Comparación de clasificaciones en bases de datos ambientales utilizando GESCONDA

Karina Gibert,[*,**] Xavier Flores, Miquel Sánchez-Marrè

D. Estadística e Investigación Operativa — Universitat Politècnica de Catalunya

Lab. d'Enginyeria Química i Ambiental — Universitat de Girona

D. Lenguajes y Sistemas Informáticos — Universitat Politècnica de Catalunya

Resumen

Las técnicas de clasificación tienen una gran importancia en los procesos de *descubrimiento de conocimiento* porque pueden identificar nuevos grupos o clases de objetos en bases datos. Así, existen métodos de aprendizaje no supervisado, que han resultado ser muy útiles en bases de datos desconocidas, no etiquetadas y poco estructuradas como las ambientales. En este artículo se analizan y comparan diferentes algoritmos de clasificación. Éstos se aplican a una base de datos ambiental real para estudiar el impacto el método en la identificación de los estados característicos de este tipo de dominios. La comparación de los diferentes métodos se lleva a cabo mediante el sistema GESCONDA, prototipo de herramienta de *data mining*.

La base de datos utilizada en este estudio proviene de una planta de tratamiento de aguas residuales situada en Catalunya, en la que se tomaron muestras en diferentes puntos durante un período de 149 días

Key Words: Adquisición y Gestión de Conocimiento, Knowledge Discovery, Data Mining, Aprendizaje Automático, Modelización Estadística, Inducción de Reglas, Bases de datos ambientales, Estaciones Depuradoras de Aguas Residuales.

1. Introducción. El problema de la validación en clustering

Un Sistema Inteligente de Soporte a la Toma de Decisiones Ambientales (SISDA) se puede definir como un sistema de información inteligente orientado a reducir el tiempo en la toma de decisiones así como mejorar la consistencia y la calidad de las decisiones en Sistemas Ambientales. Un SISDA es una herramienta orientada a la toma de decisiones ideal que sugiere recomendaciones en dominios ambientales. El núcleo de un SISDA es el conocimiento que incluye, que le proporciona habilidades avanzadas para razonar sobre el sistema ambiental de una forma más fiable. Un problema común en su desarrollo es la obtención de este conocimiento. Los enfoques clásicos están basados en adquirir este conocimiento manualmente a partir de la interacción directa con expertos. Pero cuando se dispone de bases de datos que resumen el comportamiento histórico de un sistema ambiental, existen otras aproximaciones prometedoras: Utilizar técnicas provenientes de la estadística y del aprendizaje automático para inducir este conocimiento, ya sea automática o semiautomáticamente. El uso conjunto de estas técnicas se conoce como *Descubrimiento de Conocimiento o Data Mining* [4].Toda la información y conocimiento extraído es muy importante para la predicción, control, supervisión y minimización del impacto ambiental, ya sea sobre la naturaleza o sobre los propios seres humanos.

Es importante destacar la gran cantidad de información y patrones que conocimiento implícitos en estas grandes bases de datos que provienen de sistemas ambientales. Para analizar estas bases de datos y posteriormente extraer el conocimiento que contienen, existen diferentes familias de técnicas de data mining como son las técnicas de clustering, inducción de árboles de decisión, inducción de reglas de clasificación, inducción de árboles de regresión, inducción de reglas de asociación, redes neuronales artificiales, técnicas de vector de soporte, algoritmos genéticos, razonamiento basado en casos o en instancias, métodos de regresión lineal multivarian-

[*] C. Pau i Garcia, 5, 08028 Barcelona, karina.gibert@upc.edu

[**] Esta investigación ha sido parcialmente financiada por el proyecto 11N2004-01368.

Actas del III Taller Nacional de Minería de Datos y Aprendizaje. TAMIDA2005. pp 299-308
ISBN: 84-9732-449-8

te, métodos de análisis multivariante o métodos de series temporales. Algunas de estas técnicas, como la inducción de clases, inducción de reglas de clasificación, inducción de árboles de decisión, análisis de regresión multivariante ya están implementadas en GESCONDA [10], mientras que otras técnicas serán añadidas en un futuro próximo.

De entre esos métodos, las técnicas de clustering tienen una gran importancia en el *Descubrimiento de Conocimiento* por distintas razones: por un lado, por el gran número de aplicaciones reales en *Descubrimiento de Conocimiento o Data Mining* donde o bien se requiere un clustering o bien se puede reducir el problema a uno que pase por un clustering, resultando así que el clustering es muy a menudo necesario en este contexto. Por otro lado, porque estas técnicas permiten descubrir nuevos grupos o clases de objetos en una base de datos y pueden manejar Bases de Datos no etiquetadas. Estas técnicas se sitúan en el ámbito del aprendizaje no supervisado y han resultado muy útiles frente a bases de datos desconocidas, no etiquetadas y poco estructuradas, como es el caso de las bases de datos ambientales.

