Técnicas híbridas de Inteligencia Artificial y Estadística para el descubrimiento de conocimiento y la minería de datos

Karina Gibert

Departamento de Estadística e Investigación Operativa,
Universitat Politècnica de Catalunya, C. Pau Gargallo 5, Barcelona 08028.
email: karina@eio.upc.es

Abstract. En este artículo se presenta la línea de investigación principal del grupo de *técnicas híbridas de minería de datos* de la Universidad Politécnica de Catalunya.
Keywords: data mining, statistics, clustering, artificial intelligence, metrics, qualitative and quantitative variables, ill-structured domains ...

1 Introducción

En este artículo presentamos cómo, cuándo y por qué nuestra investigación se orienta hacia el desarrollo y aplicación de técnicas híbridas de Inteligencia Artificial y Estadística para la resolución de problemas de *Knowledge Discovery y Minería de Datos*; se presentan también algunas de las técnicas que se han desarrollado en el seno del equipo de investigación.

Antes que nada, sin embargo, conviene retornar a los orígenes y dedicar un breve espacio a la *AI& Stats* (Inteligencia Artificial y Estadística), área de investigación interdisciplinar cuyo origen podríamos situar en la fundación de la *Artificial Intelligence and Statistics Society* por Douglas Fisher en 1985, ligada al *First International Workshop on Artificial Intelligence and Statistics* que impulsó Bill Gale de los AT&T Laboratories. Desde entonces la conferencia internacional de dicha sociedad se ha venido celebrando bianualmente de forma ininterrumpida. El principal objetivo de la *AI&Stats Society* es promover la comunicación entre la comunidad estadística y la de la Inteligencia Artificial.

Hace ya 10 años, en su introducción al primer volumen de las actas de la conferencia, Cheeseman y Oldford escribían:

> *We feel that there is great potential for development at the intersection of Artificial Intelligence, Computational Science and Statistics*, [3]

lo que ha sido para nosotros referencia y motivo de reflexión desde entonces, actuando como motor de nuestra investigación.

Efectivamente existen algunas familias de problemas que han sido objeto de la Estadística y paralelamente de la Inteligencia Artificial, proponiendo cada una de estas disciplinas soluciones distintas para alcanzar un mismo objetivo. Entre éstos, uno de los más conocidos es el de la *clasificación*, que básicamente consiste

en encontrar las clases en que se estructura un dominio dado [8], y que ha sido objeto de nuestros trabajos desde el principio. En la sección 1 presentaremos algunos resultados en esta línea.

Por otro lado, en 1989 se celebra en el seno del *IJCAI* (la *International Joint Conferences on Artificial Intelligence*) el primer *Workshop on Knowledge Discovery of Data*. Siete años después, en las actas de la primera *International Conference on KDD*, Fayyad da una de las definiciones más famosas de lo que se entiende por *Knowledge Discovery and Data Mining*:

> *The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* *[10]*

y la Minería de Datos se consolida rápidamente como un área de investigación también interdisciplinar donde se hace necesario combinar técnicas avanzadas de Estadística, Inteligencia Artificial, Sistemas de Información y Visualización para afrontar la obtención de conocimiento de bases de datos de dimensiones inimaginables antes del boom Internet. Según Fayyad, el término *Knowledge Discovery of Data* se acuña en 1989 para referirse a las aplicaciones de alto nivel que incluyen métodos particulares de *Data Mining*:

> *overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms* *[9]*

Es decir, *KDD* se situaría en un plano superior, combinando los métodos de *Data Mining* con otras herramientas para extraer *conocimiento* de los datos. Y aquí es exactamente donde se sitúan nuestros objetivos.

Es claro que los objetivos de la *AI&Stats Society* encajan perfectamente en esa interdisciplinariedad enmarcada en el *KDD*. De hecho, el *KDD* es uno de los tópicos de la conferencia de dicha sociedad desde 1991.

Pues bien, nuestra investigación se sitúa en la resolución de problemas de *Knowledge Discovery y Data Mining* utilizando técnicas híbridas de *Inteligencia Artificial y Estadística* en el contexto de un tipo de dominios especialmente mal condicionados, que hemos dado en llamar *dominios poco estructurados* (ver §2).

Como ya hemos dicho, la clasificación es uno de los problemas que han sido objeto de la Estadística y la Inteligencia Artificial simultáneamente. En efecto ambas disciplinas proporcionan diferentes métodos para descubrir cuáles son las clases subyacentes a un dominio. En realidad, las técnicas de *clustering* son las más populares a la hora de separar datos en grupos y una de las técnicas de Minería de Datos más utilizadas[9]. Es más, la *clasificación* está entre las tres técnicas básicas (junto a la diferenciación de la experiencia en objetos particulares y sus atributos y la distinción entre un todo y sus partes), que dominan el pensamiento humano en su proceso de comprensión del mundo[**?**]. De hecho, nosotros coincidimos en que un buen número de aplicaciones reales en *KDD* o bien requieren un proceso de clasificación o son reducibles a él [32]. Y precisamente por eso el clustering ha sido uno nuestros objetivos principales.

