Ventajas de la estructura jerárquica del *clustering* en la interpretación automática de clasificaciones

Karina Gibert, Alejandra Pérez-Bonilla *

Departamento de Estadística e Investigación Operativa

Universitat Politécnica de Catalunya

Carrer Pau Gargallo 5.

08034-Barcelona

karina.gibert@upc.edu, alejandra.perez@upc.edu

Resumen

En este artículo se presenta la aplicación de la metodología *Caracterización Conceptual por Condicionamientos Sucesivos*, orientada a la generación automática de descripciones conceptuales de clasificaciones, que puedan dar soporte a la posterior toma de decisiones, así como su aplicación a la interpretación de las clases identificadas previamente en una planta depuradora de aguas residuales, las cuales caracterizan las distintas situaciones que se presentan en el proceso de depuración. La particularidad del método es que interpreta una partición obtenida sobre un dominio poco estructurado a partir de una clasificación jerárquica.

La metodología parte del uso de algunas herramientas estadísticas (como el *boxplot múltiple*, introducido por Tukey, y que en nuestro contexto supone una herramienta ágil y potente con variables numéricas) para conocer la estructura de los datos y extraer información útil (utilizando el concepto de *variables caracterizadoras* introducido en trabajos anteriores por Gibert) para la generación automática de un sistema de reglas, que permita posteriormente identificar las clases obtenidas.

Palabras Clave: clasificación jerárquica, boxplot múltiple, distribución concidionada, variable caracterizadora, inducción de reglas, interpretaciones conceptuales.

1. Introducción

En un proceso de clasificación automática en que se descubren las clases que componen un determinado dominio, quizás uno de los problemas más importantes y menos sistematizados que hay que enfrentar es el proceso de *interpretación* de las clases, íntimamente ligado a la *validación* de las mismas, y decisivo en la *posterior* utilidad *del conocimiento adquirido. La interpretación de las clases, tan fundamental para entender* el significado de la clasificación obtenida y, en consecuencia, la estructura del dominio, se realiza habitualmente de forma muy artesanal. Pero este proceso se complica enormemente conforme el número de clases crece. En este trabajo, se pretende abordar el problema de la generación automática de interpretaciones de una clasificación, con el objetivo de consolidar, a largo plazo, una herramienta que dé apoyo a esta tarea y contribuya a la sistematización de la misma.

Se presenta, pues, la metodología de *Caracterización Conceptual por Condicionamientos Sucesivos* [10], la cual parte de una clasificación para generar una interpretación automática de la misma que dé soporte a la construcción de *sistemas inteligentes de soporte a la toma de decisiones* en plantas depuradoras de aguas residuales. La propuesta se apoya en una serie de trabajos previos, inte-

*Esta investigación ha sido parcialmente financiada por CONICYT y por el proyecto TIN2004-01368.

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.67-76
ISBN: 84-9732-449-8

grando distintos elementos en una única herramienta metodológica que aprovecha la estructura jerárquica de la clasificación para superar algunas de las limitaciones observadas en trabajos anteriores.

Este artículo está organizado como sigue: Tras la introducción se presentan los antecedentes de esta investigación §2. La §3 presenta la metodología propuesta. En §4 se presenta el dominio de aplicación, la descripción de la base de datos concreta que se ha analizado y los resultados de aplicar la metodología propuesta a dichos datos. Finalmente en § 5 las conclusiones y el trabajo futuro.

2. Antecedentes

El presente trabajo se apoya en investigaciones previas en las que se ha analizado el *proceso de caracterización automático de clases*. Inspirado en la forma cómo los expertos realizan (manualmente) el proceso de interpretación, en [3], [4] se introducen dos conceptos fundamentales que están en la base de esta investigación:

- Una variable X_k es *totalmente caracterizadora* de la clase $C \in \mathcal{P}$, si todos los valores que toma X_k en la clase C son *propios* de C, es decir, no existen objetos de otras clases que tomen esos valores.

- Una variable X_k es *parcialmente caracterizadora* de la clase $C \in \mathcal{P}$ si tiene al menos un valor propio de la clase C, aunque puede compartir algún otro valor con otras clases.

El cálculo de valores propios de una clase descansa necesariamente en las distribuciones condicionadas de cada variable en dicha clase y sirve para identificar las *variables caracterizadoras* de la misma, según [3] *"las variables más relevantes en cada una de las clases formadas; dicho de otra forma, las que han resultado más decisivas en la construcción de éstas y, eventualmente, permiten detectar la pertenencia de un objeto a una clase determinada, excluyéndolo de las restantes"*.

Los conceptos definidos en [4] son formales y generales. No existe garantía alguna de que existan valores propios en una clase cualquiera, lo que en seguida requiere plantear propuestas de solución cuando nos encontramos con este caso. Las variables caracterizadoras se utilizaron para definir un primer procedimiento de caracterización para detectar conjuntos mínimos de variables que distingan una clase de otra utilizando únicamente variables cualitativas [7].

