorar o la de los que empeoran, simultáneamente, claro indicativo de que su evolución a lo largo de la TEC no es monótona.

Hasta ahora, los psiquiatras habían tratado la TEC (el conjunto de todos los ES aplicados a un paciente durante todo el tratamiento) como una sola unidad, analizando globalmente el efecto de *toda* la terapia a través de la comparación de los valores de las mediciones neuropsicológicas y psicofisiológicas antes y después de la terapia [Abr97].

Este trabajo ha puesto de manifiesto que se estaba resumiendo demasiado la información, que se ocultaba un fenómeno relevante, como es el que se ha constatado: un mismo paciente reacciona diferente en cada sesión y si todo el tratamiento se hubiera resumido en una sola observación esto jamás habría sido descubierto.

Vistos estos resultados por los psiquiatras, todo apunta a que ello tiene que deberse a causas, externas o internas del propio paciente, pero que se dan o no en las diferentes sesiones y que por el momento están sin identificar, si bien existen ya algunas hipótesis sobre las que los psiquiatras han empezado a trabajar.

Éste es un conocimiento claramente nuevo en el ámbito de la psiquiatría que ha modificado la orientación de la investigación en este campo de forma inmediata.

En referencia a la aplicación de KDSM, descrita en el capítulo §4, al ámbito laboral, se

debe recordar que una de las funciones de la Secretaría de Trabajo y Previsión Social (STPS) a través de los Servicios Estatales de Empleo es dar seguimiento a los resultados obtenidos con el PROBECAT, midiendo su eficiencia mediante una relación costo/beneficio es decir, la inversión económica con respecto a la gente ya capacitada y que se ha incorporado al sector laboral.

En un esfuerzo por mejorar esta labor se ejecutó un programa piloto para valorar la conveniencia de utilizar la metodología KDSM como herramienta auxiliar para el análisis de los aproximadamente 180 diferentes tipos de cursos en los 2427 municipios en México. El programa piloto consistió del análisis de un sólo tipo de curso (manufactura textil) que se impartió en un conjunto de 17 municipios del Estado de Chihuahua y de los resultados obtenidos descritos en el capítulo §4 se obtuvieron las conclusiones que se detallan a continuación.

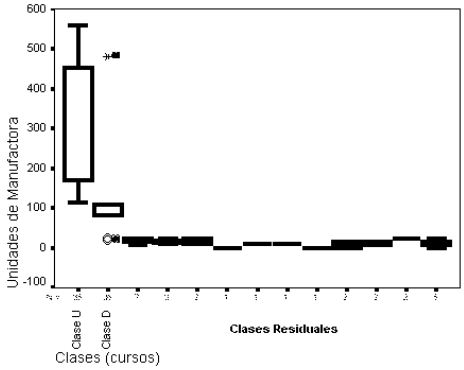


Figura 5.1: Proyección de UESM en las clases U y D.

En base al boxplot de la figura 5.1 que representa la distribución de las unidades económicas del sector manufactura (cantidad de empresas) en las clases U y D, la inversión económica (figura 4.8(c)) y la modalidad (figura 4.8(b)) se aprecia que: la inversión económica por parte del gobierno es menor cuando la modalidad del curso es *capacitación mixta* debido a la participación directa de las empresas. Es así que basándose en este hecho la STPS y los SEE pueden optimizar el PROBECAT de forma que se aproveche al máximo el nicho de oportunidad que para el gobierno, sector empresarial y trabajadores desempleados representa la coordinaciòn gobierno-empresa en la capacitación.

Por otra parte, conociendo la distribución de las empresas (figura 5.1), la modalidad del curso (figura 4.8(b)) y las características de los municipios (boxplots de la figura 4.5) la STPS y los SEE podrán realizar la planeación de la capacitación decidiendo entre las diferentes moda-

lidades de los cursos, sus contenidos, etc; de forma que respondan a las necesidades económicas de los municipios, así como las necesidades del sector empresarial y se maximice la relación costo/beneficio.

