# INSTITUTO POLITECNICO NACIONAL

## CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

# CARACTERIZACIÓN E INTERPRETACIÓN DE DESCRIPCIONES CONCEPTUALES EN DOMINIOS POCO ESTRUCTURADOS

T E S I S

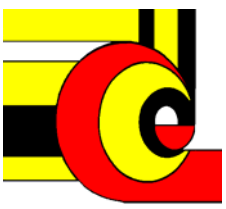
QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

**FERNANDO VÁZQUEZ TORRES**

DIRECTORES DE TESIS:

DR. CORNELIO YÁÑEZ MÁRQUEZ
DR. JUAN LUIS DÍAZ DE LEÓN SANTIAGO



MÉXICO, D.F. MAYO DE 2008

*ÍNDICE*

CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO

## RESUMEN

El proyecto de investigación que se presenta en este trabajo de tesis doctoral, tiene como objetivo fundamental: el diseño de una Familia de Algoritmos con aprendizaje inductivo para clasificación supervisada, que permita identificar las características relevantes de las clases resultantes, obtenidas de una partición de referencia, las cuales proporcionen un Sistema de Caracterización útil en la Generación Automática de las Descripciones Conceptuales de estas clases con variables numéricas, para conocer la estructura semántica de los *dominios poco estructurados*, de bajo costo y útil en tareas de predicción y/o diagnóstico. Obtener, además, una metodología que permita hacer contribuciones a la validación de clases, en relación con su representación formal y su calidad, considerando esta última como la utilidad de las clases formadas a través del nivel característico de cada variable, dando como resultado una representación difusa de los grados de pertenencia de los valores de la variable a las distintas clases, lo que constituye un cómodo soporte al significado de las clases y en consecuencia a su interpretación. Y además, la validación de campo a través de la aplicación del modelo propuesto al control y administración de una planta de tratamiento de aguas residuales usando variables numéricas, lo cual permitió contrastar los resultados obtenidos del modelo matemático con la situación real.

El modelo propuesto tiene como punto de partida el estudio del *boxplot múltiple*, la formalización de la metodología del sistema de caracterización de clases se hizo diseñando una nueva familia de algoritmos que no depende de la representación gráfica del boxplot, desde un punto de vista semántico, en comparación con otros métodos inductivos, cuando se aplica sobre datos provenientes de un *dominio poco estructurado*; además, de una nueva aproximación para discretizar el espacio de representación en términos de intervalos de longitud variable como base de la metodología, y contribuciones a la validación de la clasificación, en cuanto a su representación y calidad, en el sentido de que una clasificación es válida si probamos que las clases obtenidas tienen sentido o utilidad y a la generación automática de clases resultantes como base del proceso de predicción y/o diagnóstico.

Los resultados que se han obtenido son prometedores, constituyendo un primer paso para establecer un modelo híbrido entre dos enfoques: el estadístico y el lógico combinatorio para abordar sistemas dinámicos en el contexto de los dominios poco estructurados, que nos permita la obtención tanto la caracterización de clases como sus descripciones conceptuales para describir los objetos de estudio.



Palabras claves: Caracterización, Descripciones Conceptuales, Dominios poco Estructurados, Enfoque Estadístico, Enfoque Lógico Combinatorio.

## ABSTRACT

The project of investigation that is presented in this work of doctoral thesis, has as fundamental aim the design of a Family of Algorithms with inductive learning for supervised classification**,** which allows to identify the relevant characteristics of the resultant classes, obtained from a partition of reference, which provide a System of useful Characterization in the Automatic Generation of the Conceptual Descriptions of these classes with numerical variables, to know the semantic structure of the little constructed domains, of low cost and useful in tasks of prediction and /or diagnosis. Also to obtain, a methodology that allows to do contributions to the validation of classes, in relation with its formal representation and its quality, considering the this last one as the utility of the classes formed through the typical level of every variable, giving as a result a diffuse representation of the degrees of belonging of the values of the variable to the different classes, which constitutes a comfortable support to the meaning of the classes and in consequence to its interpretation.

Moreover, the field validity of the application of the model proposed to the control and administration of a Wastewater Treatment Plant using numerical variables, which allowed contrasting the results obtained from the mathematical model with the real situation.

The proposed model takes as a starting point the study of the multiple box plot, the formalization of the methodology of the system of characterization of classes which was done designing a new family of algorithms that does not depend on the graphical representation of the boxplot, from a semantic point of view, in comparison with other inductive methods, when it is applied to information from a little constructed domain; also, of a new approximation to discretize the space of representation in terms of intervals of changeable length as base of the methodology, and contributions to the validation of the classification, as for its representation and quality, to the effect that a classification is valid if we prove that the obtained classes make sense or are useful and to the automatic generation of resultant classes as base of the process of prediction and / or diagnostic.

The results that have been obtained are promising, constituting the first step in establishing a hybrid model between two approaches: the statistical and the combinatory logical to approach dynamic systems in the context of the little constructed domains, which allows the obtaining of both the characterization of classes and its conceptual descriptions to describe the objects of study.



Key words: Characterization, Conceptual Descriptions, Little constructed Domains, Statistical Approach, and Combinatory logical Approach.

# CAPÍTULO 1.

# INTRODUCCIÓN

En este trabajo de tesis se presenta una familia de algoritmos a través de un modelo original teórico-conceptual, que incluye una *nueva forma* de *extraer conocimiento útil* de los así llamados dominios poco estructurados. Este nuevo modelo permite identificar las características relevantes de las clases resultantes obtenidas de una partición de referencia, lo cual conlleva a la generación automática de las descripciones e interpretaciones conceptuales de estas clases; lo anterior, basado en una combinación de diferentes herramientas y técnicas de estadística (boxplot múltiple, análisis de datos), inteligencia artificial (aprendizaje automático, sistemas basados en conocimientos) y lógica difusa (modelos y razonamiento aproximado).

## 1.1 ANTECEDENTES

La comprensión de la naturaleza de los métodos que utilizamos los seres humanos para clasificar datos o conocimientos, es un problema de gran interés teórico y práctico para todas las ciencias cognitivas, ya que la acción de clasificar es una de las etapas iniciales de los procesos de adquisición de conocimiento en cualquier campo científico [1]. Teóricamente, la comprensión del concepto “clasificar” contribuye a entender mejor lo que implica el “aprendizaje”; de hecho, es difícil concebir una forma de aprendizaje sin haber pasado antes por una forma previa de clasificación [2].

Por otro lado, en la práctica, el desarrollo de sistemas automáticos de clasificación es, hoy por hoy, una necesidad imperiosa de la sociedad actual ya que en muchos procesos la cantidad de datos que se genera es tan grande, que resulta muy difícil manipularlos y transmitirlos sin el auxilio de esta clase de sistemas. La clasificación automática se desarrolla normalmente en dos grandes enfoques [3]: a) A partir de una clasificación de referencia de un universo de discurso, definir reglas para decidir la clase a la que pertenece cada elemento del discurso (*aprendizaje supervisado*); y b) Dado un universo de discurso, construir una clasificación adecuada del mismo (*aprendizaje no supervisado*). Diversas disciplinas del conocimiento humano han contribuido para la creación, diseño y desarrollo de reconocedores y clasificadores automáticos de patrones, dentro de ambos enfoques. De relevancia para este trabajo de tesis, son las contribuciones de la estadística, la inteligencia artificial y la lógica difusa.

En el campo de la estadística, personajes como Galton [58], Pearson [59], Mahalanobis [60] y Fisher [61] han sido pioneros en la aplicación de métodos estadísticos en el reconocimiento y clasificación de patrones, y actualmente se aplican diversas técnicas desarrolladas por ellos y sus seguidores, como son: boxplot múltiple, análisis de regresión, análisis de componentes principales y análisis discriminante, entre otras. Más

recientemente, se han desarrollado algunas técnicas estadísticas, siendo notable la presencia de *Clustering* [62], dentro del aprendizaje no supervisado.

En lo que respecta al aprendizaje supervisado, son dignos de mención algoritmos como el bayesiano [86], el euclidiano [87], el k-NN (los k vecinos más cercanos) [98-102], los árboles de decisión [103,104], y otros algoritmos cuyo fundamento queda en disciplinas diferentes de la estadística, como las máquinas de soporte vectorial [105], las redes neuronales [66,91,92] y las memorias asociativas [93-97].

Desde su creación a mediados de los 50 del siglo pasado, los científicos pioneros de la inteligencia artificial como Von Neumann [63] y Minsky [64], además de una pléyade de científicos afines a sus ideas, han incidido mediante sus propuestas científicas en las áreas de reconocimiento y clasificación de patrones [71]; así, a través del tiempo se han incorporado: el aprendizaje automático [4-7], los sistemas expertos [8-10,65], la inteligencia artificial evolutiva (algoritmos genéticos) [67,68], la inteligencia artificial macro-distribuida (sistemas multi-agentes) [69,70] y el descubrimiento del conocimiento en bases de datos (KDD) [32].

La lógica difusa, por su parte, como una extensión de la lógica multi-valuada de Lukasiewicz [143] que a su vez se derivó de la lógica booleana, ha permitido considerar desde una perspectiva más amplia a los problemas relacionados con el modelado del razonamiento. Formalizada por Zadeh en 1965, la lógica difusa refleja las imprecisiones de los datos generados en el mudo real [116-127]. Numerosos equipos de investigación científica actualmente trabajan en el diseño de sistemas donde se aplican de manera directa los conceptos y resultados de la lógica y el razonamiento difusos, siendo relevante su uso en áreas como control [128,129] y sistemas de clasificación basados en reglas difusas [130-141].

## 1.2 MOTIVACIÓN

Es un hecho indiscutible que la mayoría de los sistemas expertos de la primera generación como MYCIN [4] (sistema experto de la Universidad de Stanford para consulta y diagnóstico de infecciones de la sangre), INTERNIST [5], [6] (un sistema experto de la Universidad de Pittsburg usado en la identificación de infecciones en medicina interna con una base de conocimiento de más de 500 enfermedades y sus síntomas), PLANT/DS [7] (sistema experto utilizado para el diagnóstico de enfermedades y daños producidos por insectos en soya) y algunos otros son, en la práctica, sistemas clasificadores. Esta clase de sistemas utilizan un conjunto de reglas implementadas como árboles de decisiones para determinar la clase a la que pertenece un individuo de un cierto dominio de estudio [8].

En el enfoque clásico a este problema, el experto humano es el responsable de decidir cuáles son las variables “relevantes” para la formulación de las reglas de clasificación. Cuando se procede de esta forma, el diseñador

del sistema requiere información que el experto no está preparado para proporcionar debido, fundamentalmente, a la falta de familiaridad con los términos que se utilizan en el sistema informático. Esto provoca graves problemas de comunicación al tratarse de personas que tienen formaciones diferentes, por lo que la extracción de conocimiento se hace difícil de superar y consume mucho tiempo [4-9].

Por lo tanto, la clasificación de ejemplos se presenta como una herramienta alternativa posible para la extracción del conocimiento de las descripciones que los expertos podrán dar a sus dominios.

Esta es la razón de que hayan surgido diversas metodologías que permiten el análisis de la información con miras a crear agrupaciones de observaciones para su posterior caracterización e interpretación [9].

Un enfoque diferente es el de la Inteligencia Artificial, ya que se ha decidido por el uso de técnicas de aprendizaje inductivo para la automatización de procesos. De esta manera, a partir de una colección de individuos de un dominio propuestos por el experto o extraídos directamente de dicho dominio y, de estas técnicas, se puede descubrir el conocimiento oculto en los datos y en consecuencia conocer la estructura semántica del dominio, útil en la construcción de bases de conocimiento. Este mecanismo parece más viable ya que se ha observado que los expertos tienen más facilidad para dar ejemplos de instancias de su dominio, que para expresar los conceptos o reglas que les permiten identificarlas [10].

En el caso del dominio donde la estructura semántica esté claramente definida y exista una manera de discernir entre las diferentes categorías que lo componen, esta metodología es clara y provechosa a la hora de construir bases de conocimiento para sistemas basados en conocimiento, disminuyendo la interacción experto-diseñador del sistema.

Todos los problemas de adquisición de conocimiento mencionados se agravan si el dominio sobre el que se está tratando es un *Dominio poco Estructurado (ILL-Structured Domains, ISD)*. Estos dominios se caracterizan por [10], [11]:

- No existir consenso entre los expertos para la definición de todos los conceptos y objetos que los componen y las relaciones entre éstos.
- Complejidad del área de conocimiento en concreto, ya sea por la falta de una metodología de investigación aceptada por todos los expertos o por un continuo cambio en el conocimiento o en su extensión.
- Las variables que describen a los individuos pueden ser cuantitativas o cualitativas.
- Los expertos suelen disponer de grandes cantidades de conocimiento implícito, además de manejar diversos grados de especificidad, lo que hace a este conocimiento parcial y no homogéneo.

De esta forma la alternativa que parece más prometedora para resolver estas limitaciones es liberar al experto de este trabajo, mediante el desarrollo de técnicas que a partir de la evidencia empírica en forma de ejemplos, identifiquen las variables más relevantes y formulen reglas que expresen las regularidades existentes en los datos.

En las últimas décadas, el crecimiento explosivo de los avances científicos y tecnológicos ha generado sistemas complejos que han rebasado nuestra capacidad para analizarlos e interpretarlos, creando la necesidad de una nueva generación de métodos, técnicas y herramientas con la capacidad para asistir inteligente y automáticamente a los seres humanos en el análisis de estas bases de datos para *extraer conocimiento útil* que represente los dominios del mundo real [10].

Descubrir la estructura semántica o extraer conocimiento de *dominios reales y complejos* (dominios poco estructurados) no es tarea fácil y requerimos de combinar técnicas y herramientas de diversos campos para construir *Sistema Híbridos* que permitan encontrar e interpretar patrones especiales (o conceptos) en las bases de datos; y así, extraer conocimiento útil que represente estos dominios, con mejor desempeño que las técnicas tradicionales o los enfoques clásicos de los sistemas basados en conocimiento [12].

Actualmente, dada una partición (clasificación de referencia) de un conjunto grande de individuos, es necesario introducir herramientas para asistir al usuario en las tareas de interpretación, con objeto de *establecer el significado de las clases resultantes*. Frecuentemente, no es suficiente descubrir la construcción automática de clases, sino poder entender por qué se detectaron estas clases. Algunos paquetes estadísticos orientados al análisis multivariable y de propósito general como SPAD (Système Portable pour L'Analyse des Donees Textuelle) y SPSS (Statistical Package for Social Sciences) [13], incluyen varias herramientas orientadas a la interpretación de una clasificación dada, como la posibilidad de calcular la contribución de cierta variable a la formación de una clase. Sin embargo, en la etapa final, la interpretación misma deberá hacerla el usuario en una forma no-sistemática, usando su propia experiencia y la tarea llega a dificultarse cuando el número de clases resultantes aumenta, así como también cuando el número de variables también aumenta para describir los datos.

Por otro lado, no se tiene información sobre un criterio objetivo para determinar *la validación de clases*, considerándola como el grado de interpretabilidad o utilidad de éstas.

Por lo anterior, el interés de este trabajo es presentar una propuesta metodológica híbrida que combine herramientas y técnicas de estadística, inteligencia artificial y lógica difusa en forma cooperativa, considerando que, a partir de las variables cuantitativas de los datos que definen a los individuos pertenecientes a cierto dominio, podemos identificar cuáles son las situaciones características (clases resultantes) que se pueden encontrar en él, analizarlas,

estudiar su significado y, en consecuencia, conocer la estructura semántica del dominio al cual pertenecen dichos individuos [152].

Una vez identificadas e interpretadas estas situaciones típicas, el conocimiento generado puede ser usado posteriormente como herramienta de apoyo al proceso de administración o toma de decisiones. Incluso se ha llegado a decir que *la validación de una clasificación* (problema abierto) consiste, precisamente, en *probar* que las clases tienen *sentido o utilidad* [14].

En esta dirección, esta propuesta pretende llegar un poco más lejos, y establecer las bases metodológicas que faciliten la generación automática de descripciones e interpretaciones conceptuales en estos dominios reales y complejos [151].

Como parte de la motivación para realizar este trabajo de tesis, el caso de estudio es una aplicación de la metodología al proceso de tratamiento de aguas residuales de una planta depuradora [15], [16] y [17]. Cabe hacer notar que el autor de esta tesis participó activamente con el equipo interdisciplinario creador de la base de datos *Water Treatement Plant Database*, la cual se encuentra disponible en el repositorio de UCI [153].

Cuando la planta depuradora no funciona bajo condiciones normales, se deben tomar decisiones para modificar algunos parámetros del proceso de depuración y re-establecer lo antes posible la normalidad [18] y [19]; razón por la cual, es importante contar con un sistema que proporcione información relevante sobre la situación que la planta tiene en un momento específico.

## 1.3 PLANTEAMIENTO DEL PROBLEMA

A partir de una matriz de datos $X$ y una partición de referencia $P$ obtenemos un conjunto de individuos con una clase asignada de acuerdo con el proceso de clasificación dado. Se plantea desarrollar una nueva Familia de Algoritmos para establecer una metodología que identifique las características relevantes de las diferentes clases, obtenidas de tal forma que proporcionen un *Sistema de Caracterización de clases para obtener descripciones conceptuales directamente comprensibles al usuario-experto*.

Uno de los problemas principales de las técnicas de custering es que la validación de resultados es un problema de vital importancia. Es fácil evaluar un conjunto de clases en términos de criterios de exactitud siempre que exista una partición de referencia de los datos y si la comparación es posible [20]. Pero desafortunadamente, en la mayoría de las situaciones donde se requiere hacer Clustering (técnicas que intentan determinar si existen grupos) no existe y esta aproximación no es útil. Solamente la utilidad de una clasificación puede usarse para decidir si es correcta o no [21]. Evaluar la estimación de precisión de una clasificación dada requiere de un método, en esta investigación utilizamos Ten-Fols Cross-Validation; además, requerimos de un mecanismo

que permita comprender el significado de las clases identificadas para finalmente decidir si son útiles o no.

Este proceso, conocido comúnmente como *Interpretación de las clases resultantes*, tradicionalmente lo realiza el analista informático en una forma no sistemática, usando sus conocimientos y experiencia para poner de manifiesto las principales diferencias entre clases, y posteriormente, en estrecha colaboración con el experto en la materia, analiza las clases y estudia su significado para darles una interpretación. Este proceso llega a dificultarse cuando el número de clases aumenta y el número de variables utilizado para describir los datos también aumenta.

Es en esta línea de investigación donde se propone una metodología híbrida que representa una *nueva forma de extraer conocimiento útil* directamente comprensible al usuario-experto, usando una combinación de diferentes herramientas y técnicas de estadística (boxplot múltiple, análisis de datos), inteligencia artificial (aprendizaje automático, sistemas basados en conocimientos) y lógica difusa (modelos y razonamiento aproximado) para soportar la toma de decisiones en estos dominios.

## 1.4 OBJETIVO

Desarrollar una metodología para identificar las características relevantes de las clases resultantes, obtenidas de una partición de referencia, la cual proporcione un Sistema de Caracterización útil en la generación automática de las descripciones conceptuales de estas clases con variables numéricas, que permita conocer la estructura semántica de los dominios poco estructurados, de bajo costo y útil en tareas de predicción o diagnóstico. Obtener, además, un modelo conceptual que permita hacer contribuciones a la validación de clases, en relación con su representación formal y su calidad, considerando esta última como el grado de interpretabilidad o utilidad de las clases formadas.

## 1.5 CONTRIBUCIONES

**A nivel teórico conceptual:**

- Una nueva familia de algoritmos denominada Boxplot-based rule induction (BPRI), la cual representan algoritmos para clasificación supervisada con aprendizaje inductivo y que proporciona un sistema de “Caracterización de las clases en una muestra de supervisión”.
- Una herramienta para la Identificación de variables caracterizadoras.
- Método de inducción de reglas basado en intervalos de longitud variable derivada del uso del boxplot.

**A nivel metodológico:**

- Un método para determinar el nivel característico de cada atributo o variable en el contexto de cada clase como base del significado de las clases y que sirve de apoyo a la interpretación de clases.
- Comprobación de los diferentes tipos de valores de caracterización.
- Uso de criterios de agregación para clasificación: votación, probabilidad máxima y suma de probabilidades.
- Uso del paradigma difuso como soporte en la interpretación de clases.
- Una metodología con buen desempeño en cuanto a tareas de predicción y diagnóstico se refieren.
- El costo computacional de la metodología implementada es bajo en relación a la información que proporciona, puesto que se resuelve un análisis de intersecciones de grado $\xi$ (en el caso de estudio 20), calculando solamente $\xi$ máximos y $\xi$ mínimos y ordenándolos.

## 1.6 ORGANIZACIÓN DEL DOCUMENTO

La estructura de este documento consta de: el capítulo 1 que trata sobre los antecedentes, la motivación que dio origen a este trabajo; así como, el planteamiento, objetivo y contribuciones de este trabajo; el capítulo 2 describe el estado del arte que permite contextualizar el tema de tesis; el capítulo 3 trata sobre los conocimientos de los conceptos básicos de estadística, aprendizaje automático y lógica difusa que dan soporte a esta tesis; el capítulo 4 describe el diseño de la familia de algoritmos a través del modelo propuesto; en el capítulo 5 se presenta una aplicación con la implementación de la metodología a través del sistema CIADEC a un caso de estudio; en el capítulo 6 se presentan las conclusiones, el trabajo futuro y las publicaciones propias del autor de esta tesis y finalmente, un Apéndice que contiene un apartado: el apartado A se presenta de forma general, la arquitectura, funcionalidades e implementación de la metodología propuesta a través del sistema CIADEC.

# CAPÍTULO 2.

# ESTADO DEL ARTE

## 2.1 EL RECONOCIMIENTO DE PATRONES

El *Reconocimiento de Patrones* (*Pattern Recognition*) es una "disciplina de carácter multidisciplinario cuyo objetivo de estudio son los procesos de identificación, caracterización, clasificación y pronostico sobre objetos, físicos o abstractos con el propósito de extraer información que permita establecer propiedades de o entre conjuntos de dichos objetos, así como las metodologías y técnicas relacionadas con dichos procesos" [86-88]. Los objetos físicos pueden ser espaciales como: caracteres, imágenes entre otros y temporales como: formas de onda (voz), series entre otros y los abstractos como: razonamiento, soluciones a problemas, etc. Así tenemos, por ejemplo, patrones visuales basados en imágenes aéreas o satelitales, de problemas de clasificación y diagnóstico en algunos campos (como la medicina o la balística). También se puede aplicar a problemas relacionados con el campo del control inteligente, en el cual los sistemas complejos neuronales suministran la capacidad de aprendizaje y la lógica borrosa permite la extracción de las reglas de clasificación o diagnóstico [98-101]. A través del tiempo el Reconocimiento de Patrones ha ido evolucionando y tomando distintos nombres como Machine Learning (Aprendizaje Automático) o el más reciente como Data Mining (Minería de Datos) o Knowledge Discovery of Data (KDD, Descubrimiento de Conocimiento en bases de datos) ha medida que se han ido incorporando algunas otras métodos y/o técnica al reconocimiento de patrones.