Las primeras aplicaciones del método [8] se realizaron sobre bases de datos médicas previamente analizadas manualmente [14] [15].

La propuesta se fundamenta en el estudio de la distribución conjunta entre pares de clases. Lo anterior da lugar a descripciones conceptuales que contienen la conjunción de dos atributos en cada clase. A partir de ello se observó enseguida que limitarse a pares de atributos impide conseguir buenas caracterizaciones de las clases, porque muchas veces, lo típico de ciertas clases es la interacción entre más de dos variables. Además, a mayor complejidad de la estructura del dominio, mayor suele ser el orden de estas interacciones. Es más, incluso nos atreveríamos a decir que, es precisamente en esa propiedad donde radica la característica de que un dominio sea complejo y se sitúe en lo que [3] denomina *dominios poco estructurados*.

Así, en [6] se describe la forma de caracterizar una clasificación utilizando los representantes de clase a partir de variables cualitativas y se presenta la primera versión sobre el uso de condicionamientos sucesivos, en ese caso utilizando una hipótesis de mundo cerrado y conjuntando negativos en los conceptos generados, como primera aportación al manejo de interacciones de orden superior a dos con variables cualitativas.

Paralelamente, [16] aborda una primera aproximación para visualizar la distribución de una variable numérica común respecto algunos grupos (clases), haciendo uso del *box-plot* múltiple[1]. En este trabajo se constató que esta

[1] El *box-plot* múltiple es una herramienta gráfica introducida en [17] y funciona del siguiente modo: para cada clase el intervalo de valores que toma la variable se visualiza y las observaciones atípicas (outliers) se marca con "*". Se despliega una caja desde Q_1 (primer cuartil) hasta Q_3 (tercer cuartil) y la me-

herramienta proporciona toda la información necesaria para identificar las variables caracterizadoras de una clase y en [12] se consolida el uso de ésta (como alternativa al método de variables cualitativas) para el caso de variables numéricas, aplicándolo por primera vez en el análisis de plantas depuradoras.

Detectado en las variables numéricas el problema que ya se había presentado en las cualitativas, que se requiere estudiar interacciones de orden superior a dos, con el agravante de que el trabajo con rangos continuos no aconseja el análisis por casos, en [13] se plantea la posibilidad de estudiar las intersecciones entre las cajas de los *box-plot*, lo que da lugar a una primera versión de un mecanismo que genera reglas probabilizadas y se sitúa en el paradigma difuso. Un refinamiento de esta propuesta difusa se encuentra en [11].

Todo este trabajo cristaliza en la formulación del *boxplot based induction rules* que constituye un método de generación de conceptos probabilizados, con un número mínimo de atributos en el antecedente, a partir de una categorización, en intervalos de longitud variable, de las variables numéricas que tiene en cuenta los puntos donde cambian las intersecciones entre clases [5]. En [18] se incide en la implementación de estas ideas. Aunque se obtienen soluciones satisfactorias desde un punto de vista aplicado, no está clara su optimalidad en términos de cobertura.

En [2] se comparan los resultados de esta primera versión con otros métodos de aprendizaje en el ámbito de plantas depuradoras y se observó que los resultados son fácilmente comprensibles por el experto, aunque el método presenta algunas deficiencias todavía.

En [19] se estudia la sensibilidad de las interpretaciones generadas a partir de clasificaciones obtenidas por diferentes procedimientos automáticos. En este trabajo se observó que en dominios de estructura compleja existen núcleos de estructura fuerte (los días de tormenta, ...) que son reconocidos desde cualquier método de clustering, generando conceptos muy estables, mientras las situaciones de estructura mas débil sí son sensibles al método.

En [10] se mejora la propuesta [18] de tal forma de la probabilidad de los conceptos generados aumenta, lo que da lugar a interpretaciones más seguras. El objetivo final de esta investigación es apoyarse en los trabajos previos mencionados para elaborar una nueva propuesta más completa, que supere todas las limitaciones observadas en las propuestas anteriores y consolide una metodología de generación automática de interpretaciones a partir de una clasificación, que dé soporte a la construcción de sistemas inteligentes de soporte a la toma de decisiones en plantas depuradoras de aguas residuales.

3. Metodología

Sea $\mathcal{I} = \{i_1, \ldots, i_n\}$ un conjunto de individuos u objetos, que está descrito por una serie de atributos cualitativos y/o cuantitativos $X_1 \ldots X_K$, cuyos valores para cada uno de los individuos $i \in \mathcal{I}$ se representan en una matriz rectangular $\mathcal{X}$ de dimensión (n, K), como se muestra en el Cuadro 1 , donde x_{ik}, $1 \leq i \leq n$

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k-1} & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k-1} & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n-11} & x_{n-12} & \ldots & x_{n-1k-1} & x_{n-1k} \\ x_{n1} & x_{n2} & \ldots & x_{nk-1} & x_{nk} \end{pmatrix}$$

Tabla 1: *Matriz de datos* $\mathcal{X}$

$1 \leq k \leq K$, es el valor que el individuo i–ésimo toma para la k–ésima variable. Llamamos $\mathcal{D}^k = \{c_1^k, c_2^k, ..., c_s^k\}$ al dominio de X_k, si ésta es categórica, es decir, al conjunto de valores posibles que puede tomar X_k; y $D^k = r_k$ al dominio de X_k, si ésta es numérica, siendo $r_k = [minX_k, maxX_k]$, el rango de la variable X_K.