Finalmente, en base a atributos relativos a la población (boxplots de la figura 4.5(b), 4.5(c) y 4.5(d)) y los relativos al total de inscritos (figura 4.8(d)), total de egresados (figura 4.8(e)) y total de contratados (figura 4.8(f)); las dependencias podrán actuar ante una situación en la cual el mercado ya no contrate la cantidad suficiente de egresados y que se pierda el equilibrio de la relación costo/beneficio.

5.1.1 Aportación original

En lo general, KDSM es una metodología para descubrir conocimiento en dominios poco estructurados que presentan medidas seriadas muy cortas repetidas. Dicha metodología combina Inteligencia Artificial y Estadística y que ha servido para identificar conocimiento nuevo, útil y relevante en los diferentes ámbitos en que fue aplicada, de acuerdo con los requerimientos más clásicos de los métodos de Descubrimiento de Conocimiento (KDD) [FPSS96].

En cuanto a la aportación al desarrollo de la región, se han mencionado oportunamente (secciones de resultados de los capítulos §3 y §4) resultados novedosos y conclusiones relevantes (párrafos anteriores), obtenidos gracias al conocimiento nuevo que por medio de la metodología KDSM se logra encontrar. Así, es posible incidir directamente y de forma importante en la resolución de problemas, mejoramientos de procesos, etc (Ej. Mejor aprovechamiento del PROBECAT por parte del SEE del Gobierno del Estado de Chihuahua). Por esto, la metodología KDSM es parte de una línea de investigación que el(la) autor(a) asegura como muy prometedora y que con los apoyos adecuados, indudablemente, aportará desarrollo a nuestra región.

5.2 Trabajo Futuro

Este trabajo representa la primera etapa de esta línea de investigación para la cual el(la) autor(a) ha identificado ya algunas tareas como trabajo futuro.

1. Proceso continuo de mejora y actualización de la metodología KDSM.

2. Establecimiento de un método cuantitativo que permita estudiar relaciones entre las medidas seriadas (matriz Y), las características de los individuos (matriz X) y las de los eventos (matriz Z).

3. Realizar pruebas en diversos dominios o ámbitos de aplicación para posibles adaptaciones de la metodología.

4. Mejorar la representación del conocimiento, el proceso de su utilización e interacción a partir de la prueba de Kruskal-Wallis y el método CIADEC.

5. Realizar comparaciones con otros métodos de resolución de problemas similares.

6. Definir el mecanismo "definitivo" para la obtención de reglas cuando ninguno de los atributos caracterizan completamente a las clases y éstos no son relevantes, estadísticamente hablando, y

7. formalizar el proceso de obtención de la calidad de las clases en términos de "utilidad" y el proceso de caracterización multi-atributo de la matriz de basales Y_0.

Bibliografía

[Abr97] R. Abrams. *Electroconvulsive Therapy*. Ed. Oxford University Press, 1997. Third Edition. NY.US.

[AZ98] P. Adrians and D. Zantinge. *Data Mining*. Addison-Wesley, 1998. Third Edition.

[BA96] R. Brachman and T. Anand. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. *In Advances in Knowledge Discovery and Data Mining*, pages 65–78, 1996. Ed. U.Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI/MIT Press.

[Bin97] Xia B. Bin. Similarity Search in Time Series Data Sets. Master's thesis, Simon Fraser University, December 1997.

[BJR94] G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis, Forecasting and Control*. Prentice Hall, Englewood, NJ, USA, third edition, 1994.

[BO93] B. Bowerman and T. O'Connell. *Forecasting and time series: an applied approach*. Duxbury Press, Belmont, CA, USA, third edition, 1993.

[DG92] E. Diday and K.C. Gowda. Symbolic clustering using a new similaritiy measure. In *IEEE Trans. on systems, man.,and cib.*, volume 22, pages 368–378, 1992.

[Fin01] M. Fink. Convulsive therapy: a review of the first 55 years. *Journal of Affective Disorders*, (63):1–15, 2001.

[FPSS96] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases (a survey). *AI Magazine*, 3(17):37–54, 1996.