La Inteligencia Artificial y la Estadística son áreas de investigación interdisciplinaria cuyo origen podríamos situar en la fundación de la Artificial Intelligence and Statistics (AI & Stats) Society por Douglas Fisher en 1985, ligada al First International Workshop on Artificial Intelligence and Statistics que impulsó Bill Gale de los American Telephone & Telegraph Laboratorios (AT & T). Desde entonces la conferencia internacional de dicha sociedad se ha venido celebrando bianualmente de forma ininterrumpida. El principal objetivo de la Artificial Intelligence & Stats Society es promover la comunicación entre la comunidad Estadística y la de Inteligencia Artificial [12].

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...*" [80]

lo que ha sido un motivo de reflexión desde entonces; en este trabajo de tesis se toma como punto de partida.

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 estos problemas, uno de los más conocidos es el del Reconocimiento de Patrones que puede ser de clasificación, caracterización o pronóstico. En cuanto a la clasificación esta puede ser de clasificación no-supervisada (por ejemplo, clustering) que básicamente consiste en encontrar las clases en que se estructura un dominio dado [110], y el otro una clasificación supervisada que consiste, en dada una clasificación de referencia de un cierto dominio encontrar la clase a la que pertenece cualquier individuo de éste. Este último problema ha sido parte del objeto de estudio de esta investigación.

Nota. A lo largo de este documento de tesis, cuando se habla de "dominio" se entenderá como "un espacio de representación", en el sentido de una representación del conocimiento a través de una Tabla Objeto-Propiedad.

Por otro lado, en 1989 se celebra en el seno del IJCAI (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"* [33, 34].

y la Minería de Datos se consolida rápidamente como un área de investigación también interdisciplinaria 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 fenómeno abrumador denominado 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 [115]:

Así, KDD se sitúa como una línea emergente del Reconocimiento de Patrones, combinando métodos y/o técnicas de Data Mining con otras herramientas para extraer conocimiento de los datos, lo que significa que en el fondo estamos haciendo reconocimiento de patrones y es aquí exactamente donde se sitúa uno de los objetivos de este trabajo de tesis doctoral.

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.

Como ya hemos dicho, el reconocimiento de patrones y en particular la clasificación es uno de los problemas que han sido objeto tanto de la Estadística como de la Inteligencia Artificial simultáneamente. En efecto, ambas disciplinas proporcionan diferentes métodos para descubrir en un espacio de

representación cuáles son las clases subyacentes, en el problema de clasificación no supervisada. En realidad, las técnicas de clustering son las más populares para separar datos en grupos y una de las técnicas de Minería de Datos más utilizadas [34]. 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 [34]. De hecho, un buen número de aplicaciones reales en KDD, o bien requieren un proceso de clasificación, o son reducibles a él [43].

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 que aborda la clasificación desde un punto de vista formal. En la Principles of Numerical Taxonomy de Sokal y Sneath [62], [167], [168], [169], [170] y [171] 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 estudiar los procesos de clasificación. Así, Michalski [7,57] inicia con la clasificación conceptual una línea de métodos de clasificación basados en la generalización de conceptos. Estos parten de una matriz de datos donde todas las variables (atributos en el contexto de la Inteligencia Artificial) son categóricas (cualitativas). En cuanto al hecho de clasificar matrices de datos mixtas, ya Anderberg [72], propone tres estrategias principales: i) el particionamiento de variables, dividiéndolas por tipos y reduciendo el análisis al tipo dominante (si es numérico, análisis de correspondencias seguido de un Clustering sobre las componentes factoriales [73], [74], lo que produce clases en un espacio ficticio de difícil interpretación); entre otras cosas, esta aproximación pierde la información de los grupos no dominantes, ii) realizar la conversión de todas las variables a un único tipo, conservando el máximo de información posible[2] [73], [74],[75], 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. En la literatura se hallan diferentes propuestas en esta dirección Gower 71 [47], Gowda & Diday 91 [36], Gibert 91 [24, 26], Ichino & Yaguchi 94 [48], Ralambondrainy 95 [75], Ruiz-Schulcloper [76].

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

[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 $\chi^2$ [73], [74] y [75]. 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 común [35]. Ello implica la perdida 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 [29].

Actualmente el avance tecnológico informático ha posibilitado que la captura de datos sea fácil y su almacenamiento tenga un costo prácticamente nulo. Con el desarrollo del software y el hardware y la rápida informatización de los negocios, enormes cantidades de datos son recogidas y almacenados en bases de datos. El resultado es que para analizar estas enormes cantidades de datos, las herramientas tradicionales de gestión de datos junto con técnicas estadísticas no son adecuadas.

Es conocido que los datos por sí solos no producen beneficio directo a las organizaciones. Así el verdadero valor de las bases de datos radica en la posibilidad de extraer información útil para la toma de decisiones o su exploración. Tradicionalmente en la mayoría de los dominios este análisis de datos se hacía mediante un proceso manual o semiautomático: uno o más analistas con conocimiento de los datos y con la ayuda de técnicas estadísticas proporcionaban resúmenes y generaban informes, o validaban modelos sugeridos manualmente por los expertos. Sin embargo, este proceso, en especial la generación de modelos, es difícil conforme aumenta el tamaño de las bases de datos y el número de dimensiones o parámetros se incrementa. Bases de datos con un número de registros del orden de $10^9$ y $10^3$ de dimensión son un fenómeno relativamente común y sólo la tecnología informática puede automatizar el proceso [12].

Por todo lo anterior, surge la necesidad de metodologías para el análisis inteligente de datos, que permitan descubrir un conocimiento útil a partir de los datos. Este es el concepto de proceso de KDD (*Knowledge Discovery in Databases*). KDD puede ser definido como el proceso no trivial de identificar patrones en los datos con las características siguientes: válidos, novedosos, útiles y comprensibles. El proceso de KDD es un conjunto de pasos interactivos e iterativos, entre los que se incluye el pre-procesamiento de los datos para corregir los posibles datos erróneos, incompletos o inconsistentes, la reducción del número de registros o características encontrando los más representativos, la búsqueda de patrones de interés con una representación particular y la interpretación de estos patrones incluso de una forma visual [32].

El paso más importante de este proceso es conocido como Minería de Datos (*data mining)*, cuyo objetivo general de predecir resultados o descubrir relaciones entre los datos. Data mining puede ser descriptivo, por ejemplo, descubrir patrones que describen los datos, o predictivo, para pronosticar el comportamiento del modelo basado en los datos disponibles [33], [34].

Un concepto primordial y diferenciador de las técnicas estadísticas más clásicas, es el de Aprendizaje Automático (*Machine Learning*), que fue concebido hace aproximadamente cuatro décadas con el objetivo de desarrollar técnicas y/o métodos computacionales capaces de generalizar comportamientos a partir de información no estructurada y suministrada en forma de ejemplos. En particular, mecanismos capaces de inducir conocimiento a partir de datos, es decir, un proceso de inducción del conocimiento.

Ya que el desarrollo de software ha llegado a ser uno de los principales cuellos de botella de la tecnología informática de hoy, la idea de inducir conocimiento por medio de ejemplos parece particularmente atractiva en problemas de clasificación supervisada. Tal forma de inducción de conocimiento es deseable en problemas que carecen de solución algorítmica eficiente, son definidos, o imprecisamente definidos. Ejemplos de tales problemas pueden ser de diagnóstico médico, el reconocimiento de patrones visuales o la detección de regularidades en enormes cantidades de datos [153].

Los algoritmos de aprendizaje automático pueden clasificarse en dos grandes categorías: i) métodos cuyo modelo es implícito, desarrollando su propia representación del conocimiento, que no es visible desde el exterior tales como redes neuronales o los métodos Bayesianos, y ii) métodos explícitos, construyen una estructura simbólica del conocimiento que intenta ser útil desde el punto de vista de la funcionalidad, pero también descriptiva desde la perspectiva de la inteligibilidad [58], tales como los que generan árboles de decisión, reglas de asociación, o reglas de decisión.

Como es natural, las áreas del aprendizaje automático y la minería de datos se intersecan en gran medida, en cuanto a los problemas que tratan y a los algoritmos que utilizan. No obstante, la minería de datos tiene un mayor alcance en la adquisición de conocimiento a partir de grandes cantidades de datos, mientras que el aprendizaje automático se orienta más a la tarea de procesos capaces de inducir conocimiento a partir de datos, es decir, un proceso de inducción del conocimiento, buscando en algunos casos estrategias o heurísticas, más que el propio conocimiento. Por esa razón, la minería de datos tiene un espectro de aplicación más amplio, en el sentido de que interactúa mejor con diferentes dominios, pues el aprendizaje adquirido se transforma en conocimiento útil para el experto en el dominio concreto.

Los campos de investigación envueltos en un proceso de KDD son muy variados: desde bases de datos, estadística e inteligencia artificial, aprendizaje automático, lógica difusa, reconocimiento de patrones, evaluación, interpretación e incluso su visualización. Así, el KDD es una metodología adecuada para el descubrimiento y reconocimiento de patrones en las bases de datos utilizando técnicas y métodos de aprendizaje automático y minería de datos. Los investigadores de KDD incorporan técnicas, algoritmos y métodos de estos campos. Así un proceso KDD engloba todos estos campos y principalmente centra su atención en el proceso completo de extraer conocimiento de grandes volúmenes de datos incluyendo el almacenamiento y acceso, escalar el algoritmo cuando sea necesario, esto es, comenzar con una solución sub-óptima del problema y, repetidamente, mejorar la misma hasta que cierta condición sea maximizada, interpretando y visualizando los patrones y resultados. Típicamente, incluye la interacción hombre-máquina. Así, lo anteriormente expuesto constituye una relación entre el proceso KDD, Aprendizaje Automático y el Reconocimiento de Patrones.

## 2.2 Enfoques en el Reconocimiento de Patrones:

Se entiende por problemas de Reconocimiento de Patrones a todos aquellos relacionados con la clasificación de objetos y fenómenos y con la determinación de los factores que inciden en los mismos. El Reconocimiento de Patrones es una disciplina que aborda principalmente cuatro familias de problemas, a saber [78], [106]:

1. Selección de rasgos o características.
2. Clasificación con aprendizaje (supervisado) y donde el diagnóstico y pronóstico pueden modelarse como una clasificación supervisada.
3. Clasificación sin aprendizaje (no supervisado).
4. Clasificación con aprendizaje parcial (parcialmente supervisado).

A continuación se describirán brevemente los enfoques más populares en esta disciplina.

### 2.2.1 Enfoque Estadístico

Históricamente, una de las primeras herramientas empleadas en la solución de problemas de Reconocimiento de Patrones es la Estadística; utiliza el Análisis Discriminante, la Teoría Bayesiana de la Decisión, la Teoría de la Probabilidad y el Análisis de Agrupamientos (Cúmulos, cluster). El enfoque estadístico es la más simple y consiste en representar cada *patrón* mediante un *vector de números* resultantes del muestreo y cuantificación de las señales externas, y cada *clase* por uno o varios patrones prototipo. Un patrón no es más que un punto del *espacio de representación de los patrones*, que es un espacio de dimensionalidad determinada por el número de variables consideradas [13].

El estudio del conjunto apropiado de variables, la variabilidad de los patrones de una clase, las medidas de semejanza entre patrones, así como la relación entre patrones y clases constituye *el Reconocimiento Estadístico de Patrones* cuyas principales características son:

- Se basa en descripciones de objetos en términos de mediciones, es decir, variables numéricas.
- A dichas variables se le presuponen propiedades tales como las de estar definidas sobre un espacio métrico o normado, e incluso en ocasiones se asume un tipo particular de métrica.
- Es muy frecuente el uso de probabilidades, en particular cuando se considera la presencia de elementos de incertidumbre o subjetividad; pero también en estos casos es frecuente el asumir un determinado comportamiento de dichas probabilidades y con ello aparece la suposición de ajustarse a distribuciones normales.

Este enfoque ha sido aplicado en muchos problemas concretos, en particular los relacionados con imágenes y señales; sin embargo, su uso se ha

extendido indebidamente, a zonas para las cuales no fueron concebidas, en problemas donde las hipótesis que se presuponen no se cumplen. Es importante hacer notar que las herramientas matemáticas deben ser seleccionadas muy cuidadosamente para su área de aplicación. No se debería de usar una herramienta en la solución de un problema para el cual no fue diseñado [40, 45, 46].

### 2.2.2 Enfoque Sintáctico Estructural

Otro de los enfoques importantes del Reconocimiento de Patrones es el que parte de la Teoría de los Lenguajes Formales. Su origen está relacionado con el reconocimiento de imágenes y señales. Su idea central consiste en suponer que estos objetos, una señal electrocardiográfica digamos por caso, se puede descomponer (físicamente) en elementos primarios, atómicos, (en pedazos de la misma) como si fueran las letras de un cierto alfabeto; y a partir de estas letras, teniendo en cuenta la señal completa, encontrar las reglas gramaticales que nos permitan formar la señal (como si estuviéramos armando un rompecabezas). En otras palabras, el propósito es encontrar la gramática cuyo lenguaje estaría formado sólo por señales que estarían muy estrechamente vinculadas unas con las otras y aquellas señales que no tuviesen que ver con las primeras, responderían a gramáticas diferentes, por lo que pertenecerían a otro lenguaje. Algunas de las características de este enfoque, denominado Reconocimiento Sintáctico Estructural de Patrones [82], son las siguientes [78-79]:

- Se basa en las descripciones de los objetos en términos de sus partes constitutivas.
- Se apoya en la Teoría de los Lenguajes Formales, la Teoría de Autómatas, las Funciones Recursivas y la Teoría de Grafos.
- Se asume que la estructura de los objetos a ser reconocidos es cuantificable.

En forma muy general, podemos decir que en este enfoque se asocia a cada conjunto de objetos una gramática que *genera* sólo elementos de dicho conjunto, y el problema consiste en averiguar cuál de las gramáticas genera como palabra la correspondiente al objeto que se desea clasificar; o también que a cada conjunto de objetos se le asocia un grafo que describe las relaciones entre las propiedades estructurales de un objeto representante del conjunto de objetos. Aquí se compararían los grafos asociados a cada representante de las clases con el objeto que se quiere clasificar.

Esta manera de abordar un problema de Reconocimiento de Patrones es especialmente productiva cuando los objetos de estudio son objetos físicos, es decir, imágenes o señales. Ejemplos de estas aplicaciones son trabajos en identificación de impresiones digitales [83], [88], entre muchos otros.

### 2.2.3 Enfoque Lógico-Combinatorio

La Lógica Matemática, la Teoría de Testores, la Teoría Clásica de Conjuntos, la Teoría de los Subconjuntos Difusos, la Teoría Combinatoria, la Matemática Discreta en general, constituyen el basamento teórico-matemático en el que se desarrolla el denominado Enfoque Lógico-Combinatorio en Reconocimiento de Patrones. Las ideas centrales de este enfoque consisten en suponer que los objetos se describen por medio de una combinación de rasgos numéricos y no numéricos, y los distintos valores pueden ser procesados por funciones numéricas [106-109].

Este enfoque se basa en la idea de que la modelación del problema debe ser lo más cercana posible a la realidad del mismo, sin hacer suposiciones que no estén fundamentadas. Uno de los aspectos esenciales del enfoque es que las características utilizadas para describir a los objetos de estudio deben ser tratadas adecuadamente [78-79].

El enfoque lógico combinatorio es más que un conjunto de técnicas, es una filosofía, una manera de enfrentar los problemas de Reconocimiento de Patrones a partir de una determinada metodología de la modelación matemática, es decir, como deben ser modelados y resueltos los problemas reales.

El enfoque lógico combinatorio; además aborda problemas de selección de variables (determinación de síndromes de enfermedades, determinación de la relevancia de síntomas, signos de enfermedades, o del estado de una red de computadoras, etc.) y de clasificación supervisada (con aprendizaje: diagnóstico y pronóstico médicos; pronóstico de fenómenos naturales o sociales; pronóstico de perspectividad de recursos minerales, etc.) a partir del enfoque lógico combinatorio en los llamados dominios poco estructurados.

## 2.3 CONTEXTO DEL PROBLEMA SEGÚN EL RECONOCIMIENTO DE PATRONES (RP)

La presente propuesta de tesis se ubica dentro del contexto de la disciplina de Reconocimiento de Patrones como un sistema híbrido en el que, partiendo de un conjunto duro y no vacío de objetos, descritos en términos de variables llamadas atributos o rasgos, en el cual cada rasgo descriptivo tiene un dominio de definición, una relación de descripción que a cada objeto le asigna una descripción en términos de los rasgos, y donde cada objeto tiene asignada una clase o categoría con una cierta relación de pertenencia a cada clase y tomando como referencia la metodología del enfoque lógico combinatorio abordamos el problema de caracterización e interpretación de descripciones conceptuales en dominios poco estructurados, utilizando elementos estadísticos, inteligencia artificial y lógica difusa, obteniendo un modelo y su implementación en un sistema automatizado denominado CIADEC [50-56]. En especial, se considera el papel del experto en el imprescindible grupo

interdisciplinario para dar solución al caso de estudio que se trata en este trabajo de tesis, el cual se refiere al tratamiento de aguas residuales.

En sus inicios, este tema de tesis estuvo relacionado con el Proyecto Marco que se desarrolla en el Departamento de Lenguajes y Sistemas de la Universidad Politécnica de Cataluña, y actualmente se trabaja de forma paralela. La línea de investigación se inició en 1995 con el objetivo principal de estudiar los dominios poco estructurados [10], [16], [20], [24-25], [29], [31], [36-48].

El objetivo del proyecto marco es construir una plataforma integrada de soporte al análisis inteligente de *espacios de representación*, incluyendo todo tipo de herramientas, desde las más básicas de análisis descriptivo, hasta las más sofisticadas como *la clasificación basada en reglas* y herramientas de apoyo a la *interpretación de resultados*; estas últimas relacionadas con la minería de datos y el proceso Knowledge Discovery in Databases (KDD) [31-34].

La primera propuesta constituye la tesina [24] y después la tesis doctoral de Gibert [10], que se cristalizó en la formulación de la metodología de clasificación basada en reglas y una primera versión del sistema informático que la implementa, denominado Klass [25-26], el cual se ha utilizado en diversas aplicaciones [16], [22-23], [27-30].

Considerando las características especiales de este espacio de representación, se han desarrollado métodos mixtos de análisis que combinan técnicas estadísticas con técnicas de inteligencia artificial para resolver los problemas que se plantean en este contexto [35, 36].

Todo el software que se ha desarrollado en el seno del proyecto marco ha integrado lo que podemos llamar *herramienta master*, que actualmente es el *joc.Klass+*, y que aglutina herramientas de muy distinta naturaleza ofreciendo la interfaz necesaria en cada momento del análisis [26]. Esta herramienta informática ha venido evolucionando de forma continua desde su origen, en la medida en que se ha avanzado en la investigación y experimentación de la línea de investigación antes mencionada.

Así, el presente trabajo de tesis se ha desarrollado de forma paralela al proyecto marco. Además, el trabajo de tesis tiene su contexto en los enfoques Probabilístico-Estadístico y Lógico-Combinatorio; estando inmersa en el programa de Doctorado en Ciencias de la Computación del Centro de Investigación en Computación del Instituto Politécnico Nacional [151-152], [154-158].

## 2.4 MODELADO MATEMÁTICO DE RECONOCIMIENTO DE PATRONES

La aplicación de los modelos matemáticos de Reconocimiento de Patrones y las técnicas de computación a problemas de ciencias tales como la Medicina, las ciencias Biológicas, las Geociencias, las Ciencias Sociales y

otras, es hoy en día un hecho cotidiano. En estas disciplinas se presentan problemas de clasificación, de diagnóstico, de pronóstico, de determinación de factores de influencias y otros que son claramente problemas a los que estamos aludiendo. Sin embargo, existen serias limitaciones en su plena explotación debido a problemas metodológicos en el proceso de modelación matemática, en la elaboración de los sistemas automatizados que realizan dichos modelos y que son en definitiva quienes plasman las respuestas a los problemas planteados y en la explotación de dichos sistemas [79],[106].

En las disciplinas científicas que en forma genérica se denominan "ciencias poco formalizadas" [106-108], se da un conjunto de características que hacen que el problema de la modelación matemática se torne complejo [76], [78]. Entre estas características pueden mencionarse:

1. Desconocimiento de *expresiones analíticas* de los vínculos o leyes que rigen el comportamiento de los objetos en cuestión.
2. *Información incompleta* sobre cuáles son los factores que influyen en los fenómenos que se estudian.
3. De algunos factores identificados se desconoce su *grado de influencia* en el fenómeno.
4. Se usan múltiples criterios de solución y por lo general se tienen *soluciones múltiples* para un mismo problema.
5. Los especialistas no siempre tienen una explicación única de las conclusiones a las que se arriban y por lo general, se trabaja sobre la base de *analogías* (experiencias precedentes y observaciones acumuladas).
6. En la descripción de objetos están presentes *variables cuantitativas, cualitativas* y generalmente ambas; también suelen aparecer objetos de los que no se poseen todos los datos (ausencia de información en algunas de sus características descriptivas); además, los razonamientos no necesariamente se modelan con la Lógica Matemática Clásica.
7. Existe una barrera de incomprensión entre los especialistas del área y los modeladores formales (matemáticos e informáticos, entre otros) porque ocurre alguna de las situaciones siguientes:

    A. Un modelador le pide al especialista que le diga cuál es su problema matemático y generalmente los especialistas no tienen una preparación matemática suficiente, por lo tanto que no hay comunicación.
    B. Un especialista le lleva a un modelador un conjunto de datos para que se los procese con un software o con alguna técnica que haya ganado popularidad en su área. Incluso, hay quien "experimenta" con sus datos "para ver qué resultados se obtienen". Esto se hace sin tomar en cuenta los requerimientos formales de la herramienta en uso [78-79], [106] y [109].

Es muy común en los especialistas no-matemáticos, la prisa de encontrar respuestas a sus problemas; en especial, haciendo uso de la microcomputadora

personal. Sin embargo, no resulta nada fácil hacer la modelación de un fenómeno de una zona del conocimiento en otra y mucho menos si ésta es una ciencia poco formalizada. Se requiere en primer lugar del dominio profundo de la teoría en la que se va a modelar el fenómeno; en segundo lugar, no menos imprescindible que el anterior, del conocimiento del propio fenómeno en el contexto de su zona de especialidad –no se puede modelar lo que no se conoce-. Ante esta situación se adoptan actitudes diferentes, tanto por matemáticos como por los especialistas no-matemáticos [76], [78], [109].