Sea $\mathcal{P}_\xi = \{C_1, ..., C_\xi\}$, una partición en ξ clases de $\mathcal{I}$. Y $\mathcal{P}_2 = \{C_1, C_2\}$, una partición binaria de $\mathcal{I}$. Sea $\tau = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, ..., \mathcal{P}_n\}$, una jerarquía indexada sobre $\mathcal{I}$. Es importan-

diana se marca con un signo horizontal al centro de la caja. Las cajas incluyen, entonces, el 50 % de los elementos de la clase y los bigotes se extienden hasta el mínimo y máximo para cada clase.

te señalar que, en τ, entre $\mathcal{P}_\xi$ y $\mathcal{P}_\xi + 1$ siempre hay una clase y sola una, que se sudivide exactamente en 2 clases.

La *Caracterización Conceptual por Condicionamientos Sucesivos* representa una propuesta metodológica para generar interpretaciones automáticas de una partición $\mathcal{P}_\xi$ perteneciente a una jerarquía indexada τ, habitualmente τ es el resultado de una clasificacion jerárquica, y se puede representar en forma de dendograma como el de la Figura 1.(ver, [9] y [16])

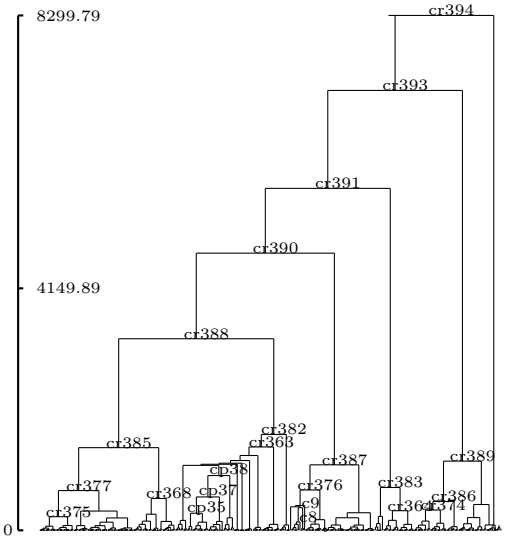

Figura 1: CAJ. Árbol general de classificación con reglas. $[\tau^{En,G}_{Gi2,R1}]$

La metodología que se propone se compone de los pasos que se describen a continuación:

1. Comenzar cortando el árbol por el nivel más alto, dejando a parte la raíz del árbol, de esta manera obtenemos $\xi = 2$ clases. La partición formada por $\mathcal{P}_2 = \{C_1, C_2\}$.

2. Detectar las *variables totalmente caracterizadoras* de cada una de las 2 clases formadas por esta partición.

   - Si existen elegir una cualquiera.
   - Si no se hallan buscar de entre las *variables parcialmente caracterizadoras* la de mayor cobertura.

3. Determinar los conceptos A^ξ_1 y A^ξ_2, en función de la variable seleccionada, asociados a C_1 y C_2 respectivamente.

4. Bajar un nivel en el árbol a analizar.

   Considerar $\xi = \xi + 1$. Necesariamente $\mathcal{P}^{\xi+1}$ está anidada en $\mathcal{P}^\xi$, es decir que las dos nuevas clases se desprenden de una y sólo una de las dos clases generadas en la partición anterior.

   Sea $C^{\xi+1}_i$ y $C^{\xi+1}_j$ las clases de $\mathcal{P}_{\xi+1}$ que subdividen una clase de $\mathcal{P}_\xi$ y $C^{\xi+1}_q$, la que ya estaba. Puesto que en el paso anterior, durante el análisis de $\mathcal{P}_\xi$, hemos separado $C^{\xi+1}_i \cup C^{\xi+1}_j$ de $C^{\xi+1}_q$, en este punto queda solamente separar $C^{\xi+1}_i$ de $C^{\xi+1}_j$, repitiendo el paso 2.

5. Integrar el conocimiento extraído de la iteración $\xi+1$ con el de la iteración ξ, determinando los conceptos finalmente asociados a los elementos de $\mathcal{P}_{\xi+1}$. Sean $B^{\xi+1}_i$ y $B^{\xi+1}_j$, Los conceptos inducidos para $C^{\xi+1}_i$ y $C^{\xi+1}_j$, en el paso $\xi + 1$. Los *conceptos compuestos* para las clases de $\mathcal{P}_{\xi+1}$ serán:

$$A^{\xi+1}_q = A^\xi_q$$
$$A^{\xi+1}_i = A^\xi_{\bar{q}} \wedge B^{\xi+1}_i$$
$$A^{\xi+1}_j = A^\xi_{\bar{q}} \wedge B^{\xi+1}_j$$

6. Hacer $\xi = \xi+1$, volver al paso 4) y repetir hasta obtener el número de clases deseado (previamente conocido).

4. Aplicación

4.1. El dominio

En este trabajo se presenta una aplicación a la interpretación de situaciones características en estaciones depuradoras de aguas residuales, entorno en el cual la extracción automática de conocimiento tiene un gran interés como apoyo a la toma de decisiones en los procesos de control y supervisión de la planta.