Si bien es cierto que el hombre clasifica por naturaleza desde siempre, no es hasta bien entrado el siglo XX que aparece el primer tratado magistral que

aborda la clasificación desde un punto de vista formal. En la *Principles of Numerical Taxonomy* de Sokal y Sneath [37] se sientan, por primera vez, las bases algebraicas de las técnicas estadísticas de *clustering*, que parten de una matriz de datos donde todas las variables son, en principio, numéricas[1]. Probablemente por ser ésta una actividad inherente a la mente humana, desde sus principios también la Inteligencia Artificial se ha ocupado de mimetizar los procesos de clasificación. Así, Michalski [31] inicia con la *clasificación conceptual* una línea de métodos de clasificación basados en la generalización de conceptos, entendida en un sentido más o menos amplio, que parten de una matriz de datos donde todas las variables (atributos en el contexto de la Inteligencia Artificial) son categóricas (qualitativas). En cuanto al hecho de clasificar matrices de datos mixtas, ya Anderberg [1], propone tres estrategias principales: ***i)*** el *particionamiento* de variables, dividiéndolas por tipos y reduciendo el análisis al tipo dominante (si el es numérico, análisis de correspondencias seguido de un clustering sobre las componentes factoriales [39], [30], lo que produce clases en un espacio ficticio de difícil interpretación). Entre otras cosas, esta aproximación, pierde la información los grupos no dominantes. ***ii)*** realizar la *conversión* de todas las variables a un único tipo, conservando el máximo de información posible[2].***iii)*** el uso de medidas de *compatibilidad* que cubran las distintas combinaciones de tipos de variables; la idea es permitir el clustering en matrices heterogéneas sin necesidad de aplicar transformaciones previas sobre las propias variables. Pasa por la definición de distancias (o disimilaridades) entre individuos que utilicen expresiones diferentes según el tipo de la variable. En la literatura se hallan diferentes propuestas en esta dirección Gower 71 [28], Gowda & Diday 91 [27], Gibert 91 [12, 16], Ichino & Yaguchi 94 [29], Ralambondrainy 95 [34], Ruiz-Schulcloper [36]. Por razones que se presentan en [13], [18] ésta es nuestra aproximación.

La estructura de este artículo es la siguiente: tras la introducción hay una sección dedicada a los *dominios poco estructurados* §2, en la que se introduce la problemática que entrañan en cuanto al análisis de su estructura. Seguidamente se presentan dos de las metodologías híbridas que hemos desarrollado, la *clasificación basada en reglas* §1 y la *inducción de reglas basada en boxplots* §4. Finalmente se presenta un capítulo de conclusiones y trabajo futuro.

2 Dominios poco estructurados

En [13], [14] aparecen caracterizados por primera vez. Citar lo imprescindible para situar al lector. Son *dominios poco estructurados* [19] aquéllos donde: **i)** Los elementos del dominio vienen descritos por conjuntos *heterogéneos* de variables, siendo las variables numéricas y las categóricas igualmente relevantes en dicha descripción. Éstas últimas acostumbran a tener mayor número de modalidades cuanto mayor es el grado de conocimiento (*expertise*) de quien origina los datos. **ii)** Existe un *conocimiento a*

[1] Existen formas de cambiar de espacio métrico para trabajar con variables qualitativas

[2] En Estadística, tradicionalmente, las variables numéricas se convierten a grupos de variables binarias, generando la *tabla de incidencia completa*. Con ella se puede realizar un clustering en la métrica de χ^2 [7]. Las dimensiones de dicha tabla hacen la clasificación muy costosa. En Inteligencia Artificial, agrupar los valores de las variables numéricas en símbolos es lo más habitual [35]. Ello implica la pérdida relevante de información, así como la introducción de un sesgo que incide en los resultados, totalmente dirigidos por la forma como se realice la codificación [26].

priori adicional sobre la estructura del dominio. Suele ser conocimiento declarativo relativo a la estructura global del mismo (relaciones entre variables, objetivos de clasificación que se persiguen, ...). **iii)** La complejidad inherente al dominio hace que el conocimiento que de él se tiene sea *parcial* (en estos dominios existe gran cantidad de conocimiento implícito y grandes incógnitas) y *no homogéneo* (el grado de especificidad del conocimiento disponible es distinto para distintas partes del dominio).

La problemática Las especiales características de estos dominios hacen que el rendimiento de las técnicas puras de Estadística o Inteligencia Artificial, y en particular en el contexto de la clasificación, sobre diversos *dominios poco estructurados* no presenten un buen comportamiento cuando se trata de descubrir su estructura subyacente al dominio[25]. En efecto, tras algunos trabajos pudimos concluir que ambas disciplinas presentaban limitaciones serias para afrontar el análisis de este tipo de dominios y hemos podido observar cómo la construcción de metodologías híbridas que combinen técnicas provenientes de ambas disciplinas en la forma adecuada permite superar las limitaciones de unas y otras y proporcionar mejores resultados. Son éstas conclusiones a las que llegamos en el contexto de los problemas de clasificación, pero cuyo alcance los transciende seguramente.

Las técnicas estadísticas de *clustering*, suelen producir en estos dominios algunas clases con gran número de objetos cuya entidad es más que dudosa, no pudiéndose encontrar características diferenciales de dichas clases respecto a las demás que puedan justificar su constitución, lo que revierte en serias dificultades para comprender el significado de las clases obtenidas.

Además, el clustering en si mismo no tiene en cuenta el conocimiento adicional sobre el dominio, lo que parece que debería ayudar a conseguir clases fáciles de interpretar, con *entidad semántica* La existencia de conocimiento adicional acerca del dominio y el deseo de aprovecharlo para inducir su estructura nos situaría en el contexto de los sistemas basados en conocimiento. Sin embargo, aunque exista conocimiento acerca del dominio, la construcción de bases de conocimiento completas es prácticamente inalcanzable, debido a la complejidad inherente a los *dominios poco estructurados*, lo que supone la existencia de grandes cantidades de conocimiento implícito, y al hecho de que los dominios a los que nos enfrentamos tendrán partes poco conocidas, cuando no desconocidas (de lo contrario habría poco que investigar).