Todo esto condiciona un hecho en el área de aplicación de los modelos matemáticos y las técnicas de computación se restringe, en general, a la de aquellos problemas cuyas soluciones se basan en la observación de un conjunto predeterminado de indicadores cuya presencia conlleva, de manera prácticamente unívoca, a una sola conclusión –o a un conjunto fijo de conclusiones. Tanto para los especialistas no-matemáticos, llamémoslos de orientación práctica, para los cuales la Matemática y las computadoras son herramientas que sólo le ayudan a aumentar la velocidad y disminuir el costo de su acción resolutiva; como para los de orientación teórica –que son muy pocos- para los cuales estos son medios para aumentar la fundamentación y confiabilidad de dicha acción resolutiva, el universo de problemas donde pueden incursionar realmente los modelos matemáticos y las técnicas de computación es el de los problemas con soluciones preconcebidas. Es decir, son respuestas esperadas [162-167].

Lo que continua está basada fuertemente en las referencias [78] y [79].

El objetivo es el de plantear la metodología para el proceso de modelación matemática de problemas de Reconocimiento de Patrones [78] consta de siete etapas, como se muestra en el esquema de la figura 2.1 [79], y son:

1. Formulación del problema inicial A (cuya solución es R).
2. Recolección de información.
3. Formalización lógica-matemática del problema A en A'. Selección del modo de solución del problema A'.
4. Solución del problema matemático (R').
5. Interpretación y validación de los resultados respecto al problema A.
6. Pruebas de campo

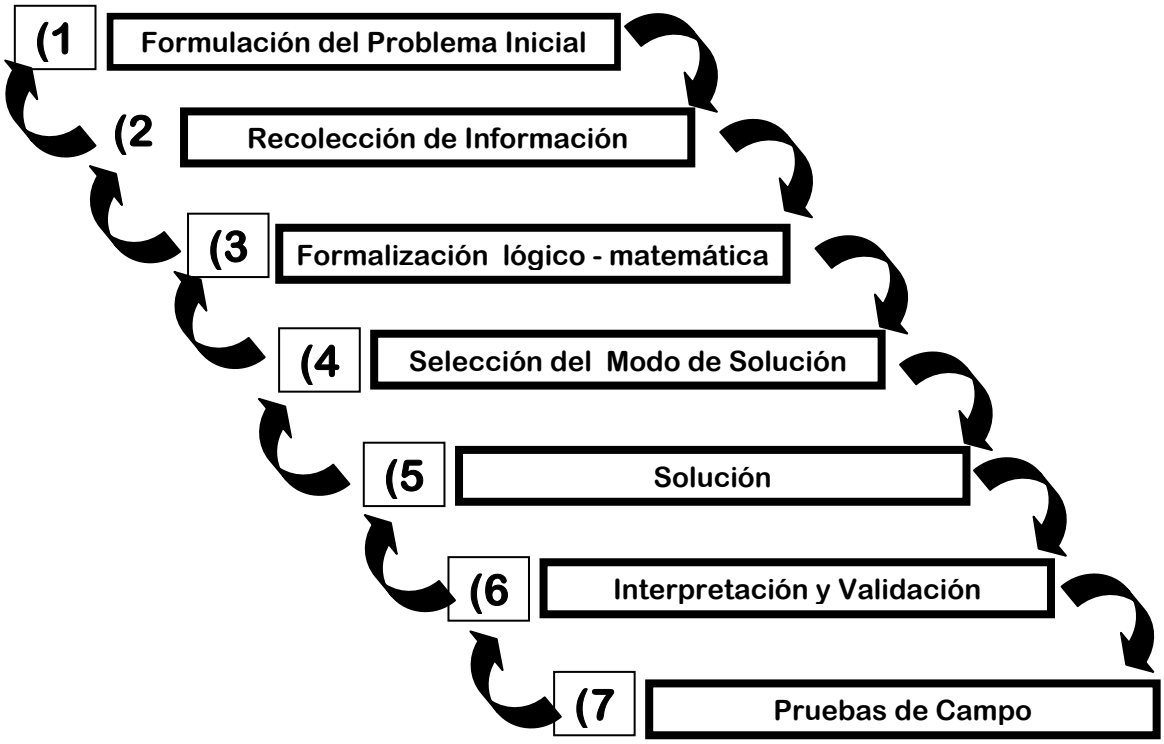



Figura 2.1 Esquema global de la modelación matemática.

Las etapas anteriores contienen un conjunto de acciones, que según la experiencia práctica, resultan fundamentales en el proceso y las cuales se describen a continuación:

**1. Formulación del problema inicial A**. En esta etapa, el especialista tiene una mayor participación porque es quien expresa en su lenguaje el problema a resolver, determinando:

A. El **objetivo** de la investigación.
B. Los **objetos** de la investigación.
C. Las **propiedades** que caracterizan a los objetos.
D. Las **características** de dichas propiedades.
E. Las **relaciones** entre los objetos y sus propiedades.
F. Las **hipótesis** en que se fundamenta el trabajo a realizar.
G. Las **fuentes** de información.
H. Qué información es **relevante**, si esto se conoce.
I. Cómo se **recolecta** la información.
J. Cómo se **interpreta** y **manipula** la información.
K. Cómo se requiere que se **presenten** los resultados.
L. La identificación de **ruidos** y **distorsiones** de la información.
M. La valoración de los **errores** en la información en su entrada, procesamiento y salida.

Es obvio que en esta etapa, el papel principal lo desempeña el especialista del área de aplicación. Sin embargo, nada tendría sentido si el

papel de los modeladores (matemáticos, ingenieros, informáticos entre otros) es pasivo, de contemplación anodina. Se trata por el contrario de cuestionar, de entender la esencia del fenómeno a explicar, si bien en el lenguaje del especialista del área, pero con la intención de alcanzar un verdadero diálogo, en el que las ideas esenciales subyacentes al problema que investigamos se vean con precisión [106], [108-109].

**2. La Recolección de datos.** La recolección de datos se refiere al uso de una gran diversidad de técnicas y herramientas que pueden ser utilizadas por el modelador para recabar información útil en el problema que se pretende resolver, dicha información se puede obtener a través de las entrevistas, la encuesta, el cuestionario, la observación, el diagrama de flujo y el diccionario de datos.

Todos estos instrumentos se aplicarán en un momento en particular, con la finalidad de buscar información que será útil a una investigación en común. Los analistas modeladores utilizan una variedad de métodos a fin de recopilar los datos sobre una situación existente, como entrevistas, cuestionarios, inspección de registros (revisión en el sitio) y observación. Cada uno tiene ventajas y desventajas. Generalmente, se utilizan dos o tres para complementar el trabajo de cada una y ayudar a asegurar una investigación completa [106-108].

**3. Formalización del problema A.** Esta etapa es posible que mentalmente se lleve a cabo a medida que el especialista formula el problema. Es compleja porque se requiere "traducir" del lenguaje del especialista al lenguaje formal de la Matemática, de tal manera que de la etapa anterior queden reflejados: *objetivos, objetos, propiedades y su escala de medición, características, relaciones entre objetos y entre propiedades, el concepto de clase de objetos, propiedades de las mismas, los conceptos de analogía, la evaluación de los errores, entre otros*. En esta etapa se realizan:

A. La selección del *espacio de representación de los objetos* de investigación;
B. La determinación de las *funciones que modelarán los criterios de comparación* de valores de cada variable, así como entre las descripciones de los objetos.
C. *El análisis* desde el punto de vista formal de los requisitos de la solución que el especialista impone a los resultados,
D. La interpretación que el especialista da a los datos.

Estos son aspectos que contribuyen en la búsqueda de la solución y en la selección de algoritmos óptimos para el problema en cuestión, y determinan en gran medida la forma en que serán elaborados los datos iniciales a partir de su organización en lo que se denomina Matriz de Aprendizaje (MA) o también Tabla de Objeto-Propiedad (TOP) [106], [108] y [109].

**4. Selección del modo de solución del problema (solución del problema** matemático A'). El proceso de formalización muchas veces restringe fuertemente el área de búsqueda de las técnicas de solución. En esta etapa un papel decisivo lo desempeña el análisis de la TOP. En esta etapa se puede reducir la cantidad de información requerida al mismo tiempo que se aumenta su calidad. Se decide el enfoque o combinación de ellos para la solución del problema A', determinando la familia de algoritmos a la que pertenece. La etapa se concluye con la elección del modo de solución que se debe aplicar; y si es el caso, el esquema de procesamiento de la información [106], [108] y [109].

**5. Solución del problema expresado en términos matemáticos** (se obtiene R'). Tomando como base los datos formalizados y el tipo de algoritmo a utilizar, se elabora el sistema computarizado (si lo amerita el caso) y se obtiene la solución R' del problema A'. Se analiza la concordancia del resultado alcanzado R' con los objetivos formalizados del problema matemático A', teniendo como herramienta fundamental la formalización de los criterios para la evaluación de resultados de la segunda etapa [106], [108] y [109].

**6. Análisis e Interpretación de los resultados respecto al problema** (de R' a R luego hacia A). Los resultados de A' (R') se interpretan expresándolos en un lenguaje o en otro, en forma similar a lo que se hizo en su contraparte en la segunda etapa. Después de la correspondencia del resultado R' con el problema A' en la etapa anterior, se hace necesario el análisis entre el resultado R y el problema A. Las acciones resolutivas obtenidas son variadas y dependen de los resultados de dicho análisis [106], [108] y [109].

El especialista del área de aplicación también es el máximo responsable de esta etapa y debe ser ejecutada en conjunto con los elementos del equipo multidisciplinario.

**7. Pruebas de campo.** Es una fase de validación científica que debe cumplir con las condiciones, requisitos y normas establecidas del problema planteado por el especialista. Estas pruebas son un ideal para evaluar los resultados lógicos obtenidos con los resultados de la realidad [106], [108] y [109].

Así, la aplicación de la metodología siempre nos llevará de manera secuencial a la solución definitiva. En ocasiones habrá que regresarse a etapas anteriores para reconsiderar algunas de las decisiones tomadas, a confirmarlas a veces, otras a modificarlas. Cabe mencionar que este proceso, que puede parecerle a algunos engorroso, aburrido, innecesario, ha dado frutos antes de llegar a clasificar, antes de procesar los datos [106], [108] y [109].

## 2.5 MODELADO ESTRUCTURAL DIFUSO

Todas las técnicas tradicionales del Reconocimiento de Patrones no supervisado tienen la desventaja de formar agrupamientos los cuales no tienen una interpretación conceptual. El problema del significado de los agrupamientos

obtenidos es dejado al especialista. Esta desventaja es significativa ya que el especialista, para los fines de su investigación o labor, requiere no sólo los agrupamientos, sino además, quiere una explicación de ellos en términos humanos. El agrupamiento conceptual surge a partir de los trabajos de Michalski [7, 57]. En este enfoque se propone encontrar, a partir de un conjunto de datos, no sólo los agrupamientos en los que éstos se estructuran sino además conformar la *explicación* de tales agrupamientos [110].

El modelo que utilizaremos para la caracterización e interpretación de descripciones conceptuales en dominios poco estructurados consta de los siguientes elementos fundamentales:

**Definición**. Los elementos del modelo estructural difuso son: Un conjunto universal Ω de individuos u objetos relacionados con el planteamiento de un problema y sea O⊆Ω. Una estructura difusa de O es una tupla ***(O, ℜ, δ, Q, π, Cc, f)***, donde se tiene que:

**O** es un conjunto duro de individuos conocidos no vacío de un dominio, esto es:

$$O = \{o_1, \ldots .o_n\}$$

$\mathfrak{R}$ es un conjunto difuso de variables o atributos que describen a cada objeto de **O** y el cual se define como:

$\mathfrak{R} = \{x_1 \backslash \mu_1, \ldots\ldots\ldots, x_r \backslash \mu_r\}$, donde cada variable $\mathbf{x_i}$ tiene un dominio de definición $\mathbf{M_i}$.

Para la descripción de los objetos en **O** requerimos la existencia de un conjunto de llamaremos conjunto descriptor de los elementos de **O** y definido como:

$$D = \{\{m_1 \backslash \eta_1, \ldots\ldots\ldots, m_r \backslash \eta_r\} / m_i \in M_i \wedge \eta_i \in [0, 1], i \in [1,r]\}$$

donde D es un conjunto difuso denominado Conjunto Descriptivo o descripción y definen los objetos que se pueden construir en $\mathfrak{R}$ y sus respectivos dominios.

$\delta$ Es una relación funcional definida en **OXD**, llamada *Relación Descriptiva*, en donde a cada objeto se la asigna una descripción en términos de las variables o atributos en $\mathfrak{R}$.

**Q** Es un conjunto duro, definido como: $Q = \{C_1, \ldots\ldots\ldots C_k\}$ donde cada $C_i$ se denomina Clase o Categoría en las cuales se agrupan los elementos de O. Cada clase queda determinada por:

$$Cc_{i=} \{ o_j \backslash \mu_{Ci}(o_j) / oj \in O \wedge \mu_{Ci}(o_j) = \pi(\delta(o_j), C_i\}$$

donde $Cc_i$ es el conjunto de todos los objetos asociados con su grado de pertenencia a dicha clase.

$\pi$ Es una relación funcional definida sobre $\delta$**(OXQ) X [0, 1]**, llamada Función de Pertenencia tal que a cada descripción de un objeto le asigna uno y sólo un grado de pertenencia normalizado a cada una de las clases.

**Cc** Es un conjunto duro definido como: Cc = {$\varphi$1,……..,$\varphi$r} de funciones llamadas Criterios de Comparación. Cada criterio puede comparar los valores de descripción de los diferentes objetos en una misma variable y está definido como:

$$\varphi_k \subseteq (M_k \; X \; M_k) \; x \; \Delta_k, \text{ para algún } \Delta_k \text{ totalmente ordenado.}$$

**f** Es una relación funcional llamada *Función de Analogía entre Patrones* definida sobre $f \subseteq (\delta(O) \; X \; \delta(O)) \; X \; Y$, para algún conjunto Y totalmente ordenado.

Resulta claro que todo algoritmo de clasificación incluye una función de analogía entre patrones que usa para clasificar [78]. No obstante, desde el punto de vista del modelo, no resulta atractivo incluir como parte de la definición de algoritmo una relación definida sobre elementos que no son parte del mismo, a decir, (***O*** y $\delta$). La misma consideración puede hacerse para incluir estos elementos y no incluirlos como parte de alguna de las entidades mencionadas.

Así, a partir del modelo podemos definir dos matrices para especificar cada uno de los elementos de la base de datos del dominio en estudio. Estas dos matrices se denominan ***Matriz de Descripción*** **($M_\delta$)** y **Matriz de Pertenencia ($M_\pi$)**, las cuales constituyen las expresiones de las dos relaciones fundamentales en la estructura de caracterización***:*** $\delta$ ***y*** $\pi$. O bien, podemos expresar estas dos matrices en una sola tabla, como se muestra en la Tabla 2.1.

| $M_\delta$, $M_\pi$ | $x_1$ | …. | ….. | $x_r$ | $C_1$ | ….. | ……. | $C_k$ |
|---|---|---|---|---|---|---|---|---|
| $O_1$ | $\delta(o_1)\|x_1$ | | | $\delta(o_1)\|x_r$ | $\pi(o_1,C_1)$ | …. | …. | $\pi(o_1,C_k)$ |
| ….. | …. | …. | …. | | ….. | ….. | ….. | ….. |
| …. | …. | | | | | | | …. |
| …. | …. | | | | | | | …. |
| …. | …. | | | | | | | …. |
| $o_n$ | $\delta(o_n)\|x_1$ | ….. | …. | $\delta(o_n)\|x_r$ | $\pi(o_n,C_1)$ | ….. | ….. | $\pi(o_n,C_k)$ |

Tabla 2.1 Estructura de las matrices $M_\delta$ y $M_\pi$

En las últimas dos décadas han sido propuestos diferentes algoritmos de agrupamiento conceptual [7], [57], [75], [168], [169], [170], [171], [172], [173], [174], [175].

### 2.5.1 Formulación del Modelo Difuso para el Problema de Caracterización Conceptual

La formulación del modelo difuso para el problema de la caracterización conceptual de espacios se establece [107, 108] en los siguientes términos: Sea M un conjunto de objetos. Una descripción I(0) es definida para cada objeto $O \in M$ y ésta es representada por una secuencia finita $x_1(O), x_2(O),....,x_m(O)$ de valores de m variables del conjunto $R = \{ x_1, x_2,..., x_m \}$, con $x_i(O) \in M_i$, siendo $M_i$ el conjunto de valores admisibles de la variable $x_i$ [162-167].

Además, asumiremos que en $M_i$ hay un símbolo * el cual denota ausencia de información, i=1,. ..,m. En otras palabras, la descripción de un objeto puede estar incompleta; esto es, para al menos una variable no se conoce el valor. Consideraremos que $I(O) \in M_1xM_2x,...,xM_m$ (este producto cartesiano es el Espacio de Representación Inicial ERI de los objetos). La naturaleza de las variables o atributos por consecuencia puede ser, simultáneamente, cualquiera (cualitativa: Booleana, multi-valuada, difusa, lingüística y otras o cuantitativas: enteras, reales). Por lo tanto, sobre el ERI no supondremos ninguna estructura algebraica o topológica.

Consideremos una función $C_i$: $M_i$ x $M_i \rightarrow L_i$ tal que:

a) $C_i(x_i(O),x_i(O')) = \min_{o' \in M} \{C_i(x_i(O),x_i(O'))\}$ si $C_i$ es un criterio de comparación de disimilaridad entre valores de la variable $x_i$ o

b) $C_i(x_i(O),x_i(O')) = \max_{o' \in M} \{C_i(x_i(O),x_i(O'))\}$ si $C_i$ es un criterio de comparación de similaridad entre cualquiera dos valores de la variable $x_i$ para i = 1,...,m; $C_i$ es una evaluación del grado de similaridad (o disimilaridad) entre cualquiera dos valores de la variable $x_i$ donde $L_i$ es un conjunto totalmente ordenado, i=1….,m.

Usualmente, la información acerca de los objetos (sus descripciones) está dada en forma de una tabla o matriz $MI= |x_i(O_j)|_{nxm}$ con n renglones (descripciones de objetos) y m columnas (valores de cada variable en los objetos seleccionados).

El problema de la estructuración conceptual de espacios sobre M consiste en determinar el conjunto cubrimiento $\{K_1, K_2,....,K_c\}$ c > l, así como el conjunto de conceptos asociados a cada $K_i$ con i = 1,...,c., en principio $K_i$; podrían ser subconjuntos duros o difusos de M y estos podrían ser disjuntos o no.

Sea una función $\Gamma(\mathrm{M1x}….\mathrm{xMm}\} \rightarrow L,$ donde L es un conjunto de resultados totalmente ordenado; $\Gamma$ será denominada *función de similaridad* y es una evaluación del grado de similitud entre dos descripciones cualesquiera pertenecientes a MI [106], [164], [165], [167]. Aunque tradicionalmente, en clustering, se ocupa más la diferencia entre patrones que la semejanza

### 2.5.2 Modelo Difuso de Algoritmos de Agrupamiento Conceptual

En la práctica profesional de especialistas en ciencias poco formalizadas (soft-sciences), frecuentemente aparece el problema de la clasificación no supervisada conceptual en el que resulta de mucho interés saber no sólo qué objetos pertenecen a los agrupamientos, sino además en qué medida o grado los objetos pertenecen a los agrupamientos encontrados o en qué medida cumplen la propiedad de cada agrupamiento [116]. Dos objetivos principales en este sentido son resueltos con el modelo de estructuración difuso:

a) obtener una estructuración extensional difusa –un conjunto de agrupamientos difusos;

b) obtener una estructuración intencional; es decir, obtener una caracterización conceptual de los agrupamientos difusos en términos de conjuntos de características apropiadas.

En lo que concierne al primer objetivo, se obtiene una estructuración del espacio de representación de los objetos (duros y difusos). De aquí que los dos métodos de agrupamiento conceptual antes mencionados quedan como caso particular del modelo de objetos simbólicos [106], [164], [165], [167].

Los objetos no necesariamente están descritos por variables que tomen un solo valor sino un conjunto de valores. Estos objetos dan la posibilidad de introducir en su definición, información más compleja como probabilidades, posibilidades y creencias. Además, los objetos simbólicos permiten describir a los objetos de manera intensional, dando flexibilidad para expresar variación en los valores que toman las variables ([$color = [rojo, blanco]$]) y también se pueden expresar restricciones semánticas (a través de reglas predicados de primer orden) entre los valores de la variable [164], [165], [167].

Existen cuatro tipos de análisis de datos dependiendo de la entrada y salida de los datos: a) análisis numérico de datos clásicos; b) análisis numérico de objetos simbólicos (por ejemplo, calculando propiedades estadísticas de las determinaciones extensionales de los objetos simbólicos); c) análisis simbólico de datos clásicos; por ejemplo, el que se realiza con agrupamiento conceptual; y d) análisis simbólico de objetos simbólicos. La nueva formulación de objeto simbólico da pie a los cuatro tipos de análisis antes mencionados, lo cual hace que el análisis clásico de datos quede como caso particular de la nueva Teoría de Objetos Simbólicos [106], [164], [165], [167].

## 2.6 ELEMENTOS ESTADÍSTICOS DEL MODELO MATEMÁTICO

Los elementos estadísticos que se consideran en el desarrollo de este trabajo son: conceptos básicos de la caracterización e interpretación de clases a partir de atributos cualitativos [10]. Además, se utilizaron técnicas sencillas de Estadística Descriptiva para describir los datos y obtener información preliminar acerca de ellos; los histogramas pueden usarse para hacer representaciones sobre la variabilidad de las mediciones, los conceptos básicos sobre el boxplot

múltiple para observar la relación entre variables y las gráficas de las series de tiempo para observar la evolución a través del tiempo [57].

### 2.6.1 Caracterización de Clases a partir de Variables Categóricas

Partiendo de los trabajos previos sobre la interpretación a partir de variables cualitativas [10], en donde se analizó la caracterización de clases a partir de conceptos fundamentales como: conjunto de valores propios (caracterizador, parcialmente caracterizador, no propio y genérico), variable caracterizadora (v.c.), variable ε-caracterizadora, sistema caracterizador, se demostró, en [50], que el *boxplot múltiple* es una herramienta ágil y potente con variables numéricas para identificar elementos útiles. De hecho, es uno de los conceptos básicos en los que se inspiró esta propuesta metodológica para la detección de las variables caracterizadoras en variables cuantitativas que se propone este proyecto de tesis [10-11], [24-25].