Por ser las técnicas de clustering apropiadas para *descubrir* la estructura *desconocida* en un determinado dominio (ambiental o no), acarrean intrínsecamente un problema de validación. De hecho, si la estructura del dominio es previamente conocida existen técnicas más adecuadas para analizarlo, como la inducción de árboles de decisión. Las técnicas de clustering realmente contribuyen a incrementar el conocimiento del dominio cuando afrontan fenómenos que no son totalmente conocidos, donde la estructura no está claramente establecida, donde existe controversia entre expertos o existen procesos con operaciones que no son bien conocidas. Y en estos casos, no hay una clasificación de referencia que permita comprobar la validez de las clases resultantes!!!. Es por esta razón que, tradicionalmente, los principales criterios de validación de las clases propuestas en un clustering, son la viabilidad de interpretar conceptualmente las clases emergentes, la posibilidad de ser *comprendidas* por un experto o la posibilidad de utilizarlas posteriormente en procesos de toma de decisiones. Sin embargo, existen muchos métodos de clustering y gran parte de ellos son parametrizados, a la vez que es también sabido que los resultados de un proceso de clustering dependen directamente del método utilizado, así como de los valores de los parámetros.

Partiendo del hecho que un clustering sugiere un modelo factible para la estructura del dominio que no es en absoluto única, y conociendo que cuando aparecen necesidades de realizar un clustering, en general no se dispondrá de una partición de referencia para realizar la validación posterior, el objetivo de este trabajo es mostrar cómo la comparación de las clasificaciones propuestas por diferentes algoritmos se puede utilizar para consolidar el conocimiento acerca del dominio a falta de un criterio más fuerte, y cómo GESCONDA permite hallar nuevos clusters en bases de datos no etiquetadas mediante el uso de diferentes familias de métodos de clustering implementadas en dicha herramienta: K-means (y algunas de sus variantes), Isodata, Nearest-Neighbour, Marata and CobWeb/3.

La mayoría se utilizan en el análisis comparativo presentado en este trabajo. La herramienta proporciona una forma fácil de validar la consistencia de las clases descubiertas utilizando distintos métodos de clustering. Este proceso de Knowledge Discovery se ilustra con una aplicación a datos de una estación de tratamiento de aguas residuales.

Este artículo empieza describiendo la arquitectura de GESCONDA 2. Los métodos de clustering se detallan en 3. La aplicación a la EDAR se presenta en la sección 4. En 5 se detallan los resultados del trabajo experimental, y finalmente, en 6 se presentan las conclusiones y el trabajo futuro.

2. GESCONDA

El proyecto de investigación *Development of an Intelligent Data Analysis System for Knowledge Management in Environmental Data Bases (DB)* trata de la construcción de un Intelligent Data Analysis System (IDAS) que dé soporte a la toma de decisiones ambientales, como ya se ha dicho. El proyecto es financiado por el Gobierno español.

GESCONDA es el nombre del IDAS desarrollado en el marco del proyecto, con el objetivo de realizar Knowledge Discovery (KD) y análisis de datos inteligente, especialmente orientado a bases de datos ambientales. De hecho KD es un paso previo y necesario para obtener IEDSS fiables. Aunque en la literatura existen otras herramientas de KD (WEKA, Intelligent Miner...), ninguna integra en si misma, como hace GESCONDA, métodos es-

tadísticos y de IA, la posibilidad de gestionar explícitamente el conocimiento producido en las Bases de Conocimiento (KB) (en el sentido clásico de la IA), técnicas mixtas que pueden cooperar entre ellas para descubrir y extraer el conocimiento contenido en los datos, análisis dinámico,... permitiendo la interacción entre todos los métodos.

En base a experiencias previas, GESCONDA se ha diseñado como una arquitectura multinivel de 4 capas, que conecta al usuario con el sistema o proceso ambiental. Estas 4 capas son: (see fig.1):

- Filtrage de datos:
  - Data cleaning;
  - Análisis y tratamiento de datos Missing;
  - Análisis y tratamiento de Outliers;
  - Análisis estadístico univariante;
  - Análisis estadístico bivariante;
  - Herramientas de visualización;
  - Transformación de Variables
- Recomendación y gestión de Meta-Conocimiento:
  - Definición del Objetivo del problema;
  - Recomendación del método;
  - Instanciación de Parámetros;
  - Gestión del Meta-conocimiento sobre las variables (o atributos);
  - Gestión del Meta-conocimiento relativo a los objetos o ejemplos;
  - Explicitación del conocimiento del dominio
- Knowledge Discovery:
  - Clustering (Aprendizaje automático y Estadística);
  - Inducción de árboles de decisión;
  - Inducción de reglas de clasificación;
  - Razonamiento basado en casos;
  - máquinas de vector de soporte;
  - Modelización estadística;
  - Análisis dinámico
- Gestión de conocimiento:
  - Integración de distintos patrones de conocimiento para predicción, o para planificación, o supervisión de sistemas;
  - Validación del conocimiento adquirido;
  - Utilización del conocimiento por los usuarios
  - User interaction.