Las *aguas residuales* se definen en [1] como *'toda combinación de líquidos o aguas que transportan residuos procedentes de residencias, instalaciones públicas y centros comerciales e industrias, a las cuales, de manera eventual, se pueden agregar aguas subterráneas, superficiales y pluviales.'*

Para tratar adecuadamente las aguas residuales son necesarias diferentes operaciones y procesos unitarios. Diferentes combinaciones de agentes físicos, químicos y biológicos forman el diagrama de proceso de cada estación depuradora. El proceso global siempre sigue una secuencia lógica de tratamiento, dividida en varias fases que pueden variar según la estructura y objetivos de la planta y se describe a continuación. Un diagrama típico del proceso de tratamiento de aguas residuales se puede observar en la figura 2 [18].

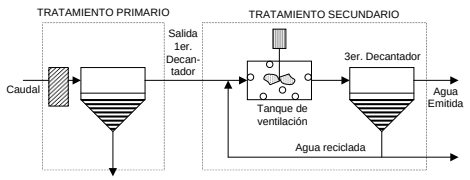

Figura 2: Diagrama típico del proceso de tratamiento de aguas residuales

El *Pretratamiento* es la primera etapa. En ella se realiza una primera separación de los sólidos arrastrados por el agua residual cuando llega al recolector. La finalidad de este pretratamiento es evitar obstrucciones posteriores, así como eliminar el efecto abrasivo de estos materiales sobre bombas o válvulas que se usan a lo largo del proceso. Esta operación física se realiza mediante una secuencia de rejas, con diferente apertura y automatismo.

El *Tratamiento Primario* es la segunda etapa donde se deja reposar el agua en un tanque de *sedimentación primaria* para que decante la materia orgánica sedimentable, así como el resto de arena o partículas inorgánicas que no han quedado retenidas por el pretratamiento.

Seguidamente el agua llega a la etapa más importante del proceso, donde se acelera el proceso biológico que se daría en la naturaleza, o sea, *la degradación*, por parte de una población multiespecífica de microorganismos, de la materia orgánica disuelta en el agua residual. Esta reacción tiene lugar en los denominados *reactores biológicos*.

La última de las etapas habituales, *sedimentación secundaria* es una nueva decantación en unos sedimentadores secundarios. El objetivo es conseguir una buena separación entre el agua tratada y la biomasa presente. Los sólidos sedimentados de ambas fases de decantación son enviados a una línea de tratamiento específico, la **línea de barros**, para reducir su volumen, peso y características.

4.2. Los datos

Los datos provienen de una Planta Depuradora situada en la costa catalana. Se analiza una muestra de 396 observaciones. Estos datos fueron obtenidos en un período de un año y un mes; desde el 1 de Setiembre de 1995 al 30 de Septiembre de 1996, correspondiendo a mediciones medias de cada día.

La descripción de la Planta para cada día consiste en caracterizar el caudal de entrada, el estado de la mezcla después del primer decantador, el caudal de salida y el estado de la mezcla en el reactor biológico. Esta caracterización se hace utilizando principalmente medidas de volumen y resultados de análisis químicos y biológicos. Las variables empleadas en la clasificación son:

- Variables de entrada:
  - Q-E: Caudal de entrada (metros cúbicos de agua por día)
  - FE-E: Pretratamiento con hierro (mg de hierro por litro de agua)
  - PH-E: pH (unidades de pH)
  - SS-E: Sólidos en suspensión (mg de sólidos por litro de agua)
  - SSV-E: Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
  - DQO-E: Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)

- DBO-E: Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)

- Variables después de la decantación:
  - PH-D: pH (unidades de pH)
  - SS-D: Sólidos en suspensión (mg de sólidos por litro de agua)
  - SSV-D: Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
  - DQO-D: Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
  - DBO-D: Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
- Variables de salida:
  - PH-S: pH (unidades de pH)
  - SS-S: Sólidos en suspensión (mg de sólidos por litro de agua)
  - SSV-S: Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
  - DQO-S: Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
  - DBO-S: Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
- Variables del tratamiento biológico:
  - V30-B: Análisis volumétrico 30; calidad de sedimentación de la mezcla (ml por litro de agua)
  - MLSS-B: Sólidos en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla)
  - MLVSS-B: Sólidos volátiles en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla)
  - MCRT-B: Edad celular (días)
  - QB-B: Caudal del reactor biológico (metros cúbicos de agua por día)
- Otras variables:
  - QR-G: Caudal de recirculación (metros cúbicos de agua por día)
  - QP-G: Caudal de la purga (metros cúbicos de agua por día)
  - QA-G: Afluencia de aire (metros cúbicos de aire por día)

4.3. Análisis

Otras muestras de datos provenientes de la misma planta fueron previamente clasificadas en un trabajo anterior [12] utilizando métodos automáticos, produciendo el árbol jerárquico de la Figura 3. La partición definitiva es el corte horizontal en 4 clases del árbol atendiendo a los criterios clásicos de homogeneidad y distinguibilidad de las clases.