### 2.6.2 El Boxplot: Base del Proceso de Discretización de Variables Numéricas

El boxplot es una herramienta de la estadística descriptiva inventada por Jhon Turkey [172], la cual es muy útil para representar gráficamente grupos de datos numéricos a través del uso de cinco medidas descriptivas de los mismos, a saber: *primer cuartil, tercer cuartil, valor mínimo, valor máximo y mediana*, las cuales pueden trabajarse en forma individual.

Así tenemos algunos conceptos básicos relacionados con el boxplot, a decir:

**Estadísticos:** Son valores representativos que proporcionan información sobre la serie en cuanto a su posición en la escala de medición, agrupamiento en torno a un valor, distribución de los datos y concentración en una región entre otros. Los estadísticos proveen información sobre una muestra. Cuando se trabaja con toda la información (población) se le denomina *parámetro*.

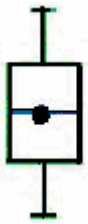
**Cuartiles**: Son valores que dividen a la distribución en cuatro partes iguales en cuanto a la cantidad de datos. Así, tenemos que el primer cuartil ($Q_1$), es el valor por debajo del cual ocurre el 25% de las observaciones y el tercer cuartil ($Q_3$) es aquel por debajo del cual ocurre el 75% de las observaciones. Siguiendo en esta línea, el segundo cuartil ($Q_2$) coincide con la mediana de la distribución.

**Dispersión:** Indica la variabilidad del conjunto de datos: cómo se distribuyen los datos de estudio. Una dispersión grande indica un conjunto de datos heterogéneos e implica poca utilidad de una medida de tendencia central únicamente para describir la distribución.

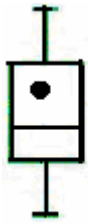
**Simetría:** Indica la forma del conjunto de datos, lo cual implica observar dónde se concentra la información. Para el estudio de la forma de una

distribución, también se usan los términos *sesgo* o *asimetría*. Una distribución puede ser:

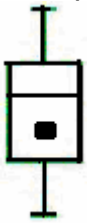
a) ***Simétrica:*** en este tipo de distribuciones la media, la moda y la mediana coinciden y los datos se distribuyen de igual forma a ambos lados de estas medidas. En el contexto, hay igual número de opiniones por encima que por debajo de la mediana.



b) ***Asimétrica positiva o sesgada a la derecha:*** los datos tienden a concentrarse hacia la parte inferior de la distribución y se extienden más hacia la derecha. La media suele ser mayor que la mediana en estos casos. En el contexto, las opiniones se concentran en un puntaje menor y las de mayor puntaje están más dispersas.



c) ***Asimétrica negativa o sesgada a la izquierda:*** los datos tienden a concentrarse hacia la parte superior de la distribución y se extienden más hacia la izquierda. La media suele ser menor que la mediana en estos casos. En el contexto, las opiniones se concentran en un puntaje mayor y las de menor puntaje están más dispersas.



d) **Medida de tendencia central:** Estadístico que procura aportar información sobre la localización central de la distribución de datos. Son: la media aritmética, la moda, la mediana, la media geométrica y la media armónica, y se emplean de acuerdo al objetivo del estudio y al tipo de dato que se tenga.

Y las medidas descriptivas que identifican las partes de un boxplot, figura 2.2 como:

1. **Valor Máximo:** Es el valor extremo superior de la distribución de datos. Los valores por encima de encima de este límite se consideran también atípicos.

2. **Tercer Cuartil:** Es aquel por debajo del cual ocurre el 75% de las observaciones o datos.
3. **Mediana:** Coincide con el segundo cuartil. Divide a la distribución en dos partes iguales y se representa por un segmento horizontal. De este modo, 50% de las observaciones están por debajo de la mediana y 50% está por encima.
4. **Primer Cuartil:** Es el valor por debajo del cual ocurre el 25% de las observaciones o datos.
5. **Valor Mínimo:** Es el valor extremo inferior de la distribución de datos. Por debajo de este valor se encuentran los valores atípicos.
6. **Valores Atípicos:** Son valores que están apartados del cuerpo principal de la distribución de datos. Pueden representar efectos de causas extrañas, opiniones extremas o en el caso de la tabulación manual, errores de medición o registro. Se colocan en la gráfica con asteriscos (*) o puntos (.) según se alejan menos o más del conjunto de datos. Se utiliza un superíndice numérico para indicar el número de veces que aparece ese dato como atípico.
7. **Media Aritmética:** Es lo que tradicionalmente se conoce como promedio. Originalmente no forma parte del boxplot, sin embargo, se considera su inclusión para dar una idea del puntaje general de los datos estudiados.

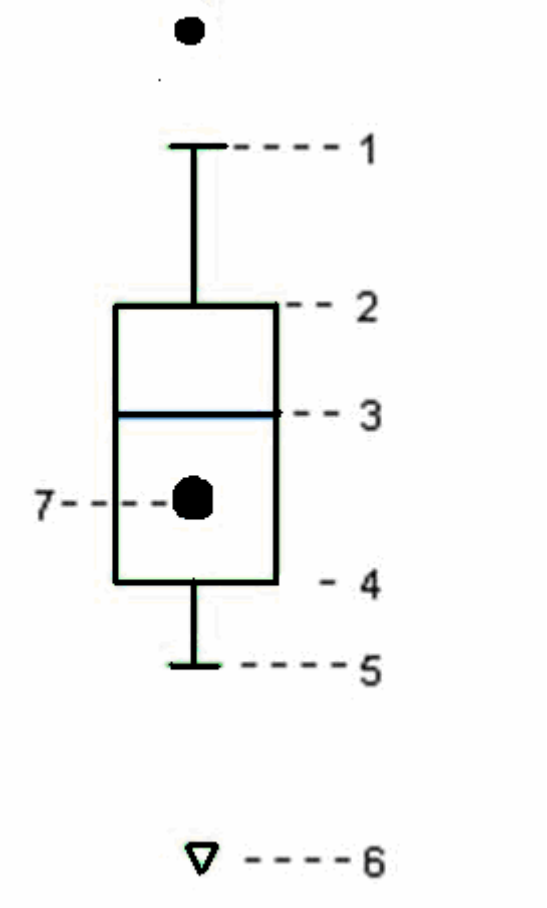



Fig. 2.2 Medidas descriptivas que identifican las partes de un boxplot

En general, los diagramas de cajas o boxplot resultan más apropiados para representar variables que presenten una gran desviación de la distribución normal y cuya representación gráfica muestra la relación entre una variable numérica y algunos grupos o clases. Para cada clase o grupo, se visualiza el intervalo de valores que toma la variable y las observaciones atípicas (outliers) se marcan con "*". Para cada clase, se despliega una caja de Q1 (primer cuartil) a Q3 (tercer cuartil) que representa el 50% de los valores de esa clase, y a partir de ésta se marcan los “bigotes” con sus extremos el mínimo y el máximo que representan cada uno el 25% de los valores de esa clase, y la mediana se marca con una línea horizontal [50-52], figura 2.3.

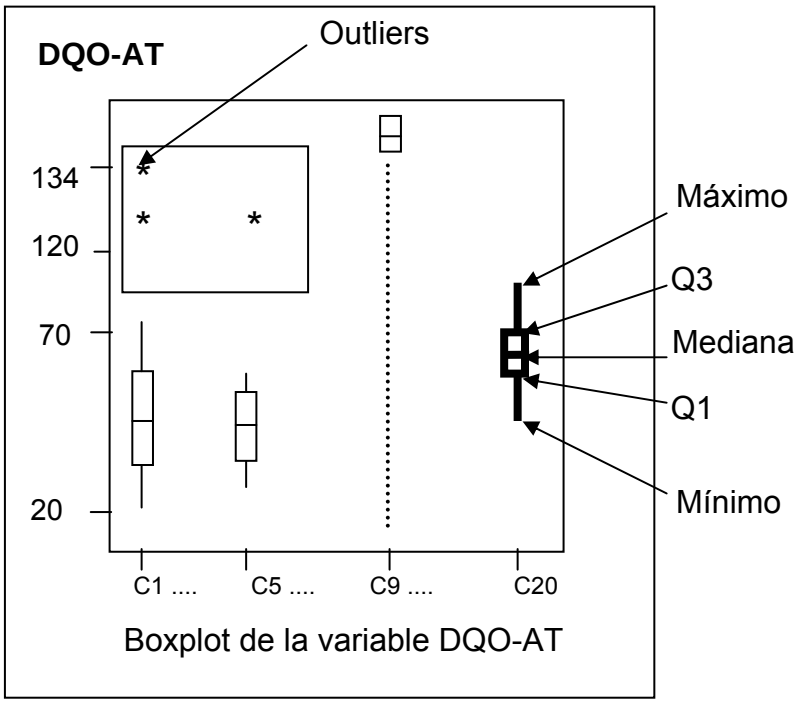



Figura 2.3 Boxplot de la variable materia orgánica química DQO-AT

Así, el boxplot muestra información visual, asociando las cinco medidas descriptivas y sobre la tendencia central, dispersión y simetría de los datos de estudio. Además, permite identificar con claridad y de forma individual, observaciones o datos que se alejan de manera poco usual del resto de los datos. A estas observaciones se les conoce como *valores atípicos*.

Por su facilidad de construcción e interpretación, permite también comparar a la vez varios grupos de datos sin perder información ni saturarse de ella. Esto ha sido particularmente importante a la hora de escoger esta representación para mostrar información.

Sobre la construcción de los límites y los valores atípicos, Turkey (1997) sugiere una regla sencilla para determinar los límites de los bigotes. Tomando en cuenta que el Rango Intercuartílico (RI) es la diferencia entre el Tercer y el Primer Cuartil, tenemos que existen límites interiores y límites exteriores. Los

primeros son barreras hasta las cuales se “permiten” datos de la muestra, por estar muy cerca del resto. Estos son los límites que definen los extremos de los bigotes. De sobrepasar esta barrera se le considera valor atípico. Los segundos límites indican cuándo un dato se aleja en exceso del resto y, siendo también atípico, se le considera fuera del límite exterior permitido y se dice que es aún más atípico. Se construyen así:

Límite interior inferior = Límite del bigote inferior = $Q_1$ - 1,5RI
Límite interior superior = Límite del bigote superior = $Q_3$ + 1,5RI
Límite exterior inferior = $Q_1$ - 3RI
Límite exterior superior = $Q_3$ + 3RI

En cuanto a la interpretación del Boxpplot se deberán tener las siguientes consideraciones a la hora de interpretarlo.

1) Mientras más larga la caja y los bigotes, más dispersa es la distribución de datos.
2) La distancia entre las cinco medidas descritas en el boxplot (sin incluir la media aritmética) puede variar, sin embargo, recuerde que la cantidad de elementos entre una y otra es aproximadamente la misma. Entre el límite inferior y $Q_1$ hay igual cantidad de opiniones que de $Q_1$ a la mediana, de ésta a $Q_3$ y de $Q_3$ al límite superior. Se considera aproximado porque pudiera haber valores atípicos, en cuyo caso la cantidad de elementos se ve levemente modificada.
3) La línea que representa la mediana indica la simetría. Si está relativamente en el centro de la caja la distribución es simétrica. Si por el contrario se acerca al primer o tercer cuartil, la distribución pudiera ser sesgada a la derecha (asimétrica positiva) o sesgada a la izquierda (asimétrica negativa) respectivamente. Esto suele suceder cuando las observaciones o datos tienden a concentrase más hacia un punto de la escala.
4) La mediana puede inclusive coincidir con los cuartiles o con los límites de los bigotes. Esto sucede cuando se concentran muchos datos en un mismo punto, por ejemplo, cuando muchos estudiantes opinan igual en determinada pregunta. Pudiera ser este un caso particular de una distribución sesgada o el caso de una distribución muy homogénea.
5) Las opiniones emitidas como *No aplica* (N/A) cuando en realidad sí aplica o las opiniones nulas (cuando el estudiante no opina en una pregunta), no son tomadas en cuenta para elaborar el boxplot de esa pregunta. Por esta razón encontrará que en ocasiones no hay igual número de opiniones para todas las preguntas.
6) Se debe estar atento al número de estudiantes que opina en cada pregunta. Lo que pareciera ser dispersión en los resultados, en ocasiones podría deberse a un tamaño de muestra muy pequeño: pocos estudiantes opinaron. Debe ser cauteloso a la hora de interpretar. En estos casos se sugiere remitirse al reporte numérico.
7) En términos comparativos, procure identificar aquellas preguntas cuyos boxplot parecen diferir del resto. Pudiera con esto encontrar fortalezas o debilidades en su actuación según la opinión de los estudiantes.

De esta forma considerando las características de esta herramienta, la tomamos como base para plantear nuestra aproximación a lo que sería un proceso automático de interpretación tiene su origen en la representación gráfica del boxplot múltiple como base del método para discretizar variables numéricas en tal forma que las particularidades de de las clases son producidas por éste. Así, la metodología expuesta en este trabajo de tesis, aunque inspirada en esta herramienta gráfica estadística ha sido automatizada usando algoritmos no gráficos [52], [54].

## 2.7 ELEMENTOS LÓGICO-COMBINATORIOS DEL MODELO MATEMÁTICO

En esta parte se exponen, de manera sucinta, la teoría y herramientas como elementos para la creación de modelos matemáticos a partir del enfoque Lógico-Combinatorio en las zonas del conocimiento poco formalizadas [109].

Con la idea de ayudar a la comprensión de las ideas básicas del reconocimiento de patrones con el enfoque lógico-combinatorio, reflexionaremos un poco en torno a los términos fundamentales, a saber: *objeto, patrón, clase, rasgo y reconocimiento*.

***Objeto*.**- Es un concepto con el cual representamos a los elementos sujetos de estudio como un fenómeno o ente y puede ser físico o abstracto.
***Patrón*.**- Representación de un objeto.
***Clase*.**- Es un conjunto de objetos. También se denomina clase a aquellos conjuntos que forman un cubrimiento; es decir, se puede tener intersección no vacía.
***Rasgo***.- Propiedad, factor, característica que debe tenerse en cuenta en el estudio de los objetos dados. De hecho, son los rasgos el vehículo para trabajar con los objetos.
***Reconocimiento*.**- Es un proceso de clasificación de un elemento en un conjunto; es decir, el procedimiento por el cual se pueden determinar las relaciones de pertenencia entre un elemento cualquiera y un conjunto de clases (dadas todas o algunas de ellas), o la formación de esos conjuntos a partir de relaciones entre los objetos, además de detectar los rasgos [106], [167].

Los anteriores conceptos dados en forma intuitiva, constituyen la base sobre la cual se realiza el estudio de los cuatro tipos de problemas fundamentales de Reconocimiento de Patrones, a saber:

1. La selección de variables, la cual tiene dos usos principales:
    - Para disminuir el número de rasgos en términos de los cuales se debe describir los objetos en modo eficiente, y
    - para encontrar los rasgos que inciden en el problema de manera determinante.
2. La clasificación supervisada. Dado un universo de objetos (por ejemplo pacientes) se agrupa en un número dado de clases (enfermedades), de las cuales se tiene una muestra de objetos (no todos) en cada una, que

pertenecen a ella (ya fueron diagnosticados). El problema consiste en que dado un nuevo objeto (un nuevo paciente), se pueden establecer sus relaciones con cada una de dichas clases.

3. Clasificación no supervisada. En estos problemas no se conoce cómo se agrupan los objetos, y es justamente el objetivo que se persigue.
4. Clasificación parcialmente supervisada. Es una de las familias de problemas menos estudiada en Reconocimiento de Patrones. Uno de estos problemas es análogo al de clasificación con aprendizaje, excepto que hay al menos una clase de objetos de la que no tenemos una muestra y el problema en general sigue siendo el mismo: dado un nuevo objeto, relacionarlo con los ya clasificados [162], [164], [166-167].

# CAPÍTULO 3.

# MARCO TEÓRICO

Para el desarrollo de este modelo de caracterización e interpretación de descripciones conceptuales en dominios poco estructurados se deberán tener en cuenta los conocimientos de los conceptos básicos sobre caracterización, de la herramienta estadística denominado *Boxplot* para observar la relación entre variables y las clases y, en especial su utilidad para representar las diferencias entre grupos; del aprendizaje supervisado que permite que a partir de una clasificación de referencia obtengamos un conjunto de reglas para decidir la clase a la que pertenece cada elemento en el Universo del discurso, del proceso Knowledge Discovery in Data Base (KDD) en el cual este modelo tiene su marco natural de referencia y conceptos básicos sobre lógica difusa, que nos permiten establecer el modelo de etiquetas lingüísticas útil para la visualización de resultados.

## 3.1 CONCEPTOS BÁSICOS

El punto de partida son los trabajos previos [10], donde se analizó la caracterización e interpretación de clases a partir de variables cualitativas, usando conceptos fundamentales como: variable caracterizadora (v.c.), variable parcialmente-caracterizadora y Sistema caracterizador (mínimo, completo).

### 3.1.1 Caracterización a partir de variables categóricas

Definimos *el conjunto de valores propios* de la variable $X_k$ para la clase $C$, representado por $\Lambda_c^k$ como: el conjunto de valores de $X_k$ que toman algunos elementos de $C$ y ningún otro elemento fuera de $C$ los toma; esto es, son valores exclusivos de $C$. Estos valores propios, cuando ocurren, identifican una clase con toda seguridad, por lo que son llamados *valores caracterizadores* de la clase $C$ [10].

Una *variable caracterizadora* $X_k$ es caracterizadora de una clase dada $C$ si

$$\exists\ \Lambda_c^k : \forall i \in C, x_{ik} \in \Lambda_c^k \wedge \forall i \notin C, x_{ik} \notin \Lambda_c^k \neq \phi$$

Para los dominios poco estructurados (dpe) en general es difícil encontrar variables caracterizadoras para las clases de una partición *P*. Para nuestros propósitos es interesante considerar las variables $X_k$ que son parcialmente caracterizadoras de una clase $C$. Estas variables se definen como $X_k$ tales que: $\Lambda_c^k \neq \phi$. Esto es, si tiene al menos un valor propio de la clase $C$, aunque puede compartir alguno con otra clase.

Así, de los conceptos básicos en la construcción de este modelo híbrido es el de la representación para identificar lo que definimos como variable caracterizadora de la clase $C$, concepto que descansa a su vez en el de valor propio de la clase $C$. Así, definimos los siguientes conceptos [25]:

- Un valor $c_s^k \in D_k$ de la variable $X_k$ es propio de la clase $C$, si cumple:

$$(\exists i \in C : x_{ik} = c_s^k) \wedge (\forall i \notin C : x_{ik} \neq c_s^{k)}$$

Estos valores, cuando ocurren, identifican una clase con toda seguridad, por lo que, los llamaremos *valores caracterizadores* de $C$ y los denotamos por $\lambda_{sc}^k$, siendo $C$ la clase y $k$ la variable.

- Una variable $X_k$ es *parcialmente caracterizadora* de la clase $C \in P$ si tiene al menos un valor propio de la clase $C$, aunque puede compartir alguno con otras clases; llamemos $V_c^k$ al conjunto de valores parcialmente caracterizadores de $C$:

$$V_c^k = \{c_s^k : c_s^k es\ valor\ \ propio\ de\ X_x\ \ para\ la\ clase\ C\}$$

- Una variable $X_k$ es *totalmente caracterizadora* de la clase $C \in P$, si todos los valores que tiene $X_k$ en la clase $C$ son *propios* de $C$. En este caso, denotamos por $\Lambda_c^k$ el conjunto de estos valores, los cuales caracterizan totalmente a la clase $C$:

$$\Lambda_c^k = \left\{c_j^k : c_j^k \in V_c^k \wedge \forall C' \neq C\ ,\ c_j^k \notin V_{c'}^k\right\}$$

### 3.1.2 Sistema de caracterización

Para la caracterización de una partición es necesario describir los que es un Sistema de Caracterización. Una partición puede ser caracterizada por lo que llamamos Sistema de Caracterización ($S$):

$$S = \{(C, X_k, \Lambda_c^k) :\ C \in P \wedge \Lambda_c^k \neq \phi\ \}$$

El Sistema de Caracterización $S$ es mínimo y completo, si contiene únicamente una tripleta para cada clase $C \in P$ [10].

$$\exists C \in P : \forall (C', X_x, \Lambda_c^k) \in S \rightarrow C \neq C'$$

Algunas veces el sistema de caracterización $S$ No es Completo:

## 3.2 EL BOXPLOT

Un boxplot múltiple es una herramienta estadística cuya representación gráfica [77], muestra la relación entre una variable numérica y algunos grupos/clases. Para cada grupo, se visualiza el intervalo de valores que toma la variable, y las observaciones atípicas (outliers) se marcan con "*". Para cada clase, se despliega una caja de $Q1$ (primer cuartil) a $Q3$ (tercer cuartil) que representa el 50% de los valores de esa clase, y a partir de ésta se marcan los "bigotes" con sus extremos, el mínimo y el máximo, que representan cada uno el 25% de los valores de esa clase; la mediana se marca con una línea horizontal.

Es muy fácil observar si el *boxplot* múltiple de cierta clase no interseca con el de las demás; en un caso así, la variable es *totalmente caracterizadora*. A veces, sólo es una parte del *boxplot* la que no interseca; en este caso se trata de una variable *parcialmente caracterizadora*.

Para identificar estas variables, estudiaremos los valores propios que toma una variable $X_k$ en una clase $C$, *en relación* a las otras y poder ver si son de la clase o no; para ello hay que analizar cómo son las interacciones entre clases.

Ejemplo. En la figura 3.1 se muestra el boxplot múltiple de la variable $DQO-AT$ (materia orgánica química), que representa una variable del dominio del tratamiento de aguas residuales, la cual nos indica la cantidad de materia orgánica química del agua tratada en una planta depuradora [153]. Como se puede observar el la figura 3.1, cada clase tiene asociada un boxplot. Al proyectar los boxplot de las diferentes clases sobre el eje vertical, la mayoría se intersecan, lo cual significa que comparten valores propios con otras clases definiendo a la variable parcialmente caracterizadora para algunas clases, aunque podemos observar que el boxplot asociado a la clase $C9$ no interseca a ningún otro boxplot; por lo cual, la variable se define como una variable totalmente caracterizadora para la clase $C9$.