GESCONDA proporciona un conjunto de técnicas mixtas que son útiles en la adquisición de conocimiento relevante sobre sistemas ambientales, a partir de las bases de datos disponibles. Este conocimiento puede ser utilizado posteriormente en la implementación de EDSS fiables. La portabilidad del software se debe al uso de una plataforma Java. En la siguiente sección se presenta una descripción más detallada del agente de clustering, que es el que está directamente involucrado en este trabajo.

Dicho agente implementa diferentes algoritmos y ofrece algunas herramientas de visualización, como histogramas por clase o representaciones bivariantes de los clusters, así como tablas con todos los elementos de un cluster. De existir una partición de referencia, puede también proporcionar el porcentaje de objetos bien clasificados, que se puede tomar como una coeficiente de calidad.

3. Métodos de Clustering

El problema del clustering está siempre relacionado con encontrar la mejor agrupación de un conjunto de objetos, atendiendo a ciertos criterios. De hecho, hay que tener en cuenta que hallar la agrupación óptima es un problema difícil de complejidad combinatoria, es decir, extremadamente caro cuando el número de objetos o variables crece. Por ello es necesario introducir heurísticos que permitan alcanzar un subóptimo en un tiempo razonable. Dependiendo del criterio y del modo en que se trate de llegar al *óptimo*, aparecen diferentes algoritmos.

En GESCONDA, se implementan algunos, como K-means, a nearest-neighbour,... y otros, como Autoclass, algunos métodos de clasificación difusa,... están previstos para futuras versiones.

Como es sabido, el punto clave de un proceso de clustering es seleccionar la distancia o medida de disimilitud a utilizar. Aunque en este trabajo el impacto de la métrica no es el principal objetivo, conviene destacar que el sistema ofrece algunas opciones, como las medidas de disimilitud de Clark, Canberra, Manhattan, Euclidean, L'Eixample dando mayor adaptabilidad al sistema. Además, los métodos se pueden utilizar con todas las variables de la base de datos, o con un subconjunto preseleccionado, segun las necesidades del usuario.

A continuación se presentan detalles de los métodos utilizados específicamente en este trabajo.

3.1. Reciprocal neighbors

Se trata de un método jerárquico. Construye un árbol binario de forma ascendente desde los objetos

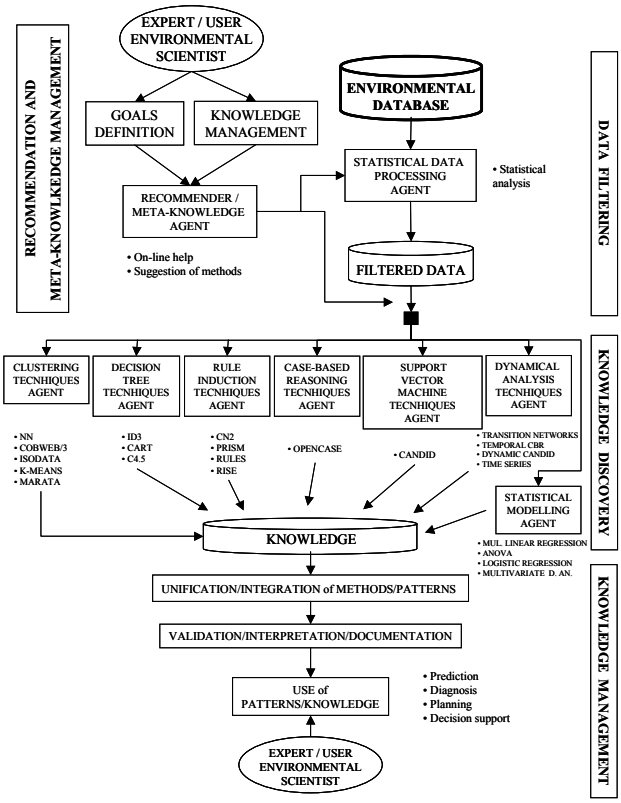

Figura 1: GESCONDA architecture.

originales hasta llegar a una clase única, fusionando a cada iteración un par de vecinos recíprocos[9].

Un par de elementos (ya sean objetos simples o subclases formadas en iteraciones previas) se definen como un par de vecinos recíprocos si uno es el más cercano al otro y viceversa, de acuerdo con cierto criterio. Es usual utilizar el criterio de Ward[11], lo que implica utilizar una transformación de la métrica euclídea relativa a la cantidad de información contenida en las clases, de modo que en cada paso se realiza la fusión más informativa.

Como es habitual en métodos jerárquicos, éste es order-independent, es decir que las clases resultantes se mantienen, sea cual sea el objeto inicial.

El número final de clases se determina, como es propio en métodos jerárquicos, a posteriori, analizando la estructura del árbol binario (*dendrograma*), y considerando algún criterio de máxima homogeneidad dentro de las clases y heterogeneidad entre ellas. Cortando el dendrograma al nivel horizontal apropiado, produce la partición final.

Aunque los métodos jerárquicos fueron inicialmente concebidos para variables numéricas, en esta investigación la distancia está parametrizada, con lo que incluso se pueden usar métricas mixtas [5], y es posible realizar el cluster sobre un conjunto heterogéneo de variables (numéricas y categóricas).