El cambio de paradigma El mal rendimiento de las técnicas de cluster en si mismas o de los sistemas basados en conocimiento en el contexto de los *dominios poco estructurados* no radica en el tamaño de los datos analizar propio de las aplicaciones en Minería de Datos, sino con la estructura subyacente al dominio en estudio.

Los métodos estadísticos en general, y en particular los de clustering, se asientan sobre un paradigma algebraico. Sus propiedades nacen, la mayoría de las veces, de las de los espacios vectoriales y el concepto de espacio métrico juega un papel fundamental. La idea de asimilar los objetos de análisis a *puntos de* $\Re^n$ es fundamental e inevitable. En realidad, la compleja estructura de los *dominios poco estructurados* resulta casi imposible de captar utilizando solamente criterios de tipo *métrico*, por muy sofisticada que sea la métrica. Por su parte, los métodos de la Inteligencia Artificial, descansan más bien sobre un paradigma lógico. Sus propiedades derivan algún procedimiento de razonamiento formal más o menos clásico y en ese contexto los objetos de estudio se suelen asimilar a *conceptos* en una base de conocimiento, que se van a manejar según el sistema de *razonamiento formal* en que queramos razonar. Así los elementos sobre los

que descansan las técnicas estadísticas o de Inteligencia Artificial tienen naturalezas bien distintas, se tratan de modos bien distintos, lo que produce, por ejemplo, que realizar una clasificación conceptual, al estilo de Michalski [31], o una de particiones, al estilo de Diday [6], por citar dos clásicos, sobre la misma base de datos pueda producir resultados bastante diferentes. Nuestra experiencia nos dice que todo lo que es mesurable se captura mejor bajo el paradigma algebraico (Estadística). No así con lo semántico, lo qualificable, que se captura mejor bajo el paradigma lógico (Inteligencia Artificial). Los *dominios poco estructurados*, en realidad, cruzan internamente dos tipos de estructuras, la lógica y la métrica, las cuales pueden incluso interactuar entre si, lo que hace que tratar de modelar su estructura usando uno solo de estos paradigmas sea poco menos que imposible y produce el mal rendimiento de las técnicas *puras* que venimos comentando.

A partir de esta idea, elaboramos una primera hipótesis de trabajo basada en que la interacción entre las componentes *mesurables* y *qualificables* sólo se podría modelar con técnicas que manejaran elementos de *ambas* naturalezas. Esto nos llevó en 1990 a plantear el desarrollo de la primera de las técnicas híbridas en las que hemos trabajado, la *clasificación basada en reglas*, que se introduce en la próxima sección. Y más tarde, observar que la hibridación producía realmente buenos resultados en aplicaciones reales de muy diversa índole, fue el punto de partida para abordar un segundo problema ligado al descubrimiento de conocimiento en *dominios poco estructurados*, la generación de interpretaciones, y así iniciamos el desarrollo de la *inducción de reglas basada en boxplots*, que se presenta en la sección §4. Todo ello nos sitúa en la intersección de los objetivos de los sistemas de *KDD* por un lado, en la línea de Fayyad, y por otro, de los de la *AI& Stats*, en la línea de las directrices marcadas por Oldford y Cheeseman en 1994 y hace que nuestra investigación tenga un marcado carácter interdisciplinar.

3 Clustering based on rules

La que llamamos *clasificación basada en reglas* [19] es una metodología de clasificación consistente en la hibridación de un proceso inductivo (Inteligencia Artificial) con uno de clustering (Estadística). El desarrollo original ha sido publicado en [14][15][19] y se desarrolla formalmente en [17] (detalles en [13]).

En este contexto, el conocimiento *a priori* proporcionado por el experto se formaliza en un conjunto de restricciones declarativas que la estructura final propuesta para el dominio ha de satisfacer ($\mathcal{R}$). Esas restricciones se utilizarán para inducir una primera *super-estructura* del dominio, que aunque parcial, guiará todo el proceso. La aproximación de la *clasificación basada en reglas* se basa en realizar clasificaciones *internas* a esta superestructura, respetando las restricciones del usuario, que pueden estar basadas (y de hecho es recomendable que así sea) en argumentos de naturaleza *semántica*.

Finalmente, todos los elementos del dominio serán integrados en una única estructura global. Los métodos de clasificación jerárquica son especialmente apropiados para nuestros propósitos, principalmente considerando que el conocimiento proporcionado por el experto va a ser heterogéneo (más específico en ciertas partes del dominio y más general en otras) y únicamente la organización jerárquica permite la generación de una única clasificación global que contemple este factor, poniendo de manifiesto qué parte de su conocimiento es más o menos genérica; y que los métodos jerárquicos, actualmente los más utilizados, permiten decidir el número de clases *a posteriori*, una vez construido el dendrograma y ya viendo la forma que presenta, lo que es muy recomendable en aplicaciones reales donde se busca la estructura subyacente a un dominio

(presumiblemente porque no se conoce, o no está clara) y donde es más bien raro que se tenga previa información segura sobre el número de clases.