[FPSSU96] U. M. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy. Advanced in Knowledge and Data Mining. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 495–515, Cambridge, Massachusetts, 1996. AAAI/MIT Press.

[GC93] K. Gibert and U Cortés. Combining a knowledge based system with a clustering method for an inductive construction of models. In *Proc. 4th Int Work. on AI and Stats.*, pages 351–360, 1993.

[GC97] K. Gibert and U. Cortés. Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing*, 4(3):251–266, 1997.

[GC98] K. Gibert and U. Cortés. Clustering based on rules and knowledge discovery in ill-structured domains. *Computación y Sistemas.*, 1(4):213–227, abril 1998. Revista Iberoamericana de Computación. IPN, México.

[Gib94] K. Gibert. *L'us de la Informació Simbòlica en l'Automatització del Tractament Estadístic de Dominis poc Estructurats.* In the statistics and operations research phd. thesis, Universitat Politecnica de Catalunya, Barcelona, Spain, 1994.

[Gow71] J.C. Gower. A General coefficient if similarity and some of its properties. *Biometrics*, pages 857–874, 1971.

[IY94] M. Ichino and H. Yaguchi. Generalized Minkowski Metrics for Mixed feature-type data analysis. *IEEE Transaction on systems, man and cybernetics*, 22(2):146–153, 1994. April.

[KP97] A. Kehagias and V. Petridis. Predictive Modular Neural Networks for Time Series Classification. *Neural Networks*, 10:31–49, 1997.

[Lez95] M.D. Lezak. *Neuropsychological Assessment*, pages 352–353. Oxford University Press, 1995.

[Lin99] J.K. Lindsey. *Models for Repeated Measurements.* Oxford University Press, Great Britain, second edition, 1999. ISBN: 0-19-850559-0.

[Mat93] J.N.S. Matthews. A refinament to the analysis of serial data using summary measures. *Statistics in Medicine*, 12:27–37, 1993. Wiley.

[Nn90] G. Núñez. *Caracterización no monótona de la inferencia inductiva y su aplicación al aprendizaje basado en similitudes (SBL).* Phd thesis, UPC, Barcelona, 1990.

[PAVL96] K. Plataniotis, D. Androutsos, A. Venetsanopoulos, and D. Lainiotis. A new time series classification approach. *Signal Processing*, 54:191–199, 1996.

[Pn89] D. Peña. *Estadística Modelos y Métodos. Modelos lineales y series temporales*, volume II. Alianza, Madrid, segunda edition, 1989.

[PW83] S.M. Pandit and M. Wu. *Time series and system analysis, with applications.* Wiley, NY, USA, 1983.

[Ral95] H.A. Ralambondrainy. A conceptual version of K-means algorithm. *Pattern Recognition Letters*, 16:1147–1157, 1995.

[RGR01a] J. Rodas, K. Gibert, and J. Rojo. El uso de la Clasificación Basada en Reglas en la identificación de distintos efectos del electroshock. *Butlleti de l'ACIA*, (25):145–153, octubre 2001. 4rt Congres Catala de'Intel.ligencia Artificial.

[RGR01b] J. Rodas, K. Gibert, and J. Rojo. Electroshock Effects Identification Using Classification Techniques. *Springer's Lecture Notes of Computer Science Series*, Crespo, Maojo and Martin (Eds.):238–244, 2001. Second International Symposium, ISMDA 2001.

[RGR01c] J. Rodas, K. Gibert, and J. Rojo. Influential factors determination on an ill-structured domain response. Research LSI-01-6-R, Technical University of Catalonia, Barcelona, Spain, March 2001. http://www.lsi.upc.es/dept/techreps/html/R01-6.html.

[RGRC01] J. Rodas, K. Gibert, J. Rojo, and U. Cortés. A methodology of knowledge discovery in serial measurement applied to psychiatric domain. Research LSI-

01-53-R, Technical University of Catalonia, Barcelona, Spain, December 2001. http://www.lsi.upc.es/dept/techreps/html/R01-53.html.

[RJ93] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, first edition, 1993.

[RV94] J.E. Rojo and J. Vallejo. *Terapia Electroconvulsiva*. Masson-Salvat Medicina, Barcelona, first edition, 1994.