3.2. K-means

El objetivo del K-means[3] es clasificar un conjunto de datos de tal forma que los objetos de un mismo grupo sean lo más similares posible.

Se trata de un método de particiones. Así, requiere a priori la especificación del número de grupos que hay que construir (k). Se generan k semillas y se reorganizan los objetos alrededor de ellas en base a las similitudes entre los objetos y las semillas, de tal forma que al final del proceso, los objetos son lo más parecidos posible dentro de las clases.

El método es order dependent y por lo tanto los resultados pueden cambiar con distintas semillas iniciales. En GESCONDA, las semillas son generadas aleatoriamente, lo que produce order-independence, pero también resultados no reproducibles, puesto que las semillas cambian en cada ejecución. Para mitigar este efecto, se pueden aplicar

técnicas de bagging para derivar clasificadores mejores, en función de la selección del usuario. Brevemente, consisten en la ejecución de un algoritmo distintas veces, utilizando una política de votaciones para decidir la clase final de un cierto objeto.

3.3. Marata

Es una modificación de K-means especialmente desarrollada para GESCONDA. El conjunto de semillas inicial se elige de tal forma que sean lo más distintas posible en términos de la distancia o disimilitud usada. En cada iteración, solo se asignan a una clase los objetos más cercanos a su prototipo, mientras en K-means todos los objetos de asignan a su prototipo más cercano y se evalúa un criterio de terminación después de cada iteración. Marata termina cuando todos los objetos han sido asignados a alguna clase. Suele producir clases más distinguibles, pero es de ejecución más cara.

3.4. Nearest Neighbor

GESCONDA también incluye un clasificador nearest-neighbor classifier [2]. El algoritmo funciona de forma incremental añadiendo distancias de la base de datos a los prototipos mas similares (o menos distantes) de las clases previamente formadas, dependiendo de la elección del usuario.

Los prototipos se actualizan iterativamente y el número de clases se determina con el umbral de similitud permitido en cada clase. Este método produce buenos resultados si las clases reales son esféricas o quasi-esféricas. Si la estructura real del dominio (que suele ser totalmente desconocida cuando se enfrenta un problema de clustering) está lejos de esta situación, las clases propuestas por el método, pueden ser realmente distintas de las reales.

3.5. Isodata

Isodata [1] (Self-Organizing Data Analysis Techniques) es un algoritmo de clustering de particiones donde las clases se dividen o fusionan a partir de las existentes si se cumple cierta condición. Se divide una clase si tiene demasiadas instancias y varianza inusualmente alta en la variable de mayor ”spread”. Se fusionan si sus centros están suficientemente cerca, en base a parámetros que da el usuario.

De hecho, este método depende de varios parámetros, que indican los umbrales de distancia para dividir o fusionar clases, así como el número inicial y final de clases, o el máximo número de uniones por iteración. El sistema propone siempre valores por defecto para todos ellos. A menos que se quiera trabajar con esos valores, el uso apropiado de los parámetros, requiere una lectura detenida del manual de usuario antes de utilizarlo.

4. Aplicación a Estaciones Depuradoras de Aguas Residuales

El principal objetivo de una Estación Depuradora de Aguas Residuales es garantizar la calidad del agua de salida (respecto a los límites legales) para restablecer el equilibrio biológico del proceso, que a menudo se ve alterado por los vertidos de aguas residuales ya sean urbanas o industriales.

El proceso más común para lograr este objetivo es el tratamiento por fangos activos. Este proceso es complejo y delicado debido a las características intrínsecas de la propia agua residual y las graves consecuencias de una mala gestión de la planta [8].

La Base de Datos analizada en este artículo procede de la línea de aguas de una Estación Depuradora de Aguas Residuales de Catalunya. A continuación se presenta una breve descripción de las principales unidades de esta EDAR siguiendo el flujo desde que el agua entra en la estación hasta que desemboca en el medio receptor:

- *Pretratamiento* El principal objetivo del pretratamiento es eliminar del agua residual elementos sólidos mediante procesos mecánicos (rejas). Sin este pretatamiento estos mismos elementos podrían dañar los equipos y disminuir la eficiencia global del tratamiento.
- *Tratamiento Biológico* En el tratamiento biológico una población de microorganismos (biomasa) degrada los contaminantes que lleva el agua residual. El reactor de esta planta es del tipo *Doble etapa* y está dividido en:
  - Un primer reactor que funciona a *alta carga* con un tiempo de retención hidráulica bajo, donde se elimina la materia orgánica fácilmente biodegradable.
  - Un reactor secundario trabajando a *baja carga*, con un tiempo de retención hidráulica mayor donde se lleva a cabo la completa eliminación de la materia orgánica y la nitrificación

Una porción del agua residual puede pasar directamente de la entrada al segundo reactor biológico para proporcionar materia orgánica y biomasa para desnitrificar (la planta alberga esta posibilidad pero no la utiliza). La razón de esta separación es que así se puede tratar las substancias tóxicas en la primera fase, y proteger las bacterias sensibles a los tóxicos.