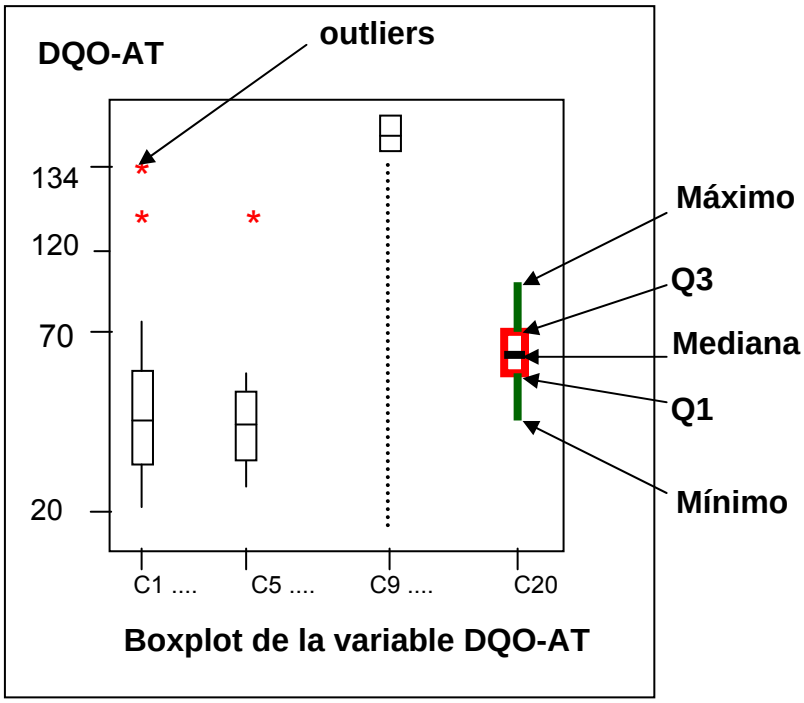



Figura 3.1 Boxplot de la variable materia orgánica química DQO-AT

## 3.3 APRENDIZAJE AUTOMÁTICO

Dado un patrón, su reconocimiento/clasificación puede consistir de una de las siguientes dos tareas [82]: (i) clasificación supervisada (ej., análisis discrimínante) en la cual el patrón de entrada se identifica como un miembro de una clase predefinida, (ii) clasificación no supervisada (ej., clustering) en la cual al patrón se le asigna una clase desconocida hasta ese momento. Aquí el problema de reconocimiento se ésta considerando como una tarea de clasificación o categorización, donde las clases están definidas por el diseñador del sistema (en clasificación supervisada). A pesar de los más de cincuenta años de investigación y desarrollo en este campo, el problema general de reconocimiento de patrones con una orientación, ubicación y escalamiento no se ha resuelto; esto es, *no se ha conseguido un diseño de un reconocedor de patrones automático de propósito general*.

El diseño de un sistema de reconocimiento de patrones incluye los siguientes tres aspectos: (i) adquisición de datos y preprocesamiento, (ii) representación de datos y (iii) toma de decisiones. El dominio del problema sugiere la selección de los sensores, la técnica de preprocesamiento, el esquema de representación y el modelo de toma de decisiones. Generalmente un problema de reconocimiento bien definido y suficientemente delimitado (pocas variaciones intra-clases y muchas variaciones inter-clases) conducen a una representación compacta de patrones y a una estrategia simple de toma de decisiones. Por lo que ningún enfoque, por sencillo que sea, será el mejor ya que se han de utilizar diferentes técnicas y métodos. En consecuencia, la

combinación de éstos es una práctica de uso común en el diseño de sistemas híbridos de reconocimiento de patrones [83].

Entre los enfoques conocidos se pueden mencionar: i) patrones de referencia [84-85], ii) clasificación estadística [86-89], iii) igualación sintáctica o estructural [83], [88-90], iv) redes neuronales [91-92], v) memorias asociativas [93-97]. La Tabla 3.1 muestra una breve descripción y comparación de estos enfoques.

| Aproximación | Representación | Función de reconocimiento | Criterio típico |
|---|---|---|---|
| Patrones de referencia | Muestras, píxeles curvas | Correlación, medida de distancia | Error de clasificación |
| Estadística | Características | Función discriminante | Error de clasificación |
| Sintáctica o estructural | Primitivas | Reglas, gramática | Error de aceptación |
| Redes neuronales | Muestras, píxeles, características | Función de la red | Error cuadrático medio |
| Memorias asociativas | Muestras, píxeles, características | Recuperación de patrones | Robustez a la alteración |

Tabla 3.1. Enfoques de reconocimiento de patrones

Los métodos clásicos de clasificación automática aplicada a dominios poco estructurados [10], muchas veces presentan resultados que no se pueden interpretar. En muchas ocasiones el experto tiene suficiente conocimiento para organizar parte del dominio en entidades que tengan sentido. Sin embargo, los métodos estadísticos clásicos prácticamente ignoran esta información. La herramienta Klass+ [24] implementa la metodología de clasificación basada en reglas cuya idea fundamental es recoger este conocimiento en forma de reglas que subdividan el espacio de clasificación en entornos coherentes y respetar esta primera estructuración sugerida directamente por el experto. Con esto se pretende cubrir dos objetivos: i) incorporación a la clasificación de información antes ignorada (como relaciones entre variables ó restricciones), recogida de los objetos de la clasificación que se pretende obtener y ii) garantizar la interpretabilidad de la clasificación obtenida [25].

**Representación del conocimiento del experto e interpretación**.

Introducir un nivel semántico en el proceso de clasificación ha de permitir una interpretación más clara de las clases finales. Incluir relaciones entre variables, condiciones de pertenencia a una clase o restricciones de incompatibilidad de grupos de objetos en un único formalismo conduce a buscar un modelo de representación muy genérico con suficiente potencia para tratar todo esto. Esta es la razón por la que el conocimiento adicional que proporciona el experto se representa a través de reglas lógicas de primer orden

[41].

La estructura de las reglas que contempla el método a usar es sencilla desde el punto de vista sintáctico y muy potente. Una regla está compuesta de una parte derecha que indica el nombre de alguna clase $C$ y una parte izquierda con la condición $A$ que ha de satisfacer un objeto $i$ para formar parte de dicha clase $C$. En resumen diremos que un objeto $i$ es seleccionado por una regla del tipo:

$$r = A(i) \rightarrow C$$

Si $A$ se evalúa como cierto para el objeto $i$.

En general, los objetos pueden satisfacer una, ninguna o más de una regla. Aquéllos que no cumplan ninguna regla no son motivo de preocupación, ya que se ha dicho que el experto proporciona sólo un conocimiento parcial sobre el dominio.

**Metodología de clasificación basada en reglas**.

Una vez construida la Base de Reglas (BR), con ayuda del experto, se puede evaluar qué objetos satisfacen cada una de las reglas. Algunos no satisfacen ninguna. El conjunto de objetos que están en esta situación forma parte de lo que se denota como *clase residual* y se integra a la jerarquía global en la última etapa del proceso de clasificación con reglas [38] y [41].

El resultado de evaluar las reglas sobre los individuos es una partición de la muestra en $k$ clases más la *clase residual*, donde $k$ es el número de partes derechas distintas en las reglas.

Con la finalidad de respetar la estructura de la clasificación jerárquica hace falta que las clases inducidas por las reglas se constituyan en forma de árbol. En primer lugar se realiza una clasificación local de cada una de las clases inducidas por las reglas. Esto genera los primeros nodos internos del árbol final. Por último, los centros de dichas clases se clasifican junto a los elementos de la clase para integrar todos los elementos en un único árbol ascendente jerárquico que es el que dará lugar a la clasificación final [38], [41] y [26].

## 3.4 PROCESO KNOWLEDGE DISCOVERY IN DATABASE (KDD)

En [81] se da un punto de vista práctico del proceso de KDD enfatizando la naturaleza interactiva e iterativa de este proceso; incluye varios pasos con decisiones que tienen que tomarse por el usuario. La Figura 3.2 muestra un diagrama del proceso KDD. A continuación se resume cada una de las etapas:

1. La comprensión del dominio de aplicación, el conocimiento a priori relevante, y las metas del usuario final.

2. La creación de un conjunto de datos destino. Seleccionar un conjunto de datos, o seleccionar un subconjunto de variables o muestra de datos, sobre los cuales se realizará el descubrimiento.
3. La preparación y pre-procesamiento de datos. Operaciones básicas, si fuera necesarias, como la eliminación de ruido, datos atípicos (outliers) o perdidos, recabar la información necesaria para modelar el ruido, decidir sobre estrategias para manejar datos perdidos.
4. La reducción y proyección de datos. Encontrar características útiles para representar los datos depende de las metas del proceso. Usar reducción de la dimensionalidad o métodos de transformación para reducir el número de variables bajo consideración o para encontrar representaciones invariantes para los datos.
5. Seleccionar la tarea de minería de datos. Decidiendo si la meta del proceso KDD es clasificación, regresión, clustering o alguna otra.
6. Seleccionar el o los algoritmo(s) de minería de datos. Seleccionar los métodos que se emplearán en la investigación para identificar patrones en los datos. Esto incluye decir qué modelos y parámetros son los apropiados y escoger un método de minería de datos compatible con el criterio del proceso de KDD.
7. La minería de datos. La investigación de patrones en una representación formal o un conjunto de representaciones como: reglas de clasificación o árboles, regresión, clustering y así sucesivamente. El usuario puede apoyar el método de minería de datos realizando correctamente los pasos previos.
8. La interpretación de los resultados obtenidos, posible retorno a cualquiera de los pasos previos del 1-7 para iteraciones posteriores.
9. La consolidación del conocimiento descubierto. Incorporación de este conocimiento en el desempeño del sistema, o simplemente documentarlo y reportarlo a las partes interesadas.

El proceso de KDD puede incluir iteraciones significativas y contener ciclos entre cualesquiera dos pasos; así, en cada etapa el “minero informático” puede volver a la etapa que él requiera para continuar su trabajo. La etapa donde se descubre la información es la denominada Minería de Datos.

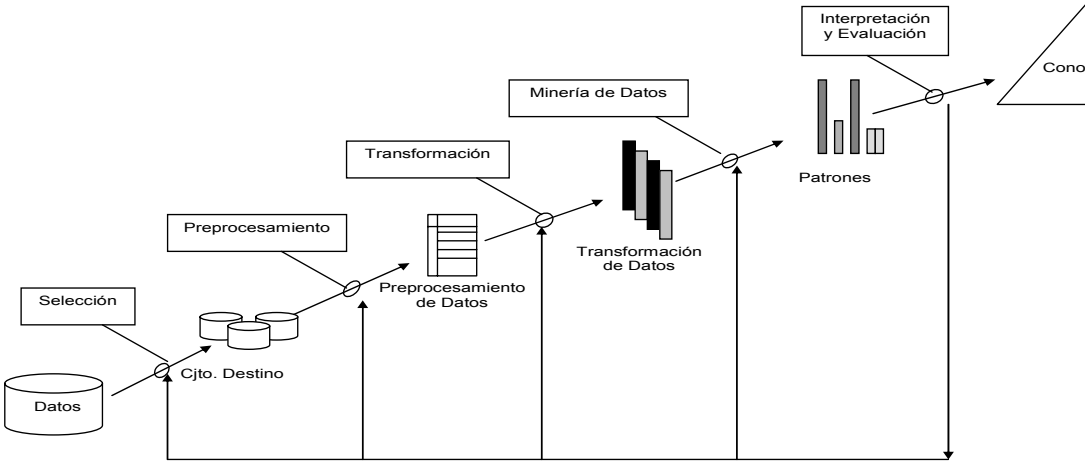



Figura 3.2. Diagrama del proceso KDD

## 3.5 CONCEPTOS BÁSICOS DE LÓGICA DIFUSA

Una gran variedad de ciencias aplican métodos de Inteligencia Artificial (IA) principalmente para modelar el razonamiento del experto. Para el diseño de tales sistemas inteligentes, la importancia de la lógica difusa ha ganando gran aceptación [129]. Publicaciones recientes han mostrado también que los sistemas híbridos en IA han conseguido buenos resultados, combinando lógica difusa e Inteligencia Artificial para la diagnosis médica en la prevención de enfermedades, redes neuronales para el reconocimiento de patrones, sistemas de inferencia difusos para incorporar conocimiento humano, realizar inferencia y tomar decisiones.

Es importante considerar que los problemas complejos del mundo real requieren sistemas inteligentes que combinen conocimiento, técnicas y metodologías de diferentes fuentes.

Estos sistemas inteligentes deberán poseer experiencia como la del humano dentro de un dominio específico, adaptándose y aprendiendo a hacer lo mejor en ambientes dinámicos y explicando cómo toman decisiones o acciones. De frente a los problemas de cálculo, es más ventajoso usar diferentes técnicas de cálculo sinérgicas que exclusivas, obteniendo como resultado la construcción de sistemas híbridos inteligentes [129].

.

### 3.5.1 Lógica difusa

La lógica como base para el razonamiento puede distinguirse por sus tres componentes principales (independientes del contexto): valores de verdad, vocabulario (operadores) y razonamiento (tautologías, silogismos). En la lógica de Boole, los valores de verdad son 0 (falso) ó 1 (verdadero) y por medio de estos valores de verdad, se define el vocabulario vía las tablas de verdad.

Una distinción entre la verdad material y la lógica [112] se hace en las llamadas lógicas extendidas: La Lógica modal [113] distingue entre verdad necesaria y posible, y la lógica temporal [114] entre enunciados que fueron verdaderos en el pasado y aquéllos que serán verdaderos en el futuro. La lógica epistémica [115] trata del conocimiento y las creencias, la lógica deóntica [101] con lo que debe hacerse y que permite ser verdadero. La lógica modal, en particular, podría ser una buena base para aplicar diferentes medidas de la lógica de la incertidumbre.

Otra extensión de la *lógica de Boole* es el cálculo de predicados, el cual es un conjunto lógico teórico que usa cuantificadores y predicados para los operadores de la lógica de Boole.

La *lógica difusa* [116] hace una extensión del conjunto teórico de la lógica multivaluada en la cual los valores de verdad son variables lingüísticas (términos de verdad de variables lingüísticas).

Lo mismo que en la lógica clásica, los operadores se definen en la lógica difusa a través de tablas de verdad, usando el Principio de Extensión para obtener las definiciones de estos operadores [117]. Hasta ahora, la teoría de la posibilidad ha empezado a ser usada para definir operadores en lógica difusa, aunque hay otros operadores que también han sido investigados [118] y que podrían usarse.

Además, podemos considerar conectivos mixtos como funciones para calcular el grado de pertenencia conjunta vía las t-normas en problemas de clasificación [119]. Un punto importante es la relación y diferencia entre los conceptos de probabilidad y posibilidad; con este último concepto se tiene una estrecha relación con el grado de pertenencia a un conjunto difuso [120]. El concepto de posibilidad juega un importante papel particularmente en la representación del significado, en la gestión o manejo de la incertidumbre en sistemas de clasificación, sistemas inteligentes y en algunas otras aplicaciones [121].

### 3.5.2 Razonamiento difuso

En realidad, cuando las personas hablamos acerca de un sistema del mundo real, lo hacemos en tres etapas [122]:

- Seleccionamos un conjunto de variables que podría ser entendido como un conjunto de entidades bien diferenciadas. Tales variables pueden estar directamente vinculados a la experiencia sensorial –y entonces expresada en una manera informal– o pueden estar determinadas por medio de procedimientos de mediciones más precisas.

- Establecemos las relaciones entre las variables, ligando sus estados particulares. Esto en realidad se hace dando reglas como *Si* (*hecho* A) *entonces* (*hecho* B), donde cada hecho describe un estado o un valor preciso de alguna variable particular.

- Finalmente, hay una tercera etapa donde los conjuntos de reglas se organizan para construir una teoría o un modelo que describe el sistema del mundo real bajo estudio.

El sistema está bien comprendido cuando su teoría no conduce a conclusiones contradictorias o a enunciados experimentalmente falsos acerca del sistema.

En este contexto el término *inferencia* se aplica a cualquier algoritmo que se use para derivar consecuencias de hechos conocidos dentro del modelo. La *inferencia* en un amplio sentido puede aparecer en diferentes formas dependiendo del contexto considerado, desde la manipulación simbólica en una base de datos lógica hasta la evaluación de una función numérica o vectorial. En el caso anterior, las reglas aparecen bajo la forma: *Si* $X = x$ *entonces* $Y = f(x), x \in X$, con los hechos conocidos como $X = x_0$ ó $X \in A$, siendo $A$ un

subconjunto de $X$. En Dubois y Prade [122] leemos: "*las reglas: si…entonces… son una herramienta clave para expresar piezas de conocimiento en lógica difusa*"

Sin embargo, cuando las variables consideradas vienen de conceptos graduales como altura, temperatura, cantidad y algunas otras, las descripciones de sus estados están algunas veces dadas también por enunciados graduales e implícitamente vagos. Ejemplos de estos enunciados son la temperatura es alta, el color es azul, entre otros. Más aún, en este caso el conocimiento acerca del sistema puede presentarse en forma de enunciados condicionales ligando estos estados vagos de las variables, tales como "*Si la temperatura es baja, entonces el color es verde*" [119].

Cuando los estados vagos de las variables están representados por *conjuntos difusos* del Universo del discurso donde las variables toman sus posibles valores, el problema surge naturalmente de cómo determinar los hechos y las reglas que se han de combinar para derivar nuevos hechos. Esta es la esencia de la *inferencia difusa*. A estos hechos vagos en el contexto de los conjuntos difusos se les denominan *enunciados difusos* ó *proposiciones difusas*, y las reglas relacionadas con estos hechos como *reglas difusas* ó *enunciados condicionales difusos* [123].

Lo que es evidente desde el punto de vista de la lógica es que la inferencia lógica tiene lugar a nivel semántico. A diferencia de los procedimientos de la lógica clásica, que derivan conclusiones por manipulación simbólica, en la lógica difusa los enunciados difusos están siempre relacionados a los conjuntos difusos que los representan, y el proceso de inferencia total se realiza por manipulación numérica de sus funciones de pertenencia. En esta forma los hechos inferidos se constituyen a partir de sus funciones de pertenencia, y no en forma inversa [124-126], [128].

La interpretación de procesos de inferencia difusa como procesos de razonamiento aproximado nos permite comparar qué tan lejanos son los hechos conocidos de los antecedentes y hechos inferidos de los consecuentes.

### 3.5.3 Las etiquetas lingüísticas en la interpretación de resultados.

Debido a la gran importancia que tiene la etapa de interpretación y evaluación del conocimiento obtenido en bases de datos del proceso KDD, es necesaria la aplicación de disciplinas, herramientas, métodos, que sean soporte para el desarrollo de interfases para el usuario en esta etapa del proceso KDD.

Las etiquetas lingüísticas son un medio atractivo de la lógica difusa para visualizar resultados para su interpretación por los usuarios que presentan los sistemas de minería de datos, con el objetivo de apoyar a la toma de decisiones.

La lógica difusa ha cobrado una gran importancia por la variedad de sus aplicaciones, las cuales van desde el control de complejos procesos industriales, hasta el diseño de dispositivos artificiales de deducción automática, pasando por la construcción de artefactos electrónicos de uso doméstico y de entretenimiento. La expedición de patentes industriales de mecanismos basados en la lógica difusa tiene un crecimiento sumamente rápido en todas las naciones industrializadas [134].

La importancia que representa la visualización de resultados para su interpretación en los sistemas de minería de datos, debe de ser tomada en cuenta por los desarrolladores de este tipo de sistemas, ya que no todos los sistemas poseen una adecuada forma o una interfaz adecuada, que visualice clara y sencillamente los resultados.

Muchos de los desarrolladores de este tipo de sistemas, están más preocupados en encontrar conocimiento en bases de datos que en visualizarlo; es por ello que se necesitan métodos que lo visualicen de tal forma que exista una semántica estrictamente cimentada entre el conocimiento obtenido por el sistema y el usuario. Existen muchos sistemas poderosos en este ramo, pero son pocos los sistemas que visualizan con un método adecuado para la interpretación de resultados [129].

En la tabla 3.2 se muestra la forma de visualizar resultados por parte de los sistemas híbridos que se han mencionado. Es de gran importancia haber entendido las descripciones mencionadas anteriormente acerca de estos sistemas, ya que la siguiente descripción está estrechamente relacionada.

La búsqueda de nuevas técnicas para la visualización de resultados, puede hacer que la potencialidad de los sistemas de minería de datos crezca a medida que estos sistemas se apeguen a la realidad, describiendo fenómenos como el ser humano tiene la capacidad de describir si el clima es caliente, frío o templado.

### 3.5.4 Etiquetas lingüísticas.

En el tratamiento de la precisión frente a la complejidad dominante de los sistemas, es natural el uso de las llamadas *variables lingüísticas*; esto es, variables cuyos valores no son números sino palabras o expresiones en lenguaje natural o artificial [144].y [147]

Una de las herramientas básicas para la lógica difusa es el concepto de *variable lingüística* que en 1973 fue llamada *variable de orden superior* más que variable difusa y definida en [128] como:

Definición. Una *variable lingüística* se caracteriza por una quíntupla:

$$(x, T(x), U, G, A),$$

| Sistema | ¿Qué visualiza? | Descripción |
|---|---|---|
| Weka *(Supervisado y no supervisado)* | Únicamente la clase de pertenencia y gráficos de los comportamientos. | Después de haber sometido una base de datos a este sistema, se visualiza únicamente la clase a la que pertenece cada uno de los individuos contenidos en la base de datos. Con la ayuda de una gráfica, se pueden observar las agrupaciones o clases encontradas por el sistema, cada una de las cuales se diferencia ya que se muestran en distinto color. |
| Clementine *(Supervisado)* | Reglas de pertenencia a las clases. | Al analizar una base de datos clasificada, este sistema visualiza una serie de reglas de pertenencia a las clases, para que el usuario de acuerdo con su criterio clasifique nuevos individuos. |
| XpertRule Miner *(Supervisado)* | Gráficos y árboles | Al analizar una base de datos clasificada, este sistema visualiza una serie de gráficos y un diagrama de árbol que indican la pertenencia a las clases, para que el usuario de acuerdo con su criterio clasifique nuevos individuos. |
| Cluesome *(No supervisado)* | Gráficos | Después de haber sometido una base de datos a este sistema, se visualiza en pantalla una gráfica multidimencional donde se pueden observar las agrupaciones o clases encontradas por el sistema; cada una de las clases se diferencia ya que se muestran en distinto color, y se muestra a que clase pertenece cada una de las observaciones en la misma gráfica. |
| Ginko *(No supervisado)* | Gráficos | Se visualiza en pantalla una gráfica donde se pueden observar las agrupaciones o clases encontradas por el sistema; cada una de las clases se diferencia ya que se muestran en distinto color, y se muestra a qué clase pertenece cada observación. La posición de dicha gráfica, se puede manipular con el objetivo de tener flexibilidad en la observación. |
| CIADEC *(Supervisado)* | La clase de pertenencia, reglas y gráficos. | Al analizar una base de datos clasificada, este sistema visualiza: sistemas reglas, gráficos de pertenencia a las clases e interpretación de resultados. Al analizar un nuevo individuo, el sistema proporciona la clase a la que pertenece para que con el sistema de gráficos interprete los resultados, observando el grado de pertenencia a las clases. |

Tabla 3.2 Métodos de visualización de resultados.