En ésta sección se ilustrará cómo la metodología propuesta se utiliza para generar la interpretación de la clasificación final.

4.3.1. Partición en 2 clases

Una vez realizada la clasificación tomamos la primera partición $\mathcal{P}_2 = \{C_{393}, C_{392}\}$. Partiendo de los boxplots múltiples, se intenta buscar variables caracterizadoras, es decir, variables que tienen valores exclusivos en una clase y la pueden caracterizar. En la figura 3 podemos observar los boxplot múltiples de todas las variable para cada clase. La estadística descriptiva puede hacerse para todas las variables, aquí presentaremos una de ellas, la variable Q-E (ver Cuadro 2).

Tabla 2: Estadísticos sumarios para Q-E versus $\mathcal{P}_2$.

	Class	C_{393}	C_{392}
Var.	N = 396	$n_c = 390$	$n_c = 6$
Q-E	$\bar{\mathbf{X}}$	42,112.9453	22,563.7988
	S	4,559.2437	1,168.8481
	min	29,920	20,500
	max	54,088.6016	23,662.9004
	N*	1	0

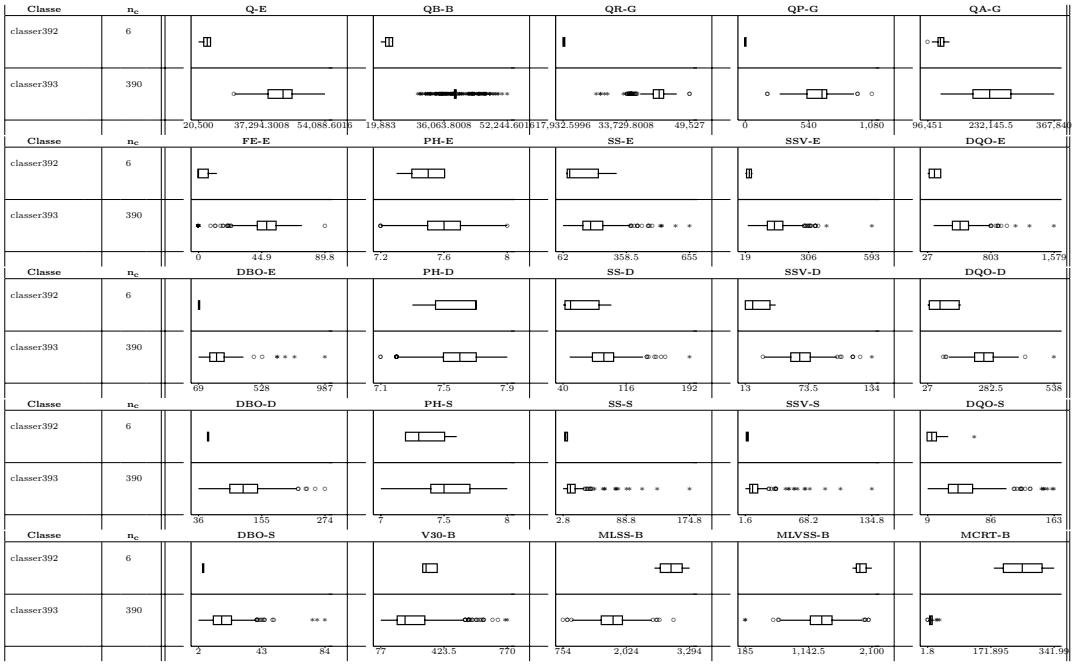

Figura 3: Análisis Descriptivo por clases para $\mathcal{P}_2$

Para hallar las variables caracterizadoras nos basamos en la propuesta previa presentada en [18], [5] que consiste en:

- Realizar la ordenación ascendente de los valores tomados por los mínimos y máximos de la variable en cada clase,
- Tomar como extremos de los intervalos los valores contiguos dos a dos. Para este caso en concreto los intervalos encontrados serían los siguientes: $I_1^{Q-E} = [20500{,}0, 23662{,}9]$, $I_2^{Q-E} = (23662{,}9, 29920{,}0]$, $I_3^{Q-E} = (29920{,}0, 54088{,}6]$.
- Cruzar la partición con el sistema de intervalos (ver Cuadro 3), lo que muestra el número de observaciones que se encuentran por clase e intervalo simultáneamente y permite calcular la probabilidad de que el valor de un intervalo determinado pertenezca a cada una de las clases.
- Generar el sistema de reglas inducido para la variable Q-E

$r_{1,2}: x_{Q-E,i} \in [20500{,}0, 23662{,}9] \xrightarrow{1,0} C_{392}$

Tabla 3: Número de observaciones classes v/s intervalos

	C_{393}	C_{392}
I_1^{Q-E}	0	6
I_2^{Q-E}	1	0
I_3^{Q-E}	388	0

$r_{2,1}: x_{Q-E,i} \in (23662{,}9, 29920{,}0] \xrightarrow{1,0} C_{392}$
$r_{3,1}: x_{Q-E,i} \in (29920{,}0, 54088{,}6] \xrightarrow{1,0} C_{393}$

Esta información puede ser representada en un diagrama de pertenencia de la variable a las distintas clases (ver figura 4.3.1).