El agua que sale de los reactores se transporta a los tanques de sedimentación, donde la biomasa es separada del agua tratada. Ésta pasará a la línea de fangos mientras que el sobrenadante estará listo para volver al medio natural.

La Base de Datos, de 149 observaciones y 18 variables, fue obtenida entre enero y mayo de 2002. Cada observación refiere una media diaria, y consta de los datos que se toman en la planta de forma rutinaria para controlar el estado de operación. Las 18 variables se agrupan como sigue (fig. 2):

- Variables de Entrada (datos obtenidos en la entrada de la planta)
  - Q-E: Caudal de agua residual (m^3 diarios de agua)
  - DQO-E: Demanda Química de Oxígeno (mg/l)
  - MES-E: Sólidos en Suspensión (mg/l)
  - P-E: Fósforo en forma de Fosfato (mg/l)
- Variables de Proceso en la Primera Etapa (datos obtenidos en el primer reactor biológico):
  - MLSS-1: Sólidos en Suspensión en el Reactor (mg/l)
  - IVF-1: Öndice Volumétrico de Fangos (ml/g)
  - CM-1: Carga Orgánica (kg materia orgánica/ Kg MLSS)
- Variables de Salida de Primera Etapa (obtenidas tras la decantación en primera etapa):
  - DQO-P: Demanda Química de Oxígeno (mg/l)
  - MES-P: Sólidos en Suspensión (mg/l)
- Variables de Proceso en la Segunda Etapa (datos obtenidos en el segundo reactor biológico):
  - MLSS-2: Sólidos en Suspensión en el Reactor (mg/l)
  - IVF-2: Öndice Volumétrico de Fangos (ml/g)
  - CM-2: Carga oOgánica (kg materia orgánica/ Kg MLSS)
  - T-2: Temperatura (C^o)
  - EF-2: Tiempo de Residencia del Fango (días)
- Variables de Salida de la Planta (obtenidas tras la decantación en segunda etapa):
  - DQO-S: Demanda Química de Oxígeno (mg/l)
  - MES-S: Sólidos en Suspensión (mg/l)
  - NO3-S: Nitrógeno en forma de Nitrato(mg/l)
  - P-S: Fósforo en forma de Fosfato (mg/l).

5. Resultados

La Base de Datos se analizó con los diversos métodos de minería de datos disponibles en GESCONDA. Aquí se comparan los resultados obtenidos utilizando diferentes métodos de clasificación. Todos los métodos utilizados generan una partición de los datos, asignando cada objeto clase única.

Como la Base de Datos estudiada sólo contiene datos numéricos, todos los métodos utilizan distancias Euclídeas. En esta aplicación concreta no hay un aporte de información por parte de expertos en la relevancia de las variables, de manera que se supone que todas las variables tienen el mismo peso en todos los algoritmos. Los métodos utilizados en este trabajo fueron los siguientes (también se indica el identificador de las clasificaciones resultantes):

- Vecinos Recíprocos, utilizando el Criterio de Ward, (P_{RN}).
- K-means, , con un valor de 12 clases y utilizando una técnica de pseudo-bagging, (P_K).
- Marata, con un valor de 12 clases, (P_M).
- Nearest-neighbor, utilizando un límite máximo de 0,48 para los valores de similitud entre observaciones de la misma clase. El algoritmo generó una partición de 12 classes (P_{NN}).
- Isodata con los siguentes parámetros: 12 como número inicial y final de classes, límite de 0,05 para las uniones, 25 iteraciones y los valores por defecto para los demás parámetros, (P_I).

De hecho, vecinos recíprocos ya se utilizó previamente [7] para identificar situaciones características en este mismo dominio proporcionando una partición en 12 clases P_{RN} (el significado de estas clases se puede encontrar en este último trabajo). Dichos resultados fueron validados de forma manual por expertos, tomando como criterio de validación el *significado* de las clases identificadas. Así, P_{RN} se consideró como partición de referencia para la valoración de las demás. Utilizando esta aproximación,

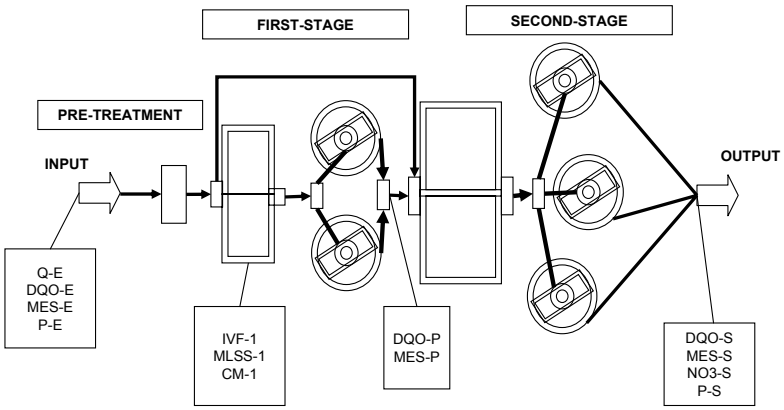

Figura 2: Architecture of the plant.