Dependiendo del carácter de la base de conocimiento proporcionada por el experto ($\mathcal{R}$), este proceso será no supervisado (como el clustering, $\mathcal{R} = \emptyset$) o supervisado (como la clasificación, $\mathcal{R}$ será entonces una teoría completa del dominio o contendrá una clasificación de referencia). El método permite trabajar en cualquier situación intermedia, que use una bases de conocimiento *parciales*, lo que nos sitúa en el contexto de los métodos semisupervisados.

La *clasificación basada en reglas* consiste en, dado un conjunto $\mathcal{I} = \{i_1 \dots i_n\}$:

1. **Construir la base de conocimiento inicial:** El objetivo principal es permitir al experto introducir el conocimiento *a priori* del dominio de que dispone en forma de *restricciones* a la formación de clases (básicamente se trataría de materializar en esa base de conocimiento las cosas que se sabe que *son* y las que se sabe que *no pueden ser*); el experto proporciona dicho conocimiento de forma *declarativa*, lo que da lugar a un conjunto inicial de reglas lógicas, sea $\mathcal{R}^0$.

- Iniciar el proceso iterativo ($\xi = 1$):

2. **Fase de proceso del conocimiento a priori**:
   (a) **Determinar la partición de $\mathcal{I}$ inducida por las reglas:** $\mathcal{P}^{\xi}_{\mathcal{R}}$ a partir de $\mathcal{R}^{\xi}$. Incluir una *clase residual* $\mathcal{C}^{\xi}_0$ en $\mathcal{P}^{\xi}_{\mathcal{R}}$ con los objetos para los que se proporcionó conocimiento inconsistente o no se proporcionó conocimiento alguno.
   (b) **Fase de resolución de conflictos**: Analizar los objetos de $\mathcal{C}^{\xi}_0$ seleccionados por reglas contradictorias:
      i. Si es satisfactorio, ir a la fase de clasificación.
      ii. Sino, volver a la construcción de $\mathcal{R}^{\xi}$ y reformularla.
3. **Fase de clasificación**:
   (a) **Clasificación *intra* restricciones del experto**: $\mathcal{P}^{\xi}_{\mathcal{R}}$ satisfará *a priori* los requerimientos del experto. Realizar la clasificación para cada $\mathcal{C} \in \mathcal{P}^{\xi}_{\mathcal{R}}$. Notar que las clases $\mathcal{C} \subset \mathcal{I}$ lo que abarata la construcción de las clases. Determinar:
      i. Los correspondientes árboles jerárquicos (*dendrogramas*, fig.1 a)) $\tau^{\xi}_{\mathcal{C}}$,
      ii. Sus prototipos $\bar{\imath}^{\xi}_{\mathcal{C}}$, vía sumarización de la clase,
      iii. Sus masas $m^{\xi}_{\mathcal{C}} = card\ \mathcal{C}$ y
      iv. Sus índices de nivel $h^{\xi}_{\mathcal{C}}$.
4. **Fase de integración**:
   (a) **Extender la clase residual:** Añadir los prototipos $\bar{\imath}^{\xi}_{\mathcal{C}}$ a la clase residual $\mathcal{C}^{\xi}_0$, como si fueran objetos ordinarios, pero teniendo en cuenta sus respectivas masas. El nuevo conjunto de datos es la llamada *clase residual extendida* $\tilde{\mathcal{I}}^{\xi}$:

$$\tilde{\mathcal{I}}^{\xi} = \left\{(\bar{\imath}^{\xi}_{\mathcal{C}}, m^{\xi}_{\mathcal{C}}) : \mathcal{C} \in \mathcal{P}^{\xi}_{\mathcal{R}}\right\} \cup \left\{(i, 1) : i \in \mathcal{C}^{\xi}_0\right\}$$

   (b) **Realizar la integración**: Clasificar $\tilde{\mathcal{I}}^{\xi}$ para integrar todos los objetos en una sola jerarquía, recuperando la estructura jerárquica de los prototipos $\bar{\imath}^{\xi}_{\mathcal{C}}$ previamente calculadas ($\tau^{\xi}_{\mathcal{C}},\ (\mathcal{C} \in \mathcal{P}_{\mathcal{R}^{\xi}})$) y descolgándolos de su raíz a nivel $h^{\xi}_{\mathcal{C}}$ en la jerarquía global. Ello da lugar a la jerarquía τ^{ξ} (fig. 1).
   (c) **Determinar el número final de clases**: Analizar el dendrograma τ^{ξ} para elegir el mejor corte horizontal, utilizando criterios heurísticos (ya sea manuales o automáticos) [15]. Realizar el corte de τ^{ξ} identifica una partición de los datos en un conjunto de clases, $\mathcal{P}^{\xi}$. Entre los k mejores cortes (siendo k pequeño), elegir aquél que permita una mejor *interpretación*.

5. **Fase de Evaluación:** El experto debe confirmar también que la partición $\mathcal{P}^{\xi}$ obtenida con $\mathcal{R}^{\xi}$ mejora la partición $\mathcal{P}^{\xi-1}$ que se obtuvo con $\mathcal{R}^{\xi-1}$ en el modo deseado. Para ello se pueden analizar qué términos contribuyen más a las diferencias entre ellas o se pueden comparar distintas clasificaciones entre sí a través de tablas; es incluso posible probar la significación de dichas diferencias, utilizando un test no paramétrico (δ-test) diseñado para este propósito y presentado en [15]. Este paso puede producir el criterio de terminación del proceso:
   (a) Si la mejora no es significativa, parar la iteración y asumir los resultados de la iteración anterior como los mejores.
   (b) Sino, analizar los resultados para reformular la base de conocimiento. Construir $\mathcal{R}^{\xi+1}$, incrementar ($\xi = \xi + 1$) y repetir.