La variable Q-E es totalmente caracterizadora, y asociamos los conceptos: "$x_{Q-E,i} \in [20500{,}0, 29920{,}0]$" a C_{393} y "$x_{Q-E,i} \in (29920{,}0, 54088{,}6]$" a C_{392}.

4.3.2. Partición en 3 clases

Consideremos ahora la *partición en 3 clases*, tenemos $\mathcal{P}_3 = \{C_{392}, C_{391}, C_{389}\}$. El primer punto es identificar la correspondencia entre las clases de $\mathcal{P}_3 = \{C_{392}, C_{391}, C_{389}\}$ y $\mathcal{P}_2 = \{C_{393}, C_{392}\}$. El Cuadro 4 muestra el

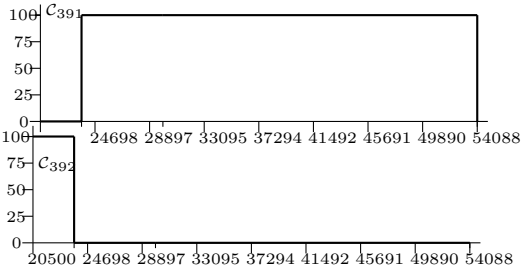

Figura 4: Diagrama de grados de pertenencia para Q-E *vs* $\mathcal{P}_2$.

Tabla 4: Intersección entre $\mathcal{P}_2$ versus $\mathcal{P}_3$

P_3 vs P_2	C_{393}	C_{392}
C_{392}	0	6
C_{391}	320	0
C_{389}	70	0

número de objetos que se encuentran en el cruce de las 2 particiones, C_{393} se divide en las clases C_{391}, C_{389} mientras que C_{392} se mantiene y el Cuadro 5 muestra la cdescriptiva para la misma variable Q-E en $\mathcal{P}_3$.

Tabla 5: Estadísticos sumarios Q-E v/s $\mathcal{P}_3$.

	Class	C_{392}	C_{391}	C_{389}
Var.	N=396	$n_c = 6$	$n_c = 320$	$n_c = 70$
Q-E	$\bar{\mathbf{X}}$	22,563.7	42,234.5	41,558.9
	S	1,168.8	4,070.2	6,336.9
	min	20,500	29,920	30,592.1
	max	23,662.9	52,255.8	54,088.6
	N*	0	1	0

De la iteración anterior se sabe que hay una característica que distingue C_{393} de C_{392} (el valor de Q-E es mayor en C_{393}). Y en la presente iteración sabemos que esta misma característica es común a C_{391} y C_{389}, lo que distingue a ambas de C_{392}, lo que arrastramos de la iteración anterior. Sólo es necesario, entonces, encontrar la característica que haga la distinción entre C_{391} y C_{389} en esta fase. Así, por un procedimiento similar al usado en el paso anterior, separamos C_{391} y C_{389} generando los sistemas de reglas inducidos por las distintas variables para C_{391} y C_{389}.

De hallar una variable totalmente caracterizadora la elegiríamos (o cualquiera de ellas si hubiera más de una). Pero en este caso no la hay y pasaremos a explorar las variables parcialmente caracterizadoras. En este caso, el poder caracterizador de los conceptos inducidos va a depender claramente de su cobertura en las clases. Estudiemos distintas variables:

La misma variable Q-E produce en esta iteración la siguiente información:

$r_1 :\ x_{Q-E,i} \in [29920{,}0, 30592{,}2] \xrightarrow{0{,}5} C_{389}$

$r_2 :\ x_{Q-E,i} \in (30592{,}2, 52255{,}8] \xrightarrow{0{,}83} C_{391}$

$r_3 :\ x_{Q-E,i} \in (52255{,}8, 54088{,}6] \xrightarrow{1{,}0} C_{389}$

La figura 4.3.2 muestra el gráfico de grados de pertenencia para las clases C_{389} y C_{391}.