Partición	**% de objetos correctamente clasificados**
P_K	54.6
P_M	38.77
P_{NN}	50.015
P_I	44.52

Cuadro 1: Coeficientes de calidad respecto P_{RN}.

Partición	P_{RN}	P_K	P_M	P_{NN}
P_K	54.6			
P_M	50.015	22.18		
P_{NN}	38.77	46.42	41.81	
P_I	44.52	39.81	28.62	47.41

Cuadro 2: Similitudes entre particiones.

Partición	P_{RN}	P_K	P_M	P_{NN}	P_I
c1	28	26	81	52	22
c2	20	25	2	2	17
c3	4	4	1	1	10
c4	2	0	1	2	10
c5	12	1	3	1	9
c6	14	24	1	30	16
c7	25	27	50	35	11
c8	4	1	3	2	20
c9	8	8	1	1	11
c10	6	6	3	5	3
c11	11	13	2	17	15
c12	9	14	1	1	5

Cuadro 3: Tamaños de clase.

la tabla 1 muestra el porcentaje de objetos correctamente clasificados, respecto a P_{RN}. Parece ser que el método que proporciona unos resultados más cercanos a los aceptados por los expertos es K-means con pseudo-bagging; ese valor, incluso siendo cercano al 50 % , queda lejos de una asignación aleatoria de la etiqueta de clase, por ser el número de clases reconocidas mucho mayor que 2.

En realidad, este coeficiente de calidad puede ser utilizado como un tipo de *similitud* entre particiones, de modo que cuanto mayor es este coeficiente más similitud existe en la forma cómo los dos métodos clasifican los objetos. Esto permite construir la tabla 2, más apropiada para un análisis global de los resultados. Esta tabla muestra que los resultados más similares corresponden a P_{RN} y P_K con un 54.6 % de objetos clasificados de la misma forma, mientras que el par más diferente fue P_K y P_M con un 73.3 % de clasificados de forma diferente.

Además, teniendo una partición de referencia, GESCONDA proporciona un *reporte de validación* para las demás particiones (fig. 3) que, a parte de los coeficientes de calidad, establece las correspondencias entre las etiquetas de cada par de particiones (basándose en las intersecciones mayores) y proporciona información sobre los objetos que interesecan en ambas clases. A partir de ello, la tabla 3 presenta el tamaño de las clases para cada partición.

Puede observarse que con Marata se concentran la mayoría de objetos en dos clases, mientras el resto de clases contienen objetos aislados en general; ésta puede ser la razón por la que, para esta aplicación en particular, presenta menor parecido con otras particiones. Viendo la Tabla 3, Isodata parece ser el método que produce una mayor distribución de los objetos en clases de tamaños similares.

Hay clases que conservan su comportamiento de

forma independiente al método:

- Clases 1 y 7 (ambas refieren situaciones normales, [7]), grandes en todas las particiones.
- Clases:
  - c3 (situaciones anormales por proliferación de bacterias filamentosas (bulking) en la primera etapa que dificulta la sedimentabilidad)
  - c4 (días con sobrecarga orgánica)
  - c8 (períodos con nitrificación parcial por desarrollo de microorganismos autotróficos)
  - c10 (bulking viscoso)

  El resto de clase suelen ser pequeñas en todos los métodos (exepto Isodata).

Esto es consiente con el hecho de que las situaciones normales son las más frecuentes, y las clases más pequeñas representan desviaciones de esta normalidad que afortunadamente no se dan muy a menudo en la planta. Las clases restantes tienen un tamaño variable dependiendo del método.

La Figura 4 representa gráficamente las cinco particiones. Se puede observar que en general, hay algunos grupos fuertes bien identificados por todos los métodos de clasificación utilizados, mientras que otros datos son asignados a una clase u otra dependiendo de la sensibilidad del método.

Es interesante destacar que estos grupos estables son siempre reconocidos, aunque la naturaleza del método de partición sea diferente. Esto ocurre por ejemplo con el método de vecinos recíprocos, originalmente un método matemático que nace de la estadística y basado en conceptos algebraicos, que se puede comparar con otros métodos originalmente concebidos en el contexto de la IA, basado en el paradigma lógico y la generación de prototipos.

6. CONCLUSIONES Y TRABAJO FUTURO

Se ha analizado una Base de Datos con diferentes métodos de minería de datos disponibles en GESCONDA; en este trabajo se comparan los resultados obtenidos con diferentes métodos de clasificación.