KLASS es el software que proporciona una implementación de esta metodología, actualmente disponible en versión LISP. A pesar de que la metodología es genérica y se podría utilizar cualquier motor de inferencia y cualquier formalismo lógico tanto para construir $\mathcal{R}^{\xi}$ como para calcular las clases inducidas por las reglas, y asimismo cualquier método jerárquico de clustering para la clasificación, **KLASS** utiliza la lógica clásica con reglas de primer orden para el manejo del conocimiento, y una reescritura del método de los *vecinos recíprocos encadenados* [5] — convenientemente adaptado para tratar con datos heterogéneos vía las medidas de compatibilidad que se detallan en [21], [22]— como método de clasificación subyacente. Una parte importante de nuestra investigación se ha centrado en las medidas de compatibilidad, lo que dio en su momento lugar a la definición de la *distancia mixta* [18].

Propiedades principales son:
- Permite tener en cuenta el conocimiento *a priori* existente sobre el dominio en estudio, incluso siendo este parcial.
- No requiere conocimiento completo sobre el dominio, admitiendo BC parciales.
- Las clases resultantes son consistentes con el conocimiento *a priori* proporcionado por el experto. Pueden generarlo o especificarlo, pero guardan consistencia con él.
- Mejora la *interpretabilidad* de las clases resultantes.

Con ello, su comportamiento frente a *dominios poco estructurados* supera las limitaciones de las técnicas clustering puras porque produce clases más *comprensibles* desde el punto de vista semántico. Lo mismo ocurre con las técnicas de aprendizaje inductivo por si mismas, puesto que reduce los efectos del conocimiento implícito, que necesariamente revierte en bases de conocimiento incompletas.

Aplicaciones Los dominios en los que esta metodología se ha aplicado y ha producido resultados satisfactorios son variados. A continuación e introducen brevemente algunos de los casos en que se ha trabajado. En todos ellos se dan las características de los *dominios poco estructurados* y, en todos, la clasificación sin utilizar la base de conocimiento *a priori* producía al menos dos clases de significado dudoso, de difícil, por no decir imposible, interpretación:
1. La clasificación de esponjas de mar [17], cuya taxonomía es motivo de controversia entre los espongiólogos. La *clasificación basada en reglas* ubicó ciertas especies objeto de discusión proporcionando argumentos objetivos para darles género.
2. La identificación de poblaciones estelares [20]. Se trabajó con datos de estrellas de la Vía Láctea tomados de la base de datos del satélite Hipparcos; los resultados mejoraban sensiblemente usando conocimiento sobre el *halo* y el *disco* de la Galaxia.

3. Disfunciones de la tiroides [25]. Se trabajó con datos de pacientes de un hospital croata y se vió cómo introducir conocimiento parcial sobre los diagnósticos clásicos producía una subdivisión más específica de utilidad clínica. Con estos mismos datos se vio cómo codificar todas las variables y repetir el análisis utilizando solamente variables qualitativas producía una sensible pérdida de información no deseable [26], lo que para nosotros supone un argumento fuerte en contra de esta práctica.
4. Plantas depuradoras de aguas residuales [24], [23]. Con datos de distintas plantas del territorio catalán de distinta estructura y función, utilizar información sobre el estado en que tiene que estar el agua que se ha de verter al río, permite identificar con claridad las situaciones más características que se operan en la planta, lo que contribuye a facilitar el control de la planta.
5. Discapacidades en ancianos[2]. Con pacientes de un hospital de Roma se ha pasado un nuevo test diseñado por la OMS para medir la discapacidad de un individuo. La *clasificación basada en reglas* ha permitido realizar una propuesta de ontología para la discapacidad, que no estaba todavía establecida, y que ha resultado obedecer a criterios funcionales, más que diagnósticos, lo que resulta muy beneficioso para la concepción geriátrica del paciente.
6. Comportamiento urbanístico de municipios [21] del área metropolitana de Barcelona a partir de las viviendas construidas de diversos tipos. Utilizar información sobre la política de protección oficial ha permitido identificar áreas de crecimiento. Con estos datos [22], de entre todas las métricas mixtas accesibles en **KLASS** la que producía clases más fácilmente interpretables era la *métrica mixta* [18].

4 Boxplot-based induction rules

Uno de los problemas abiertos ligados al clustering es el de la interpretación de resultados. De hecho, la gran crítica que hacen algunos autores a las técnicas de clustering en todas sus vertientes, es que al generar solamente una descripción extensional de las clases, queda para el analista su interpretación y caracterización [11], lo que desarrolla de forma no sistematizada, basándose en la propia experiencia, cuando a menudo, no le basta con construir automáticamente las clases, sino que necesita ayuda para entender *por qué* se detectan unas ciertas clases y no otras. Asistirle en estas tareas con herramientas automáticas es otro de los tópicos importantes para un sistema de *KDD*.

Por otro lado, uno de los mayores problemas ligado al clustering es la *validación* de resultados, que constituye un problema abierto, por no haberse encontrado aún un criterio objetivo para determinar la *calidad* de un conjunto de clases en el contexto del clustering, que se aplica en situaciones en las que *no hay* un buen conocimiento de la estructura del dominio que pueda servir de referencia (como se hace en el caso supervisado), y si lo hay, es sólo parcial. En estas situaciones, únicamente la *utilidad* de una clasificación se puede utilizar para decidir si es correcta o no [20]; evaluarla requiere un mecanismo previo de *comprensión* del significado de las clases. Esta comprensión se complica cuando aumenta el número de clases encontradas, y empeora si el número de variables a tener en cuenta es grande. Así la validación queda directamente ligada a la existencia de una interpretación clara.