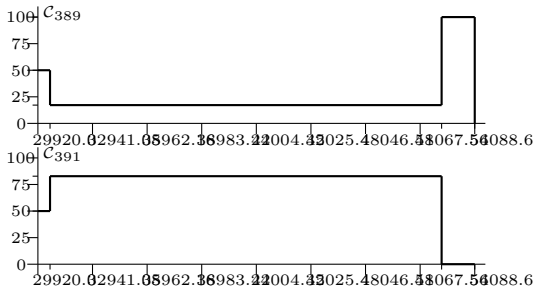

Figura 5: Diagrama de grados de pertenencia a las clases de la variable Q-E

Ahora, es posible asociar los siguientes conceptos: "$x_{Q-E,i} \in [29920{,}0, 30592{,}2]$.ª C_{389} y "$x_{Q-E,i} \in (30592{,}2, 54088{,}6]$.ª C_{391}, respectivamente. Combinando los resultados obtenidos en la iteración anterior la metodología de *Caracterización Conceptual por Condicionamientos Sucesivos*[10] genera la siguiente interpretación para $\mathcal{P}_3$:

- Clase C_{392} es tal que "Q-E está en [20500,0, 29920,0]"
- Clase C_{389} es tal que "Q-E está en [29920.0,30592.2]"
- Clase C_{391} es tal que "Q-E está en (30592.2,54088.6]"

O, en términos lingüísticos:

- Clase C_{392}, "valores bajos del Caudal de Entrada".

- Clase C_{389}, "valores medios del Caudal de Entrada".
- Clase C_{391}, "valores altos del Caudal de Entrada".

donde *bajo, medio* y *alto* se definirán de acuerdo a los límites numéricos ya indicados.

De manera similar se puede trabajar con la variable SS-E con lo que se generará el siguiente sistema de reglas reducido:

$r_1: \ x_{SS-E,i} \in [62{,}0, 82{,}0] \xrightarrow{0{,}5} C_{389}$

$r_2: \ x_{SS-E,i} \in (82{,}0, 266{,}0] \xrightarrow{0{,}75} C_{391}$

$r_3: \ x_{SS-E,i} \in (266{,}0, 655{,}0] \xrightarrow{1{,}0} C_{391}$

El gráfico de la función de pertenencia para esta variable se muestra en la figura 4.3.2.

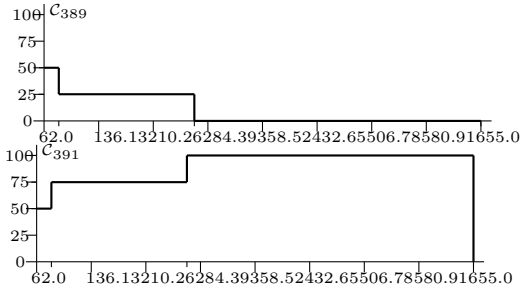

Figura 6: Diagrama de grados de pertenencia a las clases de la variable SS-E

Con respecto a esta variable, asociar el concepto "$x_{SS-E,i} \in [62{,}0, 82{,}0]$.ª C_{389} y "$x_{SS-E,i} \in (82{,}0, 655{,}0]$.ª C_{391}, desemboca en la siguiente interpretación de $\mathcal{P}_3$:

- Clase C_{392} es tal que "Q-E está en [20500,0, 29920,0]"
- Clase C_{389} es tal que "Q-E está en [29920.0,54088.6] y SS-E está en [62.0,82.0]"
- Clase C_{391} es tal que "Q-E está en [29920.0,54088.6] y SS-E está en (82.0,655.0]"

O, en términos lingüísticos:

- Clase C_{392}, " Caudal de Entrada bajo".
- Clase C_{389}, " Caudal de Entrada medio-alto y Sólidos en suspensión bajos".
- Clase C_{391}, " Caudal de Entrada medio-alto y Sólidos en suspensión altos".

O, finalmente, con respecto a QA-G, el sistema reglas que se induce es el siguiente:

$r_1: \ x_{QA-G,i} \in [124120{,}0, 136371{,}0] \xrightarrow{0{,}67} C_{389}$

$r_2: \ x_{QA-G,i} \in (136371{,}0, 324470{,}0] \xrightarrow{0{,}82} C_{391}$

$r_3: \ x_{QA-G,i} \in (324470{,}0, 367840{,}0] \xrightarrow{1{,}0} C_{391}$

Y el gráfico de la función de pertenencia para esta variable se muestra en la figura 4.3.2.

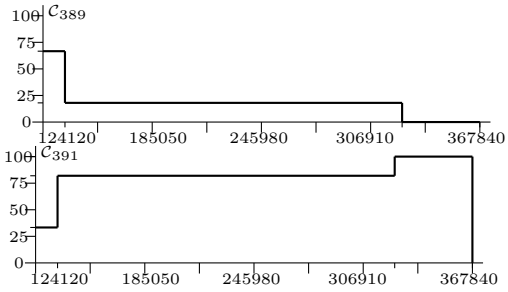

Figura 7: Diagrama de grados de pertenencia a las clases de la variable QA-G

Entonces la interpretación final es:

- Clase C_{392}, " Caudal de Entrada bajo".
- Clase C_{389}, " Caudal de Entrada medio-alto y poca Afluencia de aire".
- Clase C_{391}, " Caudal de Entrada medio-alto y Afluencia de aire medio-alta".