Una vez obtenidos los resultados, parece ser que, para este tipo de aplicación, el método que proporciona unos resultados más cercanos a los aceptados por los expertos es K-means con pseudo-bagging, siendo K-means y Marata los que proporcionan resultados más dispares entre si, difiriendo ambas clasificaciones en un 73 % de los objetos.

```
Validacio algorisme: BAGGING SOBRE K-MEANS
Respecte atribut: CLASSE
_____

PERCENTATGE CORRECTESA CLASSIFICACIO: 54.580387
PROTOTIP 1
Es correspon amb el valor 12
Valors correctes: 8/14
Percentatge valors del prototip correctes: 57.142857 %
Percentatge valors de la taula en el prototip: 88.888885 %
Percentatge correctesa: 50.793655 %

PROTOTIP 2
Es correspon amb el valor 11
Valors correctes: 11/13
Percentatge valors del prototip correctes: 84.61539 %
Percentatge valors de la taula en el prototip: 100.0 %
Percentatge correctesa: 84.61539 %

PROTOTIP 3
Es correspon amb el valor 2
Valors correctes: 20/25
Percentatge valors del prototip correctes: 80.0 %
Percentatge valors de la taula en el prototip: 76.92308 %
Percentatge correctesa: 61.538464 %

PROTOTIP 4
Es correspon amb el valor 9
Valors correctes: 0/8
Percentatge valors del prototip correctes: 0.0 %
Percentatge valors de la taula en el prototip: 0.0 %
Percentatge correctesa: 0.0 %

PROTOTIP 5
Es correspon amb el valor 1
Valors correctes: 20/26
Percentatge valors del prototip correctes: 76.92308 %
Percentatge valors de la taula en el prototip: 71.42857 %
Percentatge correctesa: 54.945057 %

PROTOTIP 6
Es correspon amb el valor 10
Valors correctes: 6/6
Percentatge valors del prototip correctes: 100.0 %
Percentatge valors de la taula en el prototip: 100.0 %
Percentatge correctesa: 100.0 %

PROTOTIP 7
Es correspon amb el valor 6
Valors correctes: 14/24
Percentatge valors del prototip correctes: 58.333332 %
Percentatge valors de la taula en el prototip: 100.0 %
Percentatge correctesa: 58.333332 %

PROTOTIP 8
.
.
.
```

Figura 3: Extracto de un reporte de validación de P_K respecto a P_{RN} generado por GESCONDA.

De hecho Marata concentra la mayoría de objetos en dos clases, mientras el resto contiene objetos aislados en general; esta puede ser la razón por la que, para esta aplicación particular presenta bajos valores de similitud respecto a otras particiones.

En oposición, Isodata parece ser el método que produce una mayor uniformidad en la distribución de los objetos en clases de tamaño similar.

También se ha observado que hay clases que se conservan independientemente del método. Algunas referentes a situaciones de normalidad son las más extensas en todas las particiones, mientras otras referentes a anomalías son siempre pequeñas. Esto es consistente con el hecho de que las situaciones normales son las más frecuentes en la operación de plantas reales y las pequeñas corresponden a episodios que afortunadamente no se dan muy a menudo.

Utilizando una de las opciones de GESCONDA que muestra gráficamente las clases, se ha observado que en general existen unos grupos fuertes bien identificados por todos los métodos de clasificación utilizados, mientras otros días se asignan a diferentes clases dependiendo de la sensibilidad del método. Es interesante notar que existen grupos estables que siempre son reconocidos, a pesar de que la naturaleza del método sea muy diferente, como aquí, donde métodos matemáticos se comparan con otros concebidos originalmente en el seno de la IA.

Dichos grupos, seguramente contribuirán a obtener una descripción más precisa de las situaciones dadas en la planta incrementando la correspondencia entre clases descubiertas y reales. Tales comparaciones pueden incluso mejorar la validación de resultados y consolidar el conocimiento adquirido del dominio, disminuyendo la incertidumbre inherente al proceso de inducción y aumentando, por tanto, la calidad de la posterior toma de decisiones.

La cuantificación de esta incertidumbre requiere más trabajo, y será tratada en posteriores investigaciones. También sería interesante consultar a expertos que interpretasen de forma manual las particiones obtenidas (P_K, P_M, P_{NN}, P_I) y proceder con una comparación conceptual en términos de significado de clases, de validando, de forma paralela a este estudio más formal, las presentes conclusiones.

También se ha visto aquí que una herramienta integrada que proporcione distintos métodos de minería de datos (en este caso particular, de clasificación), con la posibilidad de comparar resultados y compartir información, ayuda a mejorar la inducción de conocimiento. En este sentido, los autores creen que GESCONDA proporciona un buen marco para afrontar este tipo de comparaciones.

En el contexto particular de la clasificación, analizar subgrupos de objetos cuya clase cambia de un método a otro, permite obtener información acerca del rendimiento y calidad de los distintos métodos, así como las características de los grupos más estables. Éstos servirán para refinar el conocimiento descubierto a partir de los datos, e incrementar la confianza en el modelo de conocimiento descubierto a partir de los datos, como estimación del modelo de conocimiento real, normalmente desconocido.

Las posibilidades, proporcionadas por GESCONDA, de analizar las clases bajo esta aproximación global y comparativa suponen una contribución al proceso de validación de clases, planteado como un problema abierto al principio del artículo.

Posteriormente, se utilizará una técnica de clasificación hibrida (clasificación basada en reglas [6])que permite la previa introducción de conocimiento experto para que sea considerado en el proceso de clasificación sea empleado para mejorar el descubrimiento de conocimiento em el dominio.

Finalmente, los resultados resultados del proceso de clasificación se utilizarán como entrada para la generación inductiva directa de reglas de clasificación. Ello permitirá efectuar las tareas de predicción de la etiqueta de clase para nuevas observaciones.