Quizás por su naturaleza más semántica la generación automática de interpretaciones de una clasificación no se ha tratado formalmente desde el ámbito estadístico, aunque resolverlo es fundamental. Éste, de hecho, es uno de los problemas objeto del aprendizaje automático, del cual ID3[33] y sus sucesores son exponentes característicos. La construcción de árboles de decisión suele tener coste exponencial, aunque existen técnicas para reducir el espacio de búsqueda. Hemos explorado una posibilidad más

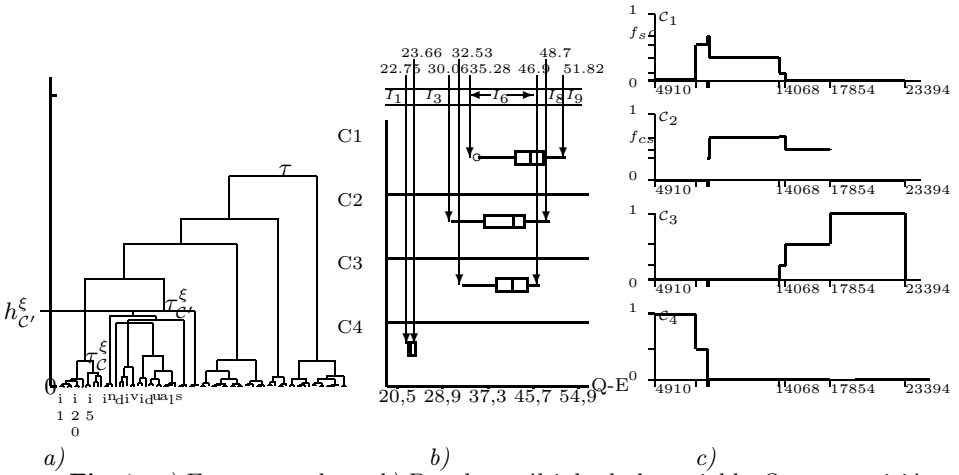

Fig. 1. *a)* Estructura de τ; *b)* Boxplot múltiple de la variable Q_E *vs* partición en 4 clases, con ventanas de longitud variable para inducir la codificación I^k de Q_E; *c)* Grados de pertenencia de Q_E a las clases.

barata, también útil con variables numéricas. Nuevamente nos hallamos ante un problema común a la Estadística y a la Inteligencia Artificial, y ligado a los sistemas de *KDD*, muy en la línea de nuestros trabajos.

Conceptos básicos Brevemente, decir que el *box-plot múltiple* (fig. 1 b)) visualiza las distribuciones de una variable numérica condicionada a un conjunto de grupos (o clases) y, en consecuencia, permite analizar la relación entre ellos. Para cada clase, se representa el intervalo de valores tomados por la variable y se marcan las observaciones extrañas con '*'. Se despliega una caja entre Q1 (primer cuartil) y Q3 (tercer cuartil) y se marca la Mediana con un signo horizontal. La caja incluye el 50% de los elementos de la clase. En nuestro caso, estos gráficos se usan para visualizar la distribución de una variable numérica en todas las clases de $\mathcal{P}$ (ver figura 1). De acuerdo con las ideas introducidas por Tukey en el campo del análisis descriptivo [38], se empezó por observar la representación gráfica de dichas distribuciones condicionales y se utilizó para extraer toda la información relevante sobre el problema. Así fue como el *box-plot múltiple* se hizo fundamental en el desarrollo del método y fue la base para identificar, primero visualmente, variables *caracterizadoras* de una clase y definir cómo calcularlas.

Sea $\Lambda^k_{\mathcal{C}}$ el conjunto de los *valores propios* de la variable X_k para la clase $\mathcal{C}$. Se define como el conjunto de valores de X_k que algún elemento de $\mathcal{C}$ toma, y que no toma ningún elemento que no esté en $\mathcal{C}$: $\exists i \in \mathcal{C} : x_{ik} \in \Lambda^k_{\mathcal{C}} \wedge \forall i \notin \mathcal{C} : x_{ik} \notin \Lambda^k_{\mathcal{C}}$.

Diremos que un valor λ de la variable X_k es *caracterizador* de la clase $\mathcal{C}$ si $\lambda \in \Lambda^k_{\mathcal{C}}$, y se puede utilizar para identificar la clase entera o parte de ella, dependiendo de si existen otros valores de X_k en $\mathcal{C}$. Así, $\{i \in \mathcal{C} : x_{ik} = \lambda\} = \mathcal{C}$ entonces, λ es *caracterizador total* de $\mathcal{C}$. Sino, $\{i \in \mathcal{C} : x_{ik} = \lambda\} \subset \mathcal{C}$ y λ es *caracterizador parcial* de $\mathcal{C}$.