En la cual $x$ es el nombre de la variable; $T(x)$ (o simplemente $T$) denota el conjunto de términos de $x$, esto es, el conjunto de los nombres de los *valores lingüísticos* de $x$, y cada valor es una variable difusa denotada

genéricamente por $X$ la cual se extiende sobre un universo de discurso $U$, que se asocia con la variable de base $u$; $G$ es una regla sintáctica (la cual comúnmente tiene la forma de una gramática) para generar el nombre, $X$, de valores de $x$; y $A$ es una *regla semántica* que asocia a cada $X$ su significado, $A(X)$, es un subconjunto difuso de $U$. Una $X$ particular, esto es, un nombre generado por $G$, se llama un *término.* Deberá notarse que la variable base $u$ puede ser un vector.

Las *etiquetas lingüísticas* son el centro de las técnicas de modelado difuso que ejemplifican la idea de variable lingüística. Desde su raíz, una variable lingüística es el nombre de un conjunto difuso. Si tenemos un conjunto difuso llamado "largo", éste es una simple *variable lingüística*, al igual que otro conjunto llamado “corto”; a cada conjunto difuso se le atribuye una etiqueta, el conjunto difuso está constituido por un rango de valores del Universo del discurso $U$. Una variable lingüística encapsula las propiedades de aproximación o conceptos de imprecisión en un sistema. Esto reduce la aparente complejidad de describir lo que debe concordar con su semántica.

En el campo de la semántica difusa cuantitativa al significado de un término "$x$" se le representa como un conjunto difuso $M(x)$ del universo de discusión. Desde este punto de vista, uno de los problemas básicos en semántica es que se desea calcular el significado de un término compuesto.

La idea básica sugerida por Zadeh [144] es que una *etiqueta lingüística* tal como "muy", "más o menos", "ligeramente". Puede considerarse como un operador que actúa sobre un conjunto difuso asociado al significado de su operando. Por ejemplo, en el caso de un término compuesto "muy alto", el operador "muy" actúa en el conjunto difuso asociado al significado del operando "alto". Una representación aproximada para una etiqueta lingüística se puede lograr en términos de combinaciones o composiciones de las operaciones básicas. En [144] se considera que las etiquetas lingüísticas pueden clasificarse en dos categorías que se definen como sigue:

*Tipo I*: las que pueden representarse como operadores que actúan en un conjunto difuso: "muy", "más o menos", "mucho", "ligeramente", "altamente", "bastante".

*Tipo II*: las que requieren una descripción de cómo actúan en los componentes del conjunto difuso (operando): "esencialmente", "técnicamente", "estrictamente", "prácticamente", "virtualmente", etc. Su caracterización envuelve una descripción de forma que afectan a los componentes del operando y por lo tanto es más compleja que las del tipo I. En general, la definición de una etiqueta de este tipo debe formularse como un algoritmo difuso que envuelve etiquetas tipo I.

En otras palabras, las etiquetas lingüísticas pueden ser caracterizadas cómo operadores más que construcciones complicadas sobre las operaciones primitivas de conjuntos difusos.

En la actualidad la mayoría de las decisiones proceden de problemas relacionados con el transcurso del tiempo (TS), el análisis económico y financiero, campos donde se relacionan generalmente con las decisiones del humano soportados por software desarrollado con técnicas de la estadística y de minería de datos. En un futuro, la importancia de estos "sistemas inteligentes" estará relacionada con la posibilidad de operarlos con información lingüística, razonando y respondiendo cuestiones prometedoras en el campo de la investigación. La Teoría de la Percepción Computacional (Computational Theory of Perceptions, CTP) [145-146] puede servir básicamente para el avance de estos sistemas. La lógica difusa constituye el cuerpo de la CTP haciendo poderosas las herramientas para el modelado y procesamiento de información lingüística de dominios cuantitativos. La metodología para el uso de palabras propone el uso de métodos de razonamiento basados en modelos difusos.

El éxito de la lógica difusa en aplicaciones de control y sistemas para el reconocimiento de patrones hace posible el uso de descripciones lingüísticas para áreas que regularmente están basadas con variables numéricas. La referencia [144] llama la atención cuando se basa en la aplicación de la lógica difusa para el apoyo a las decisiones en áreas económicas, financieras, ciencias terrenales, entre otras, con el rol central de la percepción humana. La percepciones son basadas en preposiciones como: "el precio del gas es muy alto" o "es muy improbable que suba el peso". Es normal el uso este tipo de proposiciones en las decisiones de las personas. Los términos bajo, muy improbable, alto, más o menos, normalmente están constituidos por una graduación difusa de información [147].

### 3.5.5 Ejemplo.

Los sistemas para el descubrimiento de conocimiento con respecto a su visualización de resultados, siempre están sostenidos sobre conjuntos difusos, lo que hace posible la aplicación de etiquetas lingüísticas para visualizar resultados. En este ejemplo, se plantea el uso de etiquetas lingüísticas como medio de visualización de resultados.

Sea el universo $X = \{0, 1, 2, 3,..., 198, 199, 200\}$, un conjunto de velocidades posibles de automóviles, medidas en km/Hr y $E = \{Parado, Muy\ lento, Lento, Mediano, Rápido, Muy\ rápido\}$ un conjunto de etiquetas lingüísticas que hacen referencia a modelos difusos del universo $X$. Aplicar el conjunto $E$ al universo $X$.

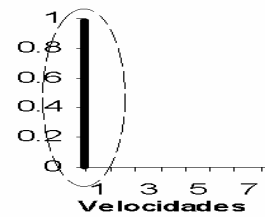
Solución:

Generando los modelos difusos para aplicar etiquetas lingüísticas que hagan referencia al universo $X$. Sea $x \in X$.

**1.- Parado**: *En este caso, analizando los elementos de $X$, se dice que el automóvil está en reposo, lo cual constituye el siguiente conjunto no difuso.*

Como se puede observar en la gráfica 3.1 la etiqueta “Parado”, sólo constituye un elemento de $X$.

$$\text{Parado} \begin{cases} 1 & si\ x = 0 \\ 0 & si\ x > 0 \end{cases}$$



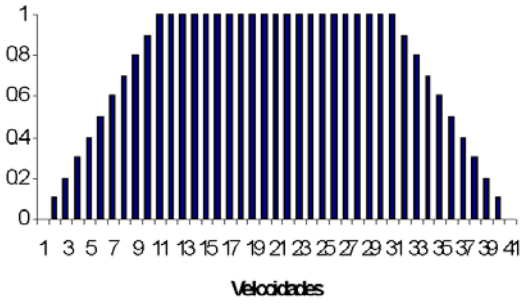

**Grafica 3.1** Etiqueta “Parado”.

Para hacer una distribución de las etiquetas lingüísticas, tomamos el camino más sencillo que es dividir los elementos $x$ restantes de $X$ entre los elementos restantes de $E$, dándonos un rango de 40 elementos del universo $X$ para cada una de las cinco etiquetas restantes del conjunto $E$.

Para poder modelar la solución a un sentido que se asemeje a la realidad, necesitamos hacer una operación que nos indique el nivel de pertenencia de los valores fronterizos de los conjuntos difusos, es decir, la velocidad 39, no le pertenece en un 100% a la etiqueta “Muy lento”, ya que ésta casi pertenece a las velocidades fronterizas de la siguiente etiqueta; es por ello que se necesita saber qué rango de valores de velocidad pertenecen a una etiqueta al 100% y cuál es el valor de pertenencia de las restantes. En las siguientes formulaciones de los conjuntos difusos para las etiquetas lingüísticas restantes, se representan los niveles de pertenencia de las velocidades a las respectivas etiquetas.

**2.- Muy lento**: Para esta etiqueta, se genera el siguiente modelo difuso.

$$\text{Muy lento} \begin{cases} \dfrac{x}{10} & si\ 0 \le x \le 10 \\ 1 & si\ 10 \le x \le 30 \\ \dfrac{40 - x}{10} & si\ 30 \le x \le 40 \\ 0 & si\ 40 \le x \le 200 \end{cases}$$



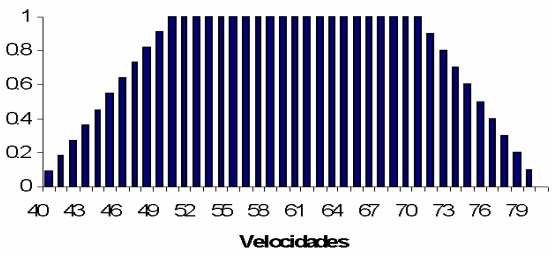

Grafica 3.2 Etiqueta “Muy lento”.

En la gráfica 3.2 se pueden observar los niveles de pertenencia de cada valor de velocidad con respecto a la etiqueta “Muy lento”.

**3.- Lento**: Para esta etiqueta, se genera el siguiente modelo difuso.

Lento

$$\begin{cases} \frac{x-39}{11} & si\ 40 \le x \le 50 \\ 1 & si\ 50 \le x \le 70 \\ \frac{80-x}{10} & si\ 70 \le x \le 80 \\ 0 & si\ 80 \le x \le 200 \end{cases}$$



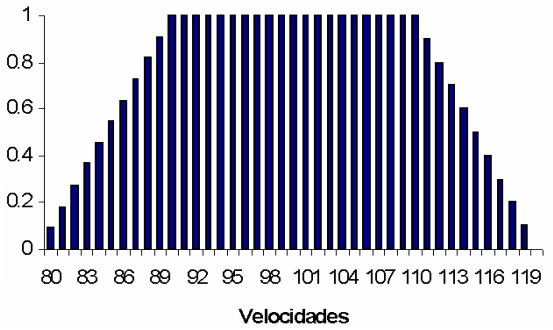

Grafica 3.3 Etiqueta “Lento”.

En la gráfica 3.3 se pueden observar los niveles de pertenencia de cada valor de velocidad con respecto a la etiqueta “Lento”.

**4.- Mediano**: Para esta etiqueta, se genera el siguiente modelo difuso.

Mediano

$$\begin{cases} \frac{x-79}{11} & si\ 80 \le x \le 90 \\ 1 & si\ 90 \le x \le 110 \\ \frac{120-x}{10} & si\ 110 \le x \le 120 \\ 0 & si\ 120 \le x \le 200 \end{cases}$$



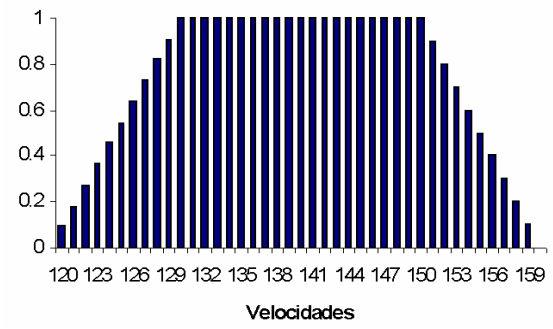

Grafica 3.4 Etiqueta “Mediano”.

En la gráfica 3.4 se pueden observar los niveles de pertenencia de cada valor de velocidad con respecto a la etiqueta “Mediano”.

**5.- Rápido**: Para esta etiqueta, se genera el siguiente modelo difuso.

Rápido

$$\begin{cases} \frac{x-119}{11} & si\ 120 \le x \le 130 \\ 1 & si\ 130 \le x \le 150 \\ \frac{160-x}{10} & si\ 150 \le x \le 160 \\ 0 & si\ 160 \le x \le 200 \end{cases}$$

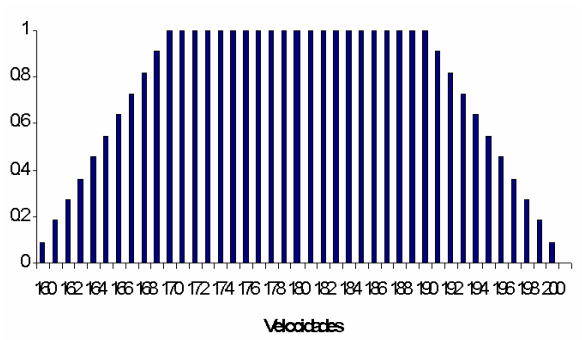



Grafica 3.5 Etiqueta “Rápido”.

En la gráfica 3.5 se pueden observar los niveles de pertenencia de cada valor de velocidad con respecto a la etiqueta “Rápido”.

**6.- Muy rápido**: Para esta etiqueta, se genera el siguiente modelo difuso.

$$\text{Muy rápido} \begin{cases} \frac{x-159}{11} & si\ 160 \le x \le 170 \\ 1 & si\ 170 \le x \le 190 \\ \frac{201-x}{11} & si\ 190 \le x \le 200 \\ 0 & si\ x > 200 \end{cases}$$

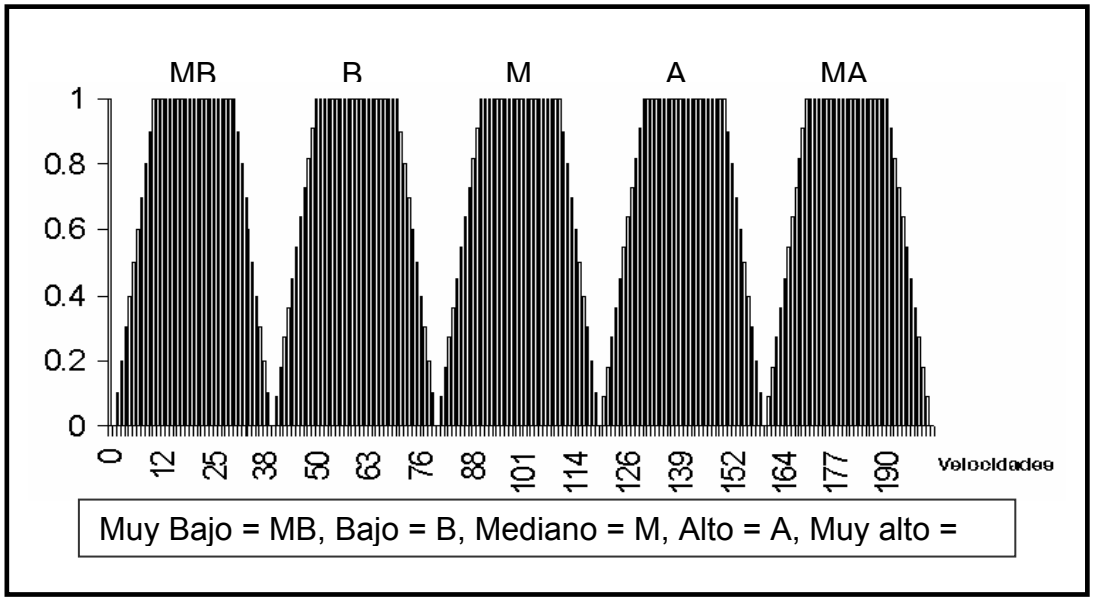



Gráfica 3.6 Etiqueta “”Muy Rápido”

En la gráfica 3.6 se pueden observar los niveles de pertenencia de cada valor de velocidad con respecto a la etiqueta “Muy rápido”.

Ésta es la forma de cómo implementar etiquetas lingüísticas sobre un universo; en este caso, se construyeron modelos difusos para poder implementar las etiquetas lingüísticas, quedando un el universo dividido en conjuntos difusos, y esta división se puede observar en la gráfica 3.7, que muestra los niveles de pertenencia de cada una de las velocidades del universo hacia las etiquetas lingüísticas.

Una mejor visualización de resultados hace que la interpretación de éste sea correcta y se tenga un mejor apoyo en la toma de decisiones. El uso de etiquetas lingüísticas para la representación de resultados en cualquier dominio, hacen generar proposiciones que forman el núcleo de nuestras relaciones con "la forma de las cosas en el mundo" e incorporar conceptos que hacen lograr que los sistemas sean potentes y se aproximen más a la realidad.




**Grafica 3.7** Comportamiento de las etiqueta lingüísticas sobre $X$.

# CAPÍTULO 4.

## MODELO PROPUESTO

La propuesta del modelo aporta una Familia de Algoritmos que establecen un Sistema de Caracterización de clases, basado en predicados de lógica de primer orden ($CP_1$), que permiten máxima potencia y flexibilidad para detectar variables cuantitativas caracterizadoras en algunas clases, permitiendo un procedimiento de generación automático de reglas, que formarán parte de la base de conocimiento de un sistema orientado a la predicción o diagnóstico. Además, la automatización de este sistema de caracterización ofrecerá un conjunto de herramientas de apoyo a la interpretación como: la construcción de un sistema de reglas, visualización de las funciones de pertenencia que determinan el nivel característico de las variables $X_k$ a las distintas clases $C$, evaluación de individuos nuevos de acuerdo con las reglas generadas y validación de la calidad de la predicción teniendo como base un conjunto de nuevos objetos.

### 4.1 El modelo propuesto

El modelo esta conformado por los siguientes nueve pasos:

1. **Uso del *boxplot múltiple* como herramienta gráfica para la detección de variables caracterizadoras**

   En esta primera etapa, se utilizan algunas técnicas descriptivas clásicas (presentación de datos, medidas de tendencia central, representación de valores muéstrales entre otros) que permiten identificar el comportamiento y naturaleza de los datos en la matriz $X$. Esta etapa sirve para obtener información preliminar acerca de la variabilidad de las mediciones y para representar los *boxplots múltiples*, que a su vez permiten observar la relación entre las variables y las clases y, en especial, es útil para representar las diferencias entre grupos.

   El modelo propuesto está inspirado en el *boxplot múltiple*, el cual es una herramienta que nos permite visualizar y comparar la distribución de una variable a través de todas las clases. Con la representación de las variables, podemos identificar lo que se denominan *variables caracterizadoras* de la clase $C$, concepto que descansa a su vez en el concepto de valor propio de una clase $C$.

   Así, definimos los siguientes conceptos:

- Un valor $c_S^K \in D_K$ de la variable $X_k$ es propio de la clase $C$, si cumple:

$$(\exists i \in C : x_{ik} = c_s^k) \wedge (\forall i \notin C : x_{ik} \neq c_s^k)$$

Estos valores, cuando ocurren, identifican una clase con toda seguridad, por lo que los llamamos *valores caracterizadores* de $C$ y los denotamos por $\lambda_{sc}^{k}$, siendo $C$ la clase y $k$ la variable.

- Una variable $X_k$ es *parcialmente caracterizadora* de la clase $C \in P$ si tiene al menos un valor propio de la clase $C$, aunque puede compartir alguno con otras clases; llamemos $V_C^k$ al conjunto de valores *parcialmente caracterizadores* de $C$:

$$V_c^k = \{c_s^k : c_s^k es\ valor\ de\ X_k\ para\ la\ clase\ C\}$$

- Una variable $X_k$ es *totalmente caracterizadora* de la clase $C \in P$, si todos los valores que tiene $X_k$ en la clase $C$ son *propios* de $C$. En este caso, denotamos por $\Lambda_C^k$ el conjunto de estos valores, los cuales caracterizan totalmente a la clase $C$:

$$\Lambda_c^k = \left\{c_j^k : c_j^k \in V_c^k \wedge \forall C' \neq C\,,\, c_j^k \notin V_{c'}^k\right\}$$

Es muy fácil observar si el *boxplot* de cierta clase no interseca con el de las demás; en un caso así, la variable es *totalmente caracterizadora*. A veces, sólo es una parte del *boxplot* la que no interseca; en ese caso se trata de una variable *parcialmente caracterizadora*.

Para identificar estas variables, estudiaremos los valores propios que toma una variable $X_k$ en una clase $C$, *en relación* a las otras y poder ver si son de la clase o no; para ello hay que analizar cómo son las interacciones entre clases.

### 2. Estudio de interacciones entre clases

En este proceso, es de nuestro interés considerar las variables en su estado natural, evitando cualquier transformación arbitraria sobre su naturaleza, que pudiera alterar el sentido de la interacción.

Esta etapa consiste en identificar todas las intersecciones que se dan entre los valores de las variables y las distintas clases, determinando en qué puntos del rango de las variables están cambiando estas intersecciones; así, podemos identificar las distintas combinaciones de clases donde se puede dar un mismo valor de cierta variable, y como consecuencia hacer emerger los valores propios (caracterizadores) de una clase; éstos nos identificarán *variables total o parcialmente caracterizadoras*.

Sin embargo, en la práctica no se puede basar un proceso automático en la interpretación de una representación gráfica, por lo que en los siguientes

apartados se propone una alternativa equivalente, pero automatizable, que representa la esencia del modelo.

### 3. Sistema de intervalos o ventanas de longitud variable

Estas intersecciones entre las distintas clases se pueden encontrar de forma exacta con un costo computacional mínimo, solamente calculando los valores mínimos y máximos por variable y clase, y ordenándolos en forma conveniente.

**NOTA**: *Este paso es una de las aportaciones fundamentales de este trabajo de tesis.*

Así, a partir de esta ordenación, se define una discretización de la variable $X_k$ en un conjunto de intervalos de longitud variable, sobre los que se podrán identificar los valores propios de dicha variable en todas las clases.

Formalizando estos conceptos, tenemos que $m_c^k \; y \; M_C^k$ son, respectivamente, los mínimos y los máximos de la variable $X_k$ en la clase $C \in P$, observados de la descriptiva o del *boxplot múltiple*, donde $m_C^k = \min_{i \in C} \{x_{ik}\}$ y $M_C^k = \max_{i \in C} \{x_{ik}\}$.

Ahora el proceso de ordenamiento ascendente consiste en:

- Definir $M^k$ como el conjunto de todos los mínimos y máximos correspondientes a la variable $X_k$, en todas las clases de *P*; esto es:

$$M^K = \left\{ m_{c_1}^k, \ldots, m_{c_\xi}^k, M_{c_1}^k, \ldots, M_{c_\xi}^k \right\}$$

siendo $card\left(M^K\right) = 2\xi$

- Ordenar los valores $M^K$ de menor a mayor valor, y para ello se construye un conjunto $Z^K$ de forma que:

$$\mathrm{Z^K} = \left\{ z_i^k \; ; i = 1 : 2\xi \right\}, \text{ tal que:}$$

i. $z_1^\kappa = \min M^k$
ii. $z_i^k = \min\left(M^k \setminus \left\{z_j^k \; ; \; j < i\right\}\right), \; 2 \le i \le 2\xi$ y $1 \le j \le 2\xi$

Dado que $\mathrm{Z^K} = \left\{ z_i^k \right\}$ es un conjunto ordenado, sus elementos tienen la siguiente propiedad:

$$Z^k = \left\{ z_j^k \middle| \, z_{j-1}^k < z_j^k \; ; 1 < j \le 2\xi \right\}$$

A este conjunto lo denominamos conjunto de *puntos de corte.*

- Construir, a partir de este conjunto ordenado, el sistema de intervalos de longitud variable $I^k$ de la siguiente forma:

$$I^k = \left\{ I_s^k : 1 \le s \le 2\xi - 1 \right\}, \text{ donde:}$$

i. $I_1^k = \left[ z_1^k , z_2^k \right]$

ii. $I_s^k = \left( z_s^k , z_{s+1}^k \right], \quad (s = 2 : 2\xi - 1)$

De ahí, definimos una nueva variable categórica $I^k$ cuyo conjunto de valores es $D^k = \left\{ I_1^k, \ldots, I_{2\xi-1}^k \right\}$; la variable $I^k$ identifica todas las intersecciones entre clases que se definen para cada variable $X_k$, y este sistema de intervalos de longitud variable está asociado a cada variable $X_k$.