Una vez, consolidado el conocimiento descubierto en esta aplicación particular, utilizando el marco proporcionado por GESCONDA, se pueden realizar otras comparaciones con otros softwares existentes como WEKA u otras herramientas. En la actualidad, las técnicas de clasificación implementadas en WEKA son Kmeans, COBWeb, EM y FarthestFirst. La primera está incluida en GESCONDA, y la segunda está mejorada con COBWeb/3, también incluido en GESCONDA. Por tanto, la comparación se restringirá a las técnicas adicionales que ofrezca WEKA. El mismo tipo de comparaciones se realizará con otras herramientas de software.

Agradecimientos Participaciones: A. Clavell en el desarrollo del software; A I. R.-Roda en la interpretación de resultados. $TIN2004-01368$ financia parcialmente la investigación.

Referencias

[1] G.H. Ball and D. J. Hall. ISODATA, a novel method of data analysis and pattern classification. Technical report, Stanford Research Institute, 1965.

[2] T.M. Cover and P.E. Hart. Nearest Neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27., 1968.

[3] R. Dubes and A.K. Jain. *Clustering Methodologies in Exploratory Data Analysis*, volume 19. Advances in Computers, 1980.

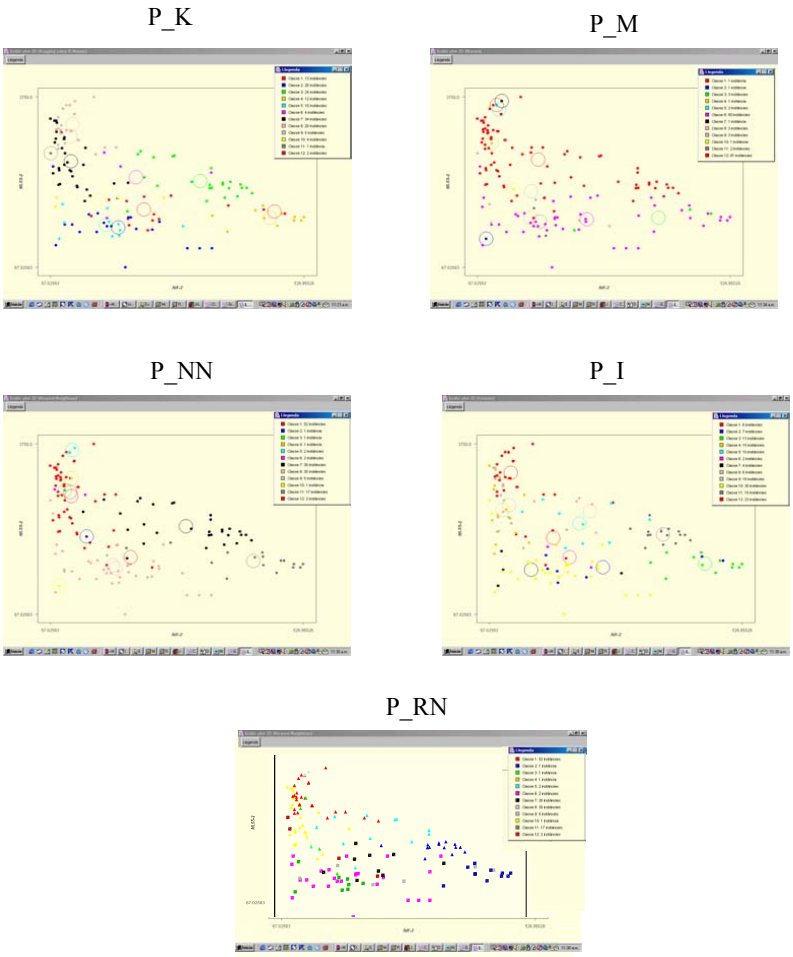

Figura 4: Visualización de las clases.

[4] K. Gibert, T. Aluja, and U. Cortés. Knowledge Discovery with Clustering Based on Rules. Interpreting results. LNAI v. 1510, pages 83–92. Springer, 1998.

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

[6] 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.

[7] K. Gibert, X. Flores, I. Rodríguez-Roda, and M. Sànchez-Marrè. Knowledge discovery in environmental databases using gesconda. In *Complexity and Integrated Resources Management, Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society*, volume I, pages 51–56, Manno, October 2004. iEMSs.

[8] J.M. Gimeno, Béjar J., and U. M. Sànchez-Marrè, Cortés. Discovering and modelling process change: An application to industrial processes. In *Practical Applications of Data Mining and Knowledge Disc.*.

[9] M. Roux. *Algorithmes de classification*. Mason, 1985. Paris, France.

[10] M. Sànchez-Marrè, K. Gibert, I. Rodríguez-Roda, and et al. Development of an intelligent data analysis system for knowledge management in environmental data bases. In *First Int'l Conf. of Int'l EMS Society*, pages 420–425, Lugano, Suiza, 2002.

[11] J.H. Ward. *Hierarchical grouping to optimize an objective function*. J. Am. Statis. Ass. 58, 1963.