Dado que cualquiera de estas interpretaciones parte de variables parcialmente caracterizadoras, se trata de interpretaciones no seguras a las que se debe asociar un grado de certeza dependiendo de las probabilidades del sistema de reglas asociado a cada partición y a cada variable. La interpretación con el grado global más alto de certeza es la que se propone considerar para ser refinada en la siguiente iteración.
El proceso seguiría en este caso separando las clases de la partición siguiente, $\mathcal{P}_4$, que si observamos el árbol de la figura 3, serían las clases C_{390} y C_{383} que aparecen como subdivisión de C_{391}.

5. Conclusiones y Trabajo futuro.

En este artículo hemos presentado la primera aproximación de un método para generar interpretaciones de forma automática sobre un conjunto de clases aprovechando las ventajas que presenta la estructura jerárquica en *clustering* para la construcción de los conceptos que se asocian a cada clase interpretada.

La *Caracterización Conceptual por Condicionamientos Sucesivos* [10]es un método rápido y eficaz que genera conceptos de longitud mínima que serán de gran apoyo al proceso de decisión posterior. Se trata de una propuesta preliminar que se ha aplicado con éxito a datos reales que vienen de una planta de tratamiento de aguas residuales.

El siguiente paso es completar la definición de un criterio que permita decidir de entre las distintas variables en cada iteración cual va a formar parte de la interpretación final. Actualmente se está trabajando en la búsqueda de un criterio que no pase por la construcción explícita de todos los conceptos y la valoración posterior de su cobertura.

Referencias

[1] Abrams and Eddy. Wastewater engineering treatament, disposal,reuse. 4th Ed. revised by George Tchobanoglous, Franklin L. Burton NY.US. McGraw-Hill., 2003.

[2] J. Comas, S. Dzeroski, K. Gibert, I. Roda, and M. Sànchez-Marrè.Knowledge discovery by means of inductive methods in wastewater treatment plant data. *AI Communications*, 14(1):45–62, march 2001.

[3] K. Gibert. L'us de la Informació Simbòlica en l'Automatització del Tractament Estadístic de Dominis Poc Estructurats. Dep. EIO phd. thesis., UPC, Barcelona, Spain, 1994.

[4] K.Gibert. The use of symbolic information in automation of statistical treatment for ill-structured domains. *AI Communications*, 9(1):36–37, march 1996.

[5] K. Gibert. *Técnicas híbridas de Inteligencia Artificial y Estadística para el descubrimiento de conocimiento y la minería de datos*. Red Nacional de MiDA, 2004.

[6] K. Gibert, T. Aluja, and U. Cortés. Knowledge Discovery with Clustering Based on Rules. Interpreting Results. *LNAI*, v. 1510. Springer-Verlag. pages 83–92, Nantes, 1998.

[7] Gibert, K. and Cortés, U. Generació automàtica de regles a partir de la caracterització de classes. Boletín de la ACIA. v. 14-15. pp 193-203. 1998

[8] Gibert, K., Cortés, U., Sonickis S. Using rules as a bias mechanism in Knowledge Discovery with clustering. Workshop on causal networks (IBERAMIA98), pp47-58, Lisboa.

[9] K. Gibert and A. Pérez-Bonilla. Clasificación de Algunas Plantas Depuradoras de Aguas Residuales de Cataluña. Research report DR 2004/13, UPC, Barcelona, 2004.

[10] K. Gibert and A. Pérez-Bonilla. Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS). Research report DR 2005/en prensa, UPC, 2005.

[11] K. Gibert and A. Pérez-Bonilla. Fuzzy boxplot based induction rules. Towards automatic generation of classes-interpretation. In Procs. EUSFLAT 2005, Barcelona, in press.

[12] K. Gibert, I. Roda. Identifying characteristic situations in wastewater treatment plants. In *Workshop BESAI* (ECAI2000), v. 1, pp. 1–9.

[13] K. Gibert and A. Salvador. Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales.*X Congreso Español sobre tecnologías y lógica fuzzy*, pp. 497–502, Sevilla, 2000.

[14] K. Gibert and Z. Sonicki. Classification Based on Rules and Medical Research. In Rocco Curto, editor, VIII International Symposium on Applied Stochastic Models and Data Analysis, pages 181–186, Italy, 1997.

[15] Gibert, K. and Sonicki, Z. Classification Based on Rules and Thyroids Dysfunctions. *Applied Stochastic Models in Business and Industry*, v. 15(4), pp 319-324, 1999.

[16] D. Rodríguez. Anàlisis de los datos de una depuradora de aguas utilizando classificación basada en reglas. PFC. FME, UPC. 1999.

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

[18] F. Vázquez and K Gibert. Generación Automática de Reglas Difusas en Dominios Poco Estructurados con Variables Numéricas. In Proc. CAEPIA 2001, v. 1, pp 143–152, Oviedo.

[19] F. Vázquez and K Gibert. Robustness of class prediction depending on reference partition in ill-structured domains. In I Workshop de MiDA (IBERAMIA2002), pp 13–22, Sevilla.