Así, si tenemos $2\xi$ puntos de corte diferentes se generan a lo más $2\xi$-1 intervalos y *card(*$D^k$*)* = $2\xi$-1, recordando que $\xi$ es el número de clases de la partición de referencia $P$ que se quiere caracterizar.

Además, siendo $D^k$ el dominio de $X_k$, éste representa una categorización del mismo, pero no es arbitraria en absoluto, y además se calcula de forma inmediata. Por último, hay que observar que para construir $I^k$ ya no hace falta realizar el *boxplot múltiple*, aunque éste sigue siendo una excelente representación de lo que se ésta haciendo.

### 4. Construcción de la tabla de contingencia de clases vs intervalos

En esta etapa se realiza la construcción de la tabla de contingencia para cada variable $X_k$, como una matriz de números $A$, en la cual los renglones están representados por los intervalos $I^k$ encontrados en la etapa anterior, y las columnas por las clases de la partición de referencia $P$; así, una cierta casilla de la matriz $A$ indica el número de elementos del dominio $I$, cuyos valores de $X_k$ se encuentran en el intervalo representado por $I_s^k$. En general, para un cierto valor de la variable $X_k$ se tienen elementos en distintas clases.

De esta forma, definimos la tabla de contingencia como:

$$A = I^k \times \mathrm{P} = (\mathrm{n_{sc}}(s = 1 : 2\xi - 1), (C \in P)), \text{ donde:}$$

$n_{sc}$ es *card* $\left\{ i \in C \wedge x_{ik} \in I_s^k \right\}$; es decir, $n_{sc}$ es el *número de elementos* de $C$ cuyo valor de $X_k$ está en $I_s^k$. La matriz $A$ es de dimensión constante $(2\xi - 1) \times \xi$, porque ésta sólo depende de $\xi$.

Usamos $I^k$ para caracterizar las clases de $P$; para ello, buscamos si $I^k$ tiene algún *valor propio o parcialmente caracterizador* en alguna clase. Intuitivamente, los valores propios son *valores exclusivos* de la clase $C$ y

gráficamente son muy fáciles de identificar en un *boxplot múltiple*, quedando la misma información reflejada en la matriz o tabla de contingencia $A$.

La característica de un *valor propio o parcialmente caracterizador* de $I^k$ en la clase $C$ sobre la tabla de contingencia $A$ es tal que cumple:

$I_s^k$ es *valor propio o parcialmente caracterizador* de la clase $C$ si:

- $n_{sc} \neq 0$
- $\forall\, C' \neq C, \quad n_{sc'} = 0$

Si además

- $\forall\, s' \neq s, \quad n_{s'c} = 0$ entonces $I_s^k$ es un valor *totalmente caracterizador* de $C$.

Como en lo habitual se encuentran pocos valores *totalmente caracterizadores*, en sentido estricto lo común son los valores *propios o parcialmente caracterizadores*; es decir, valores que determinan parte de una clase, la cual tiene que cuantificarse para poder determinar el poder de caracterización de dichos valores.

Definimos $(1-\varepsilon), \varepsilon \in [0,1]$ como el grado de caracterización de una clase $C$, para un valor. En la referencia [10] aparece la idea de $(1-\varepsilon)$*– caracterización* y se maneja en todos los trabajos posteriores a nivel de variable. Ello conduce a situaciones en apariencia complejas como el hecho de que $X_k$ sea $(1-\varepsilon_1)$*– caracterizadora* de $C$ y también $(1-\varepsilon_2)$*– caracterizadora* de $C'$ con $\varepsilon_1 \neq \varepsilon_2$.

En realidad esto sucede porque lo que determina el poder de caracterización no es la variable en sí, sino los valores que toma y su distribución a través de las clases. Así, de ahora en adelante, trasladaremos a nivel de valores este análisis.

Así definimos, dada una variable $X_k$:

- Un valor $(1-\varepsilon)$*– caracterizador* de $C$ es aquel *valor propio* de $C$ que sólo identifica $(1-\varepsilon)$ % de $C$.

Existe aún una tercera situación, que corresponde al patrón que llamamos *valor caracterizador no propio*, el cual satisface la siguiente propiedad:

- $I_s^k$ es un *valor no propio* de la clase $C$ si cumple:
  $n_{sc} \neq 0 \wedge \forall s' \neq s \;\; n_{s'c} = 0$

Para analizar los valores concretos de $\varepsilon$ en la partición $P$ será necesario un análisis previo que pasará por la tabla de contingencia $A = I^k \times P$, entre otras cosas.

## 5. Construcción de la tabla de distribuciones condicionada a los intervalos

Es fácil construir ahora la tabla de distribuciones condicionada a los intervalos, como una matriz de números $B$, en la cual los renglones están representados por los intervalos $I^k$ encontrados anteriormente, y las columnas por las clases $C$ de la partición de referencia $P$, de modo que las casillas representen una estimación de la probabilidad de que un elemento $x_{ik}$ de un cierto intervalo $I_s^k$, pertenezca a una clase específica $C$.

Así, podemos representar la tabla de distribuciones condicionada como una matriz de la forma $B = I^k \times P$, cuyos valores toman la forma:

$$B = (p_{sc}(s = 1:2\xi - 1),(c = 1:\xi))$$

siendo $\xi$ = *card*($P$), $p_{sc}$ la frecuencia relativa de los individuos de valor $x_k \in I_s^k$ que se encuentran el la clase $C \in P$ y cuyo valor está dado por:

$$p_{sc} = n_{sc} / n_{I_s^k}, \text{ donde:}$$

$n_{sc}$ es el número de individuos que pertenecen al intervalo $I_s^k$ y a la clase $C$, y $n_{I_s^k} = \sum_{c=1}^{\xi} n_{sc}$ es el número total de individuos que se encuentran en el mismo intervalo $I_s^k$.

De acuerdo con la construcción de la tabla de distribuciones condicionada $B$, es posible afirmar que para los valores de la variable $I^k$ (renglones) en cada uno de los intervalos $I_s^k$, se tienen probabilidades $p_{sc}$ en el sentido frecuentista de que a un elemento de $I$ de valor $x_{ik}$ le sea asignada la clase $C$, cumpliendo con las siguientes propiedades:

i. $p_{sc} \in [0, 1]$

ii. $\sum_{i=1}^{\xi} p_{sc_i} = 1$

En la tabla de frecuencias condicionadas $B$, los *valores caracterizadores*, de la clase $C$ son todavía más fáciles de identificar, porque se detectan observando una sola casilla de la clase y pueden ser *parcialmente caracterizadores o totalmente caracterizadores* dependiendo de si existe o no la interacción entre clases. Así, tenemos que:

- Un valor $I_s^k$ de la clase $C$ es un *valor propio o parcialmente caracterizador* si su frecuencia es $p_{sc}$ = 1
- Un valor $I_s^k$ de la clase $C$ es un valor *totalmente caracterizador* si su frecuencia $p_{sc}$ =1 y $p_{s'c}$ = 0, $\forall\, s' \neq s$
- $I_s^k$ es un valor *caracterizador no propio* si $p_{sc} \in$ (0, 1)

Visto como se identifican los valores *caracterizadores*, vamos ahora a cuantificar al *grado de caracterización* tal y como ya se definió en el paso 5).

El valor $I_s^k$ de la variable $X_k$ será $(1-\varepsilon)$– *caracterizador* de $C$ si

$$n_{sc} = (1-\varepsilon)\cdot n_c$$

El *grado de caracterización* en este contexto, se interpreta como la parte proporcional (porcentaje) de individuos de $C$, cuyos valores de la variable $X_K$ se encuentran en el intervalo $I_s^k$.

**6. Generación del sistema de reglas $\Re(X_k, P)$**

Para cada *valor propio* (*total* o *parcial*) de la clase $C$, se puede extraer una regla que *identifica* la clase con el mínimo de información, de la forma:

$$\left(X_k \in \Lambda_c^k\right) \longrightarrow C$$

donde $X_k$ es la *k*-ésima variable y $\Lambda_c^k$ es el conjunto de *valores propios* de la clase $C$.

Ahora bien, si un valor es *caracterizador no propio* entonces, cuando se da ese valor, la clase de asignación puede ser una u otra con distintos grados de certeza, de ahí que la regla

$\left(X_k \in I_s^k\right) \longrightarrow C$, deje ser segura

Podemos definir $p_{sc}$ como el grado de certeza de esa regla, entendiendo que $p_{sc}$ (frecuencia relativa sobre la muestra) constituye una buena estimación puntual de la probabilidad de que un individuo $i$ que toma valores en ese intervalo $I_s^k$, pertenezca realmente a la clase $C$.

De aquí si $I_s^k$ es *un caracterizador no propio* de $C$, podemos generar una regla de la forma:

$$x_{ik} \in I_s^k \xrightarrow{\ p_{sc}\ } i \in C$$

donde $p_{sc}$ lo podemos definir en forma equivalente como una probabilidad condicional $P(C|I^k = I_s^k)$ de la siguiente manera:

$$p_{sc} = P(C|I^k = I_s^k) = card\{i \mid \mathrm{x_{ik}} \in I_s^k \wedge i \in C\}/\mathrm{n_{I_s^k}}$$

De hecho, $p_{sc}$ está indicando con qué probabilidad el elemento $i$ pertenece a la clase correcta $C$ a partir del valor $X_k$, considerando que existen otros individuos que toman valores en $I_s^k$ y se dispersan en las demás clases.

El esquema en la Tabla 4.1 establece la relación entre el conjunto antecedente $I_s^k$ donde se encuentra el valor de la variable $I^k$, la forma de la regla de asociación y el valor de su probabilidad de asignación $p_{sc}$ a la clase *C*.

| REGLA | CONJUNTO ANTECEDENTE | | PROBABILIDAD |
|---|---|---|---|
| | $I_s^k \subset C$ | $I_s^k = C$ | |
| $x_{ik} \in I_s^k \longrightarrow C$ | *propio parcialmente caracterizador* | *propio totalmente caracterizador* | $p_{sc} = 1$ |
| $x_{ik} \in I_s^k \xrightarrow{p_{sc}} C$ | | *total caracterizador no propio* | $p_{sc} \in (0, 1)$ |

Tabla 4.1: Relación entre reglas de asociación y valores propios

De ahí se observa que los valores propios siempre generan reglas seguras, pero el poder de caracterización depende de la cardinalidad del conjunto antecedente. Si éste coincide con toda la clase entonces hay una *caracterización completa* de la misma. De otra forma, es *parcial*.

Como se observa en la Tabla 4.1, ésta tiene una casilla vacía; esta casilla identifica un cuarto caso que corresponde a un otro patrón. Se trata de la situación más general que la llamamos *valor genérico* y que permite generar caracterizadoras parciales y no seguras, representando éste el caso más débil de todos. Así, definimos:

- Un valor $I_s^k$ de la variable $I^k$ es un *valor genérico* de la clase $C$ si:
  i) $p_{sc} \in (0, 1)$ y
  ii) $\exists s'$ tal que $p_{s'c} \neq 0, \quad s' \neq s,$ y
  iii) $\exists c'$ tal que $p_{sc'} \neq 0, \quad c' \neq c.$

Estos valores los podemos interpretar como el subconjunto de individuos *i* de la clase $C$ que comparten su valor $I_s^k$ con las demás clases, existiendo a su vez en la misma clase $C$ algunos otros elementos que pertenecen a otros intervalos.

A partir de los conceptos anteriores, se puede realizar la siguiente identificación, en relación con los *valores caracterizadores*:

- Si $I_s^k$ es el valor de la variable $I^k$ (intervalos de $X_k$), y $p_{sc} \in$ (0, 1] es su frecuencia condicionada para la clase $C$, entonces podemos generar para cada elemento de la Tabla $B$ reglas de la forma:

  $$\textit{Si } x_{ik} \in I_s^k \text{ para el elemento } i \xrightarrow{p_{sc}} i \in C$$

  donde: $x_{ik}$ es el valor de la *k-ésima* variable para el *i-ésimo* elemento, $I_s^k$ es el intervalo al que pertenece dicho valor y $C$ es la clase caracterizada a partir de $I_s^k$ con probabilidad $p_{sc}$.

Esta definición es general y cubre como casos particulares las reglas resultantes de los valores propios de $C$, que incluye los valores $p_{sc}$ = 1, que corresponden a las reglas seguras.

Por ello, para cada tabla de distribución condicionada a intervalos $B$ se puede derivar el siguiente sistema de reglas asociado a $X_k$ para identificar cierta partición $P$.

$$\mathcal{R}(X_k, P) = \{ r_l : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C \quad con\ p_{sc} > 0,\ p_{sc} \in B, \\ l = \{1, \ldots, (2\xi - 1)\xi\}, \quad s = \{1, \ldots, (2\xi - 1)\}\}$$

Este sistema ha de permitir identificar las distintas clases a partir de $X_k$.

Fijando una sola clase $C$ que se quiere caracterizar, las probabilidades de todas reglas que representan a $C$ como parte derecha pueden verse como una *distribución de posibilidades* [122] y [143] que asigna a cada valor de la variable $I^k$ su grado de pertenencia a la clase $C$ y que se representa como un gráfico (ver figura 4.1) con cada una de las funciones horizontales. Cabe mencionar que el área bajo las curvas que corresponden a estas funciones ya no es 1, puesto que se componen de probabilidades que provienen de distintas distribuciones condicionadas ($C \mid \mathrm{I} = \mathrm{I}_\mathrm{s}^\mathrm{k}, \forall s$).

Así definimos la función:

$$\pi_k^C(x_{ik}) \overset{def}{=} p_{sc}, \quad x_{ik} \in I_s^k$$

Para cada elemento de la partición $P$ (columnas de las matrices $A$ y $B$) que son las distintas clases, se tiene una distribución de posibilidad $\pi_K^C$, que

indica el *grado de compatibilidad* del valor de $X_k$ con la asignación a $C$. En esta distribución se tiene un número finito de *niveles de posibilidad de* $C$, distinguiendo valores entre lo "imposible" (codificado por 0) y lo "completamente posible" (codificado por 1).

A partir de lo anterior, se tiene que para toda $x_{ik} \in I_s^k$, $\pi_k^C(x_{ik})$ representa hasta qué punto es posible que cierto valor de $X_k$ implique la pertenencia a $C$.

La función $\pi_k^C$ representa una restricción flexible de los valores de la variable $X_k$ con las siguientes convenciones:

- $\pi_k^C(x_{ik}) = 0$, significa que la pertenencia a la clase $C$ es imposible;
- $\pi_k^C(x_{ik}) > 0$, significa que la pertenencia a la clase $C$ es posible a distintos grados (ejemplos: débil, fuerte, muy fuerte y otros) tanto más intenso cuanto más se acerque a 1, valor que representa la pertenencia segura.

Finalmente, se obtiene un sistema global que contiene reglas difusas o posibilistas, a partir del cual, para cierto valor de la variable $X_k$ se da, con mayor o menor grado, la pertenencia a cada clase de cierta partición de referencia $P$.

## 7. Descripción conceptual de las clases

La descripción conceptual de las clases es importante en la interpretación de las mismas y usar el conocimiento generado como herramienta de apoyo a la posterior toma de decisiones. Uno de los problemas principales de las técnicas de clustering es que la *validación* de la clasificación es un problema que tiene múltiples soluciones, ya que no existe un criterio objetivo para determinar la calidad de las clases de una clasificación. Es fácil evaluar un conjunto de clases en términos de *criterios de exactitud* siempre que exista una partición de referencia de los datos, asumiendo que la comparación es posible. Pero desafortunadamente, en la mayoría de las situaciones donde se requiere hacer clustering, la partición de referencia no existe y este enfoque no es útil: solamente la *utilidad* de una clasificación puede usarse para decidir si es correcta o no [21]. Evaluar la utilidad de una clasificación dada requiere de un mecanismo que permita comprender el significado de las clases identificadas, para finalmente decidir si son útiles o no.

Este proceso, conocido comúnmente como *Descripción Conceptual de las clases* resultantes, habitualmente lo realiza el analista informático en una *forma no sistemática*, usando sus conocimientos y experiencia para poner de manifiesto las principales diferencias entre clases; posteriormente, en estrecha colaboración con el experto en la materia, analiza las clases, estudia su significado y les da una *Interpretación*. Este proceso llega a dificultarse cuando

el número de clases aumenta y el número de variables utilizadas para describir los datos también aumenta.

Así, podemos decir que la validación de una clasificación se puede considerar como el grado de interpretabilidad o utilidad de ésta, sin ningún otro criterio que el de un especialista que observa y analiza las clases resultantes de una clasificación.

Teniendo, como base la tabla de distribuciones condicionadas a los intervalos analizados en el apartado paso 6), se puede asociar a un individuo cualquiera $i$ su grado de pertenencia a cada clase. Esto da lugar a un gráfico de grados de pertenencia difusos para cada clase y para cada variable, como se muestra en la figura 4.1. En el gráfico, el eje horizontal es común y representa el rango de $X_k$, y para cada clase se representa el grado de pertenencia de los valores de $X_k$ según las reglas. La forma escalonada de dichas funciones de pertenencia se debe a la categorización de $X_k$ en $I_k$. Así, dado un valor de $X_k$, se visualiza fácilmente su relación con las otras clases.

Se observa que a partir de esta representación gráfica, el paradigma difuso [52] constituye un excelente soporte al proceso de interpretación a través de un sistema de etiquetas lingüísticas para visualizar los resultados.

Lo anterior, porque el sistema $\Re(\mathrm{X_k}, \mathrm{P})$ contendrá reglas con el mismo antecedente $\left(I_s^k\right)$ y partes derechas diferentes (clases) con distintos grados de pertenencia. Por otro lado, una clase $C$ se reconoce por muchas reglas, lo que trae consigo problemas de imprecisión e incertidumbre en el modelo de razonamiento asociado a la caracterización de las clases. Esto es claramente visible en la representación gráfica de la Figura 4.1 y evidencia que se presenta una situación compleja que por sus características se presta a su contextualización en el paradigma de los conjuntos difusos [120-121], la lógica difusa y la teoría de la posibilidad; los que constituyen un excelente soporte para representar y manejar piezas de información, que contienen tanto la imprecisión como la incertidumbre, como es el caso en la determinación de la clase $C$ de un objeto $i \in \mathrm{I}$.

A partir de aquí, debemos fundamentar el proceso con un método de creación de etiquetas lingüísticas que genere descripciones conceptuales de las clases, con el siguiente del estilo:

Si la variable $X_k$ toma valores muy altos entonces ese objeto $i$ se asocia con $c_3$, donde, el grado de pertenencia de una variable específica $X_k$ al concepto “muy altos” vendría determinado precisamente por el gráfico de $c_3$ como el de la Figura 4.1. Así, una vez que se ha asignado la clase $C$ a un nuevo individuo, podemos analizar los gráficos de distribución variable por

variable para obtener conocimiento útil y comprensible en la interpretación conceptual de la clase identificada y su relación con otras clases.

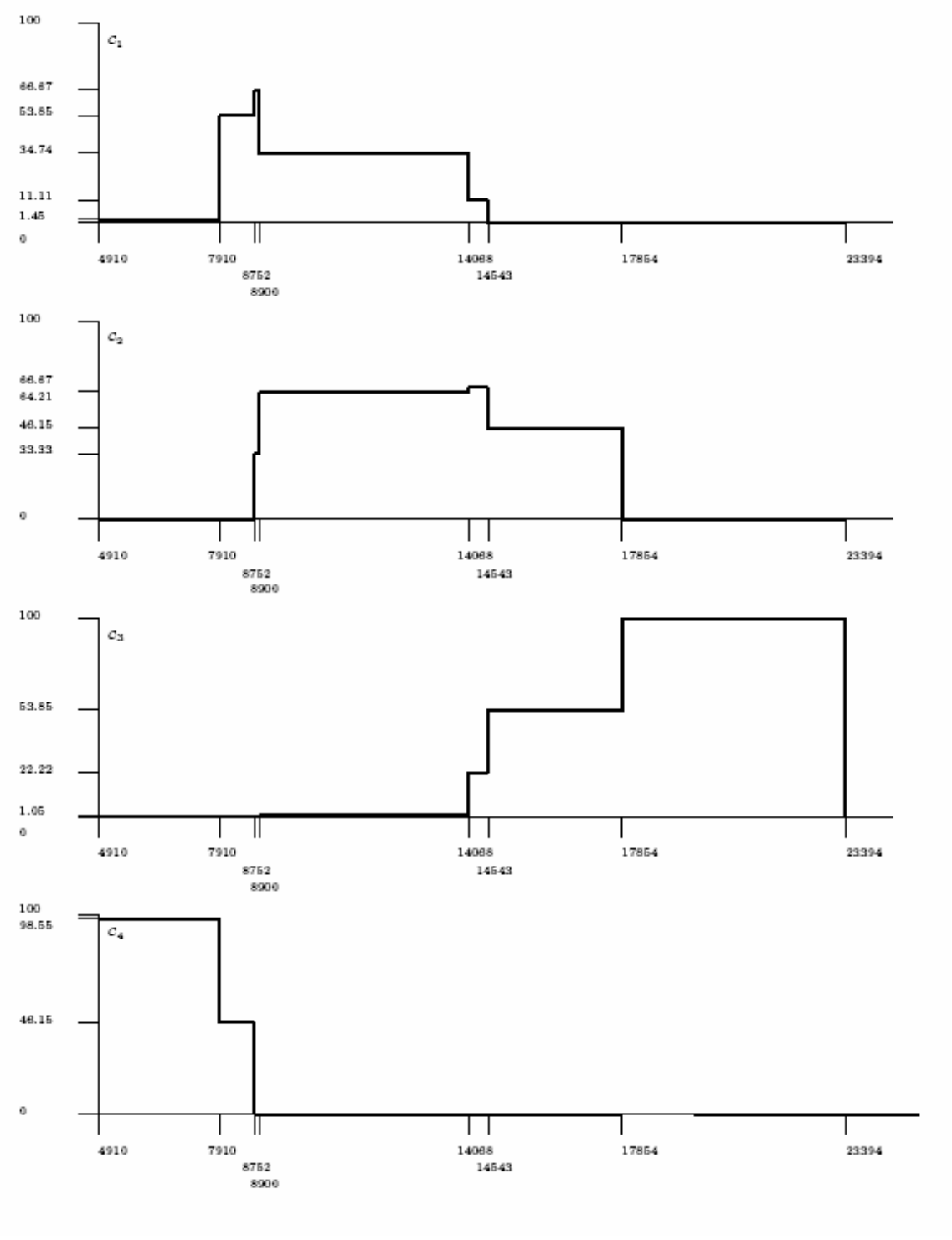



Figura 4.1 Diagrama de grados de pertenencia a las clases de la variable Q-AB

## 8. Validación del sistema de caracterización

En el modelo propuesto, el boxplot múltiple se usa como un elemento básico gráfico para la determinación de los *valores característicos*, considerándolo como la base del sistema de intervalos de longitud variable generado para cada variable $X_k$. Esto permite identificar cuál es la estructura natural que subyace en la base de datos del dominio de estudio, variable por variable. Esto ha permitido desarrollar un método rápido para construir un sistema de reglas difusas asociadas a cada variable $X_k$, el cual queda reflejado en la tabla de distribuciones condicionadas a intervalos $B = P \mid I^k$. Un primer propósito fue reducir la ambigüedad inherente al sistema de reglas $\Re(\mathrm{X_k}, \mathrm{P})$ considerando el criterio de grado más grande de asociación (con el consecuente de la regla con la probabilidad máxima, *PM*), el cual nos conduce a un sistema reducido $\Re^*(\mathrm{X_k}, \mathrm{P})$ mucho más pequeño en número de reglas, sin ambigüedad pero conservando incertidumbre.