El método La *inducción de reglas basada en box-plots* (*Boxplot-based Induction Rules*), se basa en una idea muy simple, que mimetiza bastante bien lo que los expertos realmente hacen cuando interpretan un boxplot múltiple. La idea central se presenta en la figura 1: Identificando los valores extremos de las distribuciones condicionales de X_k en cada clase (lo que es directo sobre el boxplot múltiple) y ordenándolos en una

lista global, se puede determinar un conjunto de intervalos en el rango de la variable numérica que se puede utilizar como codificación de la misma. Son intervalos de longitud variable tales que: **i)** Identifican *todos* los valores de las variables numéricas donde la superposición entre clases varía; donde cambia, en definitiva, la *aridad* de la intersección entre clases, **ii)** se pueden computar independientemente de la representación gráfica fácilmente y con poco coste computacional, tras una simple ordenación, sin necesidad de entrar en el estudio de intersecciones $n \times n$, de coste combinatorio. A partir de estos intervalos, se pueden calcular las *variables caracterizadoras* de una clase, ya sea totales o parciales, y en ese último caso, las probabilidades condicionales se pueden utilizar para expresar la incertidumbre ligada a la interpretación generada (a la inducción). Una descripción breve del método [14] [24] es la siguiente:

1. Para cada $C \in \mathcal{P}$ calcular el mínimo y máximo de X_k local a cada clase:
   $x_{mC} = min(X_k|\mathcal{C}) = min_{\forall i \in C}\{x_{ik}\}$ $\qquad$ $x_{MC} = max(X_k|\mathcal{C}) = max_{\forall i \in C}\{x_{ik}\}$
2. Construir un conjunto nuevo con todos los valores extremos calculados en 1:
   $$E = \{x_{mC}, x_{MC} : C \in \mathcal{P}\}$$
3. Ordenar E en E^*
4. Construir una recodificación de X_k utilizando como puntos de corte los elementos de E^*: $\quad \mathcal{I}^k = \{I_s^k\}$, donde $I_s^k = [e_s^*, e_{s+1}^*), \forall s = \{1 \dots 2\xi - 1\}$;
   Así, $\cup_{s=1}^{2\xi-1} I_s^k = [e_1^*, e_{2\xi-1}^*) = [min_{\forall i \in \mathcal{I}} x_{ik}, max_{\forall i \in \mathcal{I}} x_{ik}) = r_k$.
5. Construir la tabla de frecuencias de las clases condicionadas a los intervalos:

$\mathcal{P}\|\mathcal{I}^k$	I_1^k	I_2^k	$\dots$	$I_{2\xi-1}^k$
C_1	f_{11}	f_{12}		
C_2				
$\vdots$				
		f_{cs}		
C_ξ				$f_{\xi(2\xi-1)}$
	1	1		1

   donde $f_{cs} = \frac{card\{i : i \in C \& x_{ik} \in I_s^k\}}{card\{i : x_{ik} \in I_s^k\}}$

   y el caracterizador total es tal que $f_{cs} = 1$
6. Identificar las frecuencias condicionadas empíricas con grados de certeza. Esto se puede representar gráficamente y puede ser utilizado como una herramienta de soporte a la interpretación. La figura 1 muestra los grados de pertenencia de una variable X_k en una partición de 4 clases.
7. Para cada casilla no vacía de la tabla $\mathcal{P}|\mathcal{I}^k$ construir una regla probabilística tal como: $\quad \mathcal{R} = \{r : x_{ik} \in I_s^k \xrightarrow{f_{sc}} C : \forall f_{sc} > 0\}$
8. Finalmente, si se requieren decisiones crisp, decidir el nivel de incertidumbre α y cortar de $\mathcal{R}$ todas las reglas con un grado de incertidumbre menor que α para tener una interpretación automática de $\mathcal{P}$, *a nivel* α.

Existe un prototipo en versión β funcionand. El método ha sido probado con los datos de las plantas depuradoras, para generar la interpretación de las clases, así como con los datos de pacientes con disfunción de la tiroides. En [24] se presentan resultados utilizando solo variables cualitativas. En [4] se comparan los resultados de este método con otros métodos de aprendizaje inductivo. En [23] aparece una versión previa.

5 A modo de conclusión

Después de todo este trabajo, nuestra conclusión es que la hibridación de técnicas estadísticas con técnicas de Inteligencia Artificial produce una mejora en la calidad

de los resultados obtenidos que es mayor a la que se obtiene por aplicación separada de unas y otras seguida del contraste de resultados posterior, al menos en el contexto específico de los *dominios poco estructurados*.

De hecho, nuestra tesis es que dicha hibridación permite analizar las interacciones entre la estructura algebraica y la estructura lógica de los dominios en estudio. Esto es especialmente adecuado para analizar *dominios poco estructurados*, donde dichas interacciones juegan un papel fundamental en la estructura subyacente a estos dominios.

Hemos comprobado en algunos casos que la aplicación de dichas técnicas a dominios de estructura algebraica fuerte coincide con la aplicación de métodos estadísticos puros, mientras que si el dominio es de estructura lógica fuerte, coincide con la aplicación de métodos puros de IA, lo que es presumiblemente generalizable.

En concreto, la *clasificación basada en reglas* supera las limitaciones de los métodos puros de Estadística o Inteligencia Artificial en el análisis de *dominios poco estructurados* [19], como obtener clases sin sentido o requerir una base de conocimiento *a priori* completa [15]. Por su parte, la *inducción de reglas basada en box-plots* es más barata que los métodos inductivos clásicos y aporta al clustering una interpretación conceptual que permite *comprender* el porqué de la partición obtenida.

De hecho pensamos que ésta es una línea de desarrollo actualmente muy fértil, donde todavía queda mucho por hacer y que conduce a metodologías de gran potencia. En primer término nos proponemos insertar ambas técnicas en una metodología genérica que concatene la primera con la segunda para generar directamente conocimiento explícito acerca de la estructura del dominio a partir de la matriz de datos y la base de conocimiento *a priori*. En [23] se presenta una propuesta preliminar en esta línea, lo que nos acerca a la concepción genérica de sistema de *KDD* propuesto por Fayyad [9], que debe incluir módulos de soporte a: definición del problema (que incluya el conocimiento *a priori*, recolección de datos, depuración y preproceso, reducción de datos, selección de la técnica de data mining, data mining, interpretación y producción del conocimiento descubierto.