[SC88] S. Siegal and N.J. Castellan. *Nonparametric statistics for the Behavioral Sciences*, chapter Minimum mean Rank differences, pages 206–215. McGraw Hill, 2 edition, 1988. ISBN 0-07-057357-3-0.

[Sch92] G. Schuhfried. *Wiener Testsystem. Vienna Reaction Unit, Basic Program*, 1992. Development and production of scientific equipment. Mödling, Austria.

[Ste69] S. Sternberg. Memory scanning: mental process revealed by reaction time experiments. *American Scientist*, 57:421–457, 1969.

[Tuk77] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[VÓ2] F. Vázquez. Automatic Characterization and Interpretation of Conceptual Descriptions in Ill-Structured Domains using Numerical Attribute. Research LSI-02-51-R, Technical University of Catalonia, Barcelona, Spain, June 2002. http://www.lsi.upc.es/dept/techreps/html/R02-51.html.

[VG01a] F. Vázquez and K Gibert. Automatic generation of fuzzy rules in ill structures domains with numerical variables. Research LSI-01-51-R, Technical University of Catalonia, Barcelona, Spain, December 2001. http://www.lsi.upc.es/dept/techreps/html/R01-51.html.

[VG01b] F. Vázquez and K Gibert. Generación Automática de Reglas Difusas en Dominios Poco Estructurados con Variables Numéricas. In *Actas de la Conferencia de la Asociación Española para la Inteligencia Artificial*, volume 1, pages 143–152, España, nov 2001. CAEPIA 01.

[Vol85] M. Volle. Analyse des données, 1985. Ed. Economica, Paris, France.

[War] J.H. Ward. *Hierarchical grouping to optimize an objective function*, pages 236–244.

[WMS98] R. Walpole, R. Myers, and Myers S. *Probability and Statistics for Engineers and Scientists*, volume 1. Prentice Hall, sixth edition, 1998.

[Wu94] B. Wu. Pattern Recognition and Classification in Time Series Analysis. *Applied Mathematics and Computation*, 62(1):29–45, april 1994.

Apéndice A

Soporte a la aportación del presente trabajo

Este apéndice, tiene como fin aclarar que la aportación original de este trabajo es la metodología KDSM, que es un producto teórico innovador. Sin embargo, resulta obvio que para llevar la teoría a la práctica (capítulos §3 y §4) fue necesario desarrollar un programa computacional; al cual se le denominó *COLUMBUS*[1].

Es así, que *COLUMBUS* es un sistema para el descubrimiento de conocimiento en dominios poco estructurados con medidas seriadas muy cortas y repetidas con factor de bloque. Este sistema, desarrollado en lenguaje C, implementa la metodología KDSM (capítulo §2) y trabaja conjuntamente con la herramienta *KLASS*[2] [Gib94] y el método CIADEC[3] [Vó2] para la obtención de las reglas y la posterior clasificación basada en dichas reglas.

El(la) autor(a) de este trabajo considera que la verdadera aportación innovadora, el denominado trabajo creativo,... en resumen lo verdaderamente importante es: la metodología KDSM. Por este motivo no consideró la inclusión del programa computacional que implementa a KDSM al presente documento.

En otro orden de ideas es necesario resaltar, que la disciplina tecnológica en que se participa, es difícil establecer que un producto o trabajo está terminado; pues, todo lo referente a la computación y a las técnicas relativas a ésta, se encuentran en constante cambio, es decir esta disciplina es sumamente dinámica. En pocas palabras, el programa, técnica o metodología computacional que no se encuentre en constante revisión y adaptación, perece.

[1] En homenaje a un gran descubridor de nuevas tierras.

[2] Módulo de COLUMBUS que realiza la Clasificación Basada en Reglas.

[3] Método para caracterizar clases por medio de reglas.

Por tal motivo, la metodología KDSM y el programa computacional que la implementa son productos actualmente operativos (o bien, productos terminados en su primera versión), que se encuentran sujetos a un proceso permanente de revisión, ajuste y actualización.