References

1. M. R. Anderberg. *Cluster Analysis for aplications.* Academic Press, 1973.
2. R. Annichiarico and K. Gibert *et al.* Qualitative profiles of disability. *JRRD*, 2004.
3. P. Cheeseman and R. W. Oldford (eds.). volume 89 of *LNS*. Springer, NY., 1994.
4. J. Comas, S. Dzeroski, and K. Gibert *et al.* KD by means of inductive methods in wastewater treatment plant data. *AI Communications.*, 14(1):45–62, march 2001.
5. C. De Rham. La Classif. Hierarch. selon la méthode des voisins réciproques. *Cahiers d'Analyse des Données*, V(2):135–144, 1997.
6. E. Diday. La méthode des nuées dynamiques. *StatsAp*, 2(19):19–34., 1997.
7. W.R. Dillon and M. Goldstein. *Multivariate analysis...* . Wiley., 1984. USA.
8. B. Everitt. *Cluster Analysis.* Heinemann, London, 1981.
9. U. Fayyad. *From Data Mining to Knowledge Discovery: An overview.* 1996.
10. 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.
11. D. Fisher. Machine Learning. Including a discussion on neural networks. In Florida, editor, *4th Int'l Work. on AI&Stats*, 1993.
12. K. Gibert. Klass. estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Master's thesis, UPC, 1991.
13. K. Gibert. *Ph. D. Thesis.* Eio dep., UPC, Barcelona, Spain, 1994.

14. K. Gibert. The use of symbolic information in automation of statistical treatment for ill-structured domains. *AI Communications*, 9(1):36–37, march 1996.
15. K. Gibert, T. Aluja, and U. Cortés. Knowledge Discovery with Clustering Based on Rules. In Quafafou Eds., editor, *Principles of Data Mining and Knowledge Discovery*, volume 1510 of *Lecture Notes in Artificial Intelligence*, pages 83–92, Nantes, 1998. Springer-Verlag. Interpreting Results.
16. K. Gibert and U. Cortés. KLASS: Una herramienta estadística para la creación de prototipos en ISD. In *IBERAMIA-92.*, pages 483–497, México, 1992.
17. K. Gibert and U Cortés. Combining a knowledge based system with a clustering method for an inductive construction of models. In *4th AISTATS*, 1993. Fl, USA.
18. K. Gibert and U. Cortés. Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing*, 4(3):251–266, 1997.
19. 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.
20. K. Gibert, M. Hernández, and U. Cortés. Classification based on rules: an application to Astronomy. In U. Tokio. Japón, editor, *5th. IFCS*, 1996.
21. K. Gibert and R. Nonell. Impact of mixed metrics on clustering. *Lecture Notes on Computer Science*, 2905:464–471, Nov 2003.
22. K. Gibert, R. Nonell, and *et al.* KDD with clustering:: Impact of metrics and reporting phase by using KLASS. In *COMPSTAT*, page in press, 2004.
23. K. Gibert, I. Rodríguez-Roda, and U. Cortés. Identifying characteristica situations in wastewater treatment plants with kdisd. *Applied Intelligence*, in press, 2004.
24. K. Gibert and A. Salvador. Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas... In *X ESTYLF*, pages 497–502, 2000.
25. K. Gibert and Z. Sonicki. Classification based on rules and thyroids dysfunctions. *Applied Stochastic Models in Business and Industry*, 15(4):319–324, october 1999.
26. K. Gibert, Z. Sonicki, and J. C. Martín. Impact of data encoding and thyroids dysfunctions. *Technology and Informatics*, 90:494–503, 2001.
27. K. C. Gowda and Diday E. *Symbolic clustering using a new similarity measure.* IEEE Tr. SMC, Vol 22, No 2, March/April., 1991.
28. J.C. Gower. A General coefficient of similarity ... *Biometrics*, 27:857–874, 1971.
29. M. Ichino and H. Yaguchi. Generalized Minkowski Metrics for Mixed feature-type data analysis. *IEEE Tr. on SMC*, 22(2):146–153, 1994. April.
30. L. Lebart. *Traitement statistique des données.* Ed.???, Dunod, Paris., 1990.
31. R.S. Michalski. Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data... *IJPAIS*, 4:219–243., 1980.
32. G. Nakhaeizadeh. Classification as a subtask of of Data Mining experiences form some industrial projects. In *IFCS, v-I*, pages 17–20, Kobe, JAPAN, march 1996.
33. J.R. Quinlan. Learning logical definitions from relations. *ML*, 5(3):239–266, 1990.
34. H. Ralambondrainy. *A conceptual version of the K-means algorithm.* Lifetime Learning Publications, Belmont, California, 1995.
35. M. Roux. *Algorithmes de classification.* Mason, 1985. Paris, France.
36. J. Ruiz-Shulcoper *et al.* Data analysis between sets of objects. In *8th ICSRIC*, volume III, pages 85–81, Baden Baden, august 1996.
37. R.R. Sokal and P.H.A. Sneath. Principles of numerical taxonomy. 1963. Freeman.
38. J.W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.
39. M. Volle. Analyse des données, 1985. Ed. Economica, Paris, France.