Como una aplicación práctica, la evaluación del sistema de reglas consiste en considerar un conjunto de elementos de prueba $P_0$ y evaluarlos en el correspondiente sistema de reglas de la partición de referencias $P$. Así, considerando la variable $X_k$ y una participación de referencia $P$, tomamos cada valor $x_{ik}$ para toda *i* en el conjunto de prueba y lo evaluamos en el sistema de reglas reducido $\Re^*(\mathrm{X_k}, \mathrm{P})$.

En cada caso, para cada valor $x_{ik}$ se localizan los intervalos $I_s^k$, la clase $C$ y la probabilidad correspondiente $p_{sc}$; es decir, si existe una regla:

$$r : x_{ik} \in I_s^k \xrightarrow{Psc} C,$$

la clase $C$ se asigna al individuo $i$ con un grado de pertenencia $p_{sc}$ considerando únicamente la variable $X_k$. El resto de las variables se evalúan de igual forma.

Este proceso continúa hasta agotar todas las variables de todos los individuos en el conjunto de prueba $P_0$.

El siguiente paso es considerar otros criterios de agregación de información como: criterio de votación (Vot), suma máxima de probabilidades (Sum) que nos permitan un mejor desempeño en la clasificación de nuevos individuos; así, de acuerdo con el criterio de agregación de información elegido, la combinación de todas las variables por individuo del conjunto de entrenamiento $P_0$, determina el número de individuos mal clasificados y se calcula el error de predicción del sistema de reglas como un parámetro de validación del propio sistema de reglas generado.

# CAPÍTULO 5.

# RESULTADOS Y DISCUSIÓN

En este capítulo presentaremos la aplicación del modelo propuesto al dominio complejo de la base de datos de una planta depuradora de aguas residuales (WWTP) de la ciudad de Lloret en Cataluña, España, en donde se aplicó la metodología completa de este trabajo de tesis

## 5.1 CASO DE ESTUDIO. DOMINIO DE UNA PLANTA DEPURADORA DE AGUAS RESIDUALES

Las grandes áreas urbanas producen gran cantidad de aguas residuales, y cuando el medio ambiente está contaminado, la calidad del agua empeora debido a que el proceso residual llega a superar el desempeño del auto regulación de las aguas recibidas. En este caso, se deben tomar ciertas medidas provisorias para restaurar el equilibrio del medio ambiente [148].

Las plantas depuradoras de aguas residuales proporcionan un importante equilibrio entre el medio ambiente y las aguas residuales concentradas de las áreas urbanas. Si estas últimas se liberan de forma descontrolada, se degradaría el medio ambiente, elemento esencial para el bienestar de los seres humanos [44].

Para tratar adecuadamente las aguas residuales son necesarias distintas operaciones y procesos unitarios. El diagrama del proceso de una estación depuradora incluye diferentes combinaciones de agentes físicos, químicos y biológicos, cuyo proceso global se presenta en la Figura 5.1.1, donde se incluye un esquema típico, así como la secuencia lógica de tratamiento, dividida en diferentes fases, las que son resumidas brevemente a continuación; para mayor detalle referirse a [17] y [18].

El **pretratamiento** es la primera etapa para la depuración de aguas residuales. En esta fase, se realiza una primera separación de los sólidos, arrastrados por el agua residual cuando llega al recolector. Con ello se pretende evitar obstrucciones posteriores y otros problemas sobre las bombas ó válvulas utilizadas a lo largo de todo el proceso. Esta operación física se realiza mediante una secuencia de rejas, que se abren y cierran automáticamente.

El **tratamiento primario** corresponde a la segunda etapa del proceso. En esta fase, se deja reposar el agua en un tanque de sedimentación primaria, para que decante la materia orgánica sedimentable, el resto de la arena o partículas inorgánicas, que no se han retenido en el pretratamiento.

Posteriormente, se lleva a cabo la etapa más importante del proceso, conocido como **tratamiento secundario.** Aquí se degrada la materia orgánica disuelta en el agua residual, y esto ocurre por la acción de una población multi-específica de microorganismos, conocida como biomasa. La reacción que se

produce, tiene lugar en uno o varios reactores biológicos, dependiendo del número de reactores que tenga la planta. Finalmente, una nueva decantación se lleva a cabo en los sedimentadores secundarios, para emitir posteriormente el agua. El objetivo del proceso antes descrito, es conseguir una buena separación, entre el agua ya tratada y la biomasa.

Los sólidos sedimentados de ambas fases de decantación son enviados (purga) hacia una línea de tratamiento específico, conocida como “línea de barros” y eventualmente, realimentan la biomasa del reactor biológico.

Cuando la planta depuradora no funciona bajo condiciones normales, se deben tomar decisiones para modificar algunos parámetros del proceso de depuración y restablecer lo antes posible la normalidad. Para esto, es importante contar con un sistema automatizado que nos proporcione la información relevante sobre la situación que la planta tiene en un momento específico. *En nuestro caso, la aplicación de la metodología de esta tesis está orientada a hacer aportaciones en ese sentido.*

Como se mencionó anteriormente, un buen conocimiento sobre la situación de la planta en tiempo real constituye un excelente apoyo a la gestión de la misma. Por ello, el objetivo de esta aplicación es presentar el modelo expuesto en el capítulo 5 para la generación automática de descripciones conceptuales, que caractericen las distintas situaciones que se pueden presentar en un cierto día (registro promedio en distintos puntos) en la planta depuradora.

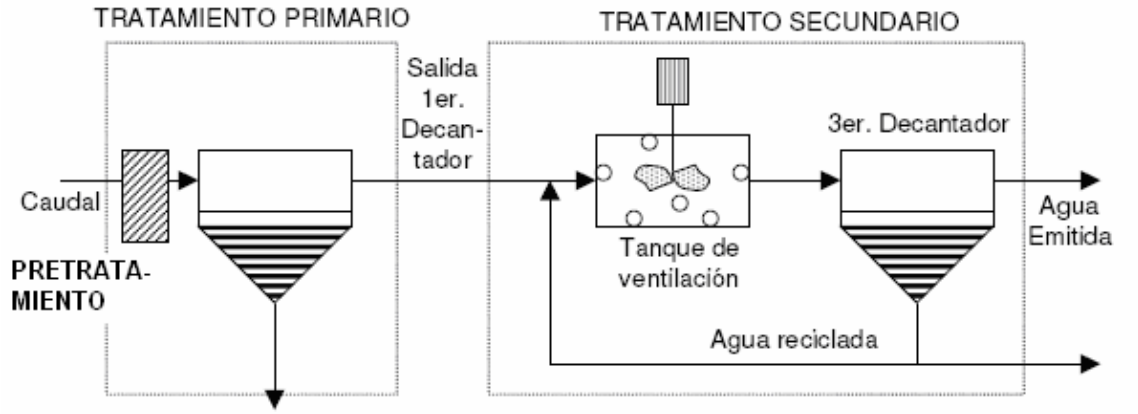



Figura 5.1.1 Diagrama típico del proceso de tratamiento de aguas residuales

Se partirá de dos clasificaciones de referencia para identificar situaciones típicas. A partir de estas clasificaciones, se propone un modelo conceptual que determine las variables relevantes involucradas en el proceso y describa e interprete las diversas situaciones que se presentan en un cierto día en cada una de ellas y mostrar que la predicción de clases depende de la partición de referencia [151].

### 5.1.1 Presentación de los datos de la planta depuradora de aguas residuales (WWTP)

Los datos analizados en esta aplicación provienen de una planta depuradora de la Costa Catalana (España), y están formados por un total de 243 observaciones, obtenidos consecutivamente el mismo número de días. Cada observación corresponde a la media diaria de repetidas mediciones sobre un conjunto de 63 variables. Cabe hacer notar que el autor participó en forma activa en la generación de la base de datos correspondiente, la cual se encuentra disponible en línea en la Universidad en Irvine en California [153].

El conjunto de datos con variables cuantitativas y cualitativas que se recogen en la planta depuradora, describe el estado de la planta a través de un conjunto de 63 variables, algunas de las cuales se midieron en distintos puntos de la planta (AB: a la entrada de la planta, SP1: después del primer decantador, B: en el reactor biológico, SP3: después del tercer decantador, AT: a la salida de la planta) y otras se obtuvieron por cálculos a partir de las primeras [15].

Los expertos recomiendan trabajar con un subconjunto de 19 variables, 17 de las cuales son numéricas a decir: Q-AB (caudal a la entrada de la planta), DQO-AB (materia orgánica química a la entrada), COND-AB (conductividad eléctrica a la entrada de la planta), DQO-SP1 (materia orgánica química en el primer decantador), Q-SP3 (caudal a la salida del tercer decantador), DQO-AT (materia orgánica química en el agua tratada, a la salida), SST-AT (total de sólidos en suspensión, a la salida), NH4-AT (amonio sobre el agua tratada), NO3-AT (nitrato sobre el agua tratada), IVF (índice volumétrico de fangos), CM (carga másica), ESC-B (presencia de espuma en el tanque de ventilación), ASP-AT (calidad del agua tratada), ZOO (Zooglea), NFILAM (número de bacterias filamentosas diferentes), BIODIV-MIC (biodiversidad de la micro fauna en el fango activo), P-FLAG (Flagelados > 20μm) y 2 variables categóricas, FILAM (bacteria filamentosa dominante) y FLOC (copos de fango activado). En este trabajo, hemos tomado como referencia estas variables, las cuales se representan como $X_k$, tanto las cuantitativas como las cualitativas.

Este conjunto de datos $I$ de 243 días ha sido previamente clasificado por la herramienta *Linneo*$^+$ y el sistema híbrido denominado *Klass*$^+$, y lo consideraremos como el conjunto de entrenamiento $T_0$; otro conjunto, de 25 nuevos individuos también previamente clasificados, será usado para validación de nuestro sistema de reglas y le denominaremos conjunto de prueba $P_0$.

### 5.1.2 Particiones de referencia: *Linneo*$^+$ y *Klass*$^+$

***Partición de Linneo***$^+$ $P_L$. El estado de la planta (la variable clase) fue previamente identificado por medio de un proceso de clasificación semi-automático usando la herramienta *Linneo*$^+$ y el criterio del experto [159].

*Linneo*$^+$, que es una herramienta de adquisición de conocimiento semi-

automática utilizada en la construcción de clasificaciones para dominios poco estructurados, fue el software utilizado para particionar los datos.

Después de un proceso iterativo de clasificación supervisada por el experto, los 243 individuos fueron clasificados en 20 situaciones típicas que ocurren en la planta, como se muestra en la Tabla 5.1.1. Estas 20 clases corresponden a los clusters obtenidos con la clasificación de *Linneo*$^+$, usando un radio igual a 10 excepto para dos clases no detectadas con este radio. Otras clasificaciones con diferentes radios descubrieron otros dos nuevos clusters que corresponden a dos estados de la planta.

| Situación Característica | Clase | días |
|---|---|---|
| Normal de la depuradora en días de invierno | c01 | 73 |
| normal de la depuradora en días de verano | c02 | 73 |
| días lluviosos | c03 | 3 |
| días de tormentas | c04 | 3 |
| carga baja | c05 | 10 |
| sobrecarga orgánica | c06 | 1 |
| nitrificación | c07 | 2 |
| desfloculación | c08 | 5 |
| aumento de sedimento debido a la triotixina | c09 | 3 |
| sedimento espumoso debido a la microtrixina con bidiversidad de microfausa | c10 | 15 |
| días de verano con operación óptima de la planta | c11 | 21 |
| aumento de cloro | c12 | 1 |
| desnitrificación en el segundo decantador | c13 | 6 |
| transición a un volumen de sedimento debido a triotixina | c14 | 2 |
| débil sedimento espumoso debido a la nocardia | c15 | 4 |
| Severo sedimento espumoso debido a la nocardia | c16 | 4 |
| Sedimento espumoso debido a la nocardia y la desfloculación | c17 | 7 |
| sedimento espumoso debido a la microtrixina con baja diversidad de microfauna | c18 | 1 |
| sedimento espumoso debido a la microtrixina y aumento de viscosidad debido a la zooglea | c19 | 5 |
| Cambio de configuración de la planta de invierno-verano | c20 | 2 |

Tabla 5.1.1 Clases obtenidas con la clasificación de *Linneo*$^+$

Aunque los expertos han dado una interpretación válida para la clasificación de *Linneo*$^+$ $P_L$, también se ha observado que como en todo proceso de clasificación automática, los propios expertos reconocen que no es la única y que incluso podría ser mejorada. La validación de esta clasificación no ha sido contrastada previamente por medios objetivos, razón por la cual se propone una segunda partición obtenida por *Klass*$^+$.

**Partición de *Klass*$^+$** $P_K$. El sistema *Klass*$^+$ [24] presenta diferencias importantes con respecto a otros clasificadores: el procesamiento de información simbólica y una metodología específica con restricciones declarativas, de ahí que, *Klass*+ se considera como una herramienta de ayuda a la adquisición de conocimiento, cubriendo un doble propósito:

- Implementar un método de clasificación con restricciones basado en el conocimiento.
- Una herramienta de ayuda a la adquisición de conocimiento basada en métodos estadísticos, orientada a la generación de reglas para un sistema de diagnóstico o predicción.

La base de la metodología de *Klass*$^+$ es un método de clasificación ascendente jerárquico, que utiliza el algoritmo de *vecinos recíprocos encadenados* [26]. La estrategia de clasificación consiste en detectar los pares de vecinos recíprocos que han sido fusionados y construir el árbol de agregaciones.

En esta aplicación Klass + la métrica mixta y el criterio de *Ward* [42] se han usado para clasificar el conjunto de entrenamiento $T_0$ usando además la metodología basada en reglas [38]. Básicamente, se realizan dos procesos de agrupamiento: uno por las reglas del experto y el otro para los objetos que no satisficieron las reglas del experto llamada clase residual. Ambas clasificaciones jerárquicas son integradas en una sola partición para el conjunto total $T_0$. La Figura 5.1.2 representa el dendograma final.

La clasificación de *Klass*$^+$ para los 243 individuos del conjunto de entrenamiento $T_0$ se hizo tomando en cuenta las reglas dadas por el experto y haciendo un corte del árbol igual a 20. Las clases que se obtuvieron son las que muestra la Figura 5.1.3.

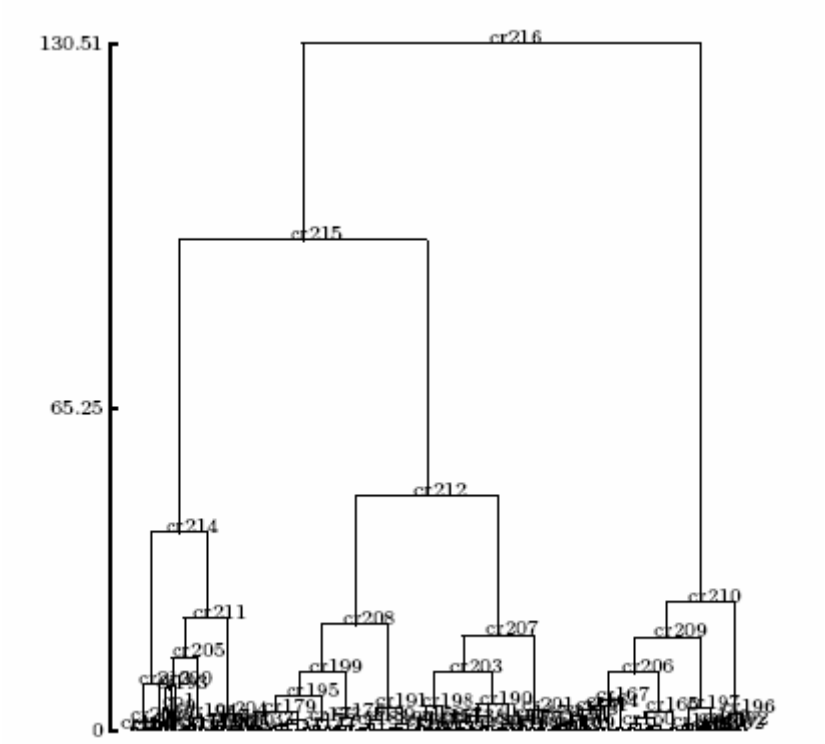



Figura 5.1.2 Árbol del clustering basado en reglas para un corte de 20 clases

$$\begin{aligned}
Clase\hat{C}01 &= Classer164, & Clase\hat{C}02 &= Classer179,\\
Clase\hat{C}03 &= Classer118, & Clase\hat{C}04 &= Classer176,\\
Clase\hat{C}05 &= Classer197, & Clase\hat{C}06 &= Classer140,\\
Clase\hat{C}07 &= Classer165, & Clase\hat{C}08 &= Classer196,\\
Clase\hat{C}09 &= ClasserCp1, & Clase\hat{C}10 &= Classer194,\\
Clase\hat{C}11 &= Classer191, & Clase\hat{C}12 &= D50,\\
Clase\hat{C}13 &= Classer204, & Clase\hat{C}14 &= Cs0,\\
Clase\hat{C}15 &= Classer170, & Clase\hat{C}16 &= Classer198,\\
Clase\hat{C}17 &= Classer202, & Clase\hat{C}18 &= D07,\\
Clase\hat{C}19 &= Classer201, & Clase\hat{C}20 &= Classer173
\end{aligned}$$

Figura 5.1.3 Clases obtenidas por *Klass*$^+$

**Comparación entre las particiones** *Linneo*$^+$ **y** *Klass*$^+$ $P_K$. Analizando los elementos que contienen cada una de las 20 clases en la clasificación de *Klass*$^+$ $P_K$ y comparándola con la obtenida con *Linneo*$^+$ $P_L$ se han obtenido los siguientes resultados:

De las 20 clases entre ambas clasificaciones, siete clases se identifican como muy similares, y las relaciones de estas clases entre ambas clasificaciones están dadas en la Tabla 5.1.2; entre las dos clasificaciones obtenemos una matriz donde los elementos de la diagonal, representan los elementos comunes entre las diferentes clases de ambas clasificaciones, teniendo un total de 49 objetos en clases similares, lo cual representa un 43.11 % de elementos coincidentes, y el resto de los objetos se dispersan en las otras clases, formando clases diferentes con características diferentes.

### 5.1.3 Análisis por variable

El sistema CIADEC (Caracterización e Interpretación Automática de Descripciones Conceptuales en Dominios Poco Estructurados usando Variables Numéricas) [54, 56], el cual es descrito en el Apéndice A, es un sistema que permite la caracterización e interpretación automática de descripciones conceptuales en dominios poco estructurados previamente clasificados, combinando conceptos y técnicas de estadística e inteligencia artificial y lógica difusa.

La automatización de esta metodología ofrece un conjunto de funcionalidades que permiten:

- Construir un sistema de reglas para la predicción de clases, diagnóstico de situaciones características
- Visualizar las funciones de pertenencia de la variable $X_k$ a las distintas clases
- Evaluar un conjunto de objetos nuevos de acuerdo con las reglas

generadas. Estimar la exactitud de asignación teniendo un conjunto de prueba $P_0$

| CLASES | $\hat{C}01$ | $\hat{C}02$ | $\hat{C}03$ | $\hat{C}04$ | $\hat{C}05$ | $\hat{C}06$ | $\hat{C}07$ | $\hat{C}08$ | $\hat{C}09$ | $\hat{C}10$ | $\hat{C}11$ | $\hat{C}12$ | $\hat{C}13$ | $\hat{C}14$ | $\hat{C}15$ | $\hat{C}16$ | $\hat{C}17$ | $\hat{C}18$ | $\hat{C}19$ | $\hat{C}20$ | $Klass^+$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C01 | *19* | | | | 6 | | 18 | 4 | 1 | 1 | | | 8 | | 5 | 5 | | 1 | 4 | | 72 |
| C02 | | *23* | | 3 | | 6 | | | | | 5 | | 1 | | 1 | 2 | | | 4 | 4 | 49 |
| C03 | | | *1* | 2 | | | | | | | | | | | | | | | | | 3 |
| C04 | | | | *2* | | | | | | | | | | | | | | | | | 2 |
| C05 | 1 | | 3 | | *4* | | | | | | | | 2 | | | | | | | | 10 |
| C06 | | | | | | | | | | | | | | | | | 1 | | | | 1 |
| C07 | | | | | 1 | | | 1 | | | | | | | | | | | | | 2 |
| C08 | | | | | | | | *2* | | | | | 1 | | | | 2 | | | | 5 |
| C09 | | | | | | | | | *2* | | | | | | | | | | | | 2 |
| C10 | 2 | | | | | 1 | | | | *8* | | | | 1 | | 4 | | | | | 16 |
| C11 | | | | 1 | | | | | | | *15* | | | | 6 | | | | | | 22 |
| C12 | | | | | | | | | | | | *1* | | | | | | | | | 1 |
| C13 | 1 | | | | | 1 | | | | | | | *2* | | 1 | | | | 1 | | 6 |
| C14 | | | | | | | | | | | | | | *1* | | | | | | 1 | 2 |
| C15 | | | | | | | 1 | 1 | | | | | | | *1* | 1 | | | | | 4 |
| C16 | | | | | | | | 1 | | | | | | | | *3* | | | 1 | | 5 |
| C17 | | | | | | | | | | 1 | | | | | | | *7* | | | | 8 |
| C18 | | | | | | | | 1 | | | | | | | | | | | | | 1 |
| C19 | | | | | 4 | | | | | | | | | | | | | | *1* | | 5 |
| C20 | | | | | | | | | | | | | | | | | | | | *2* | 2 |
| $Linneo^+$ | 23 | 23 | 4 | 8 | 15 | 8 | 19 | 10 | 3 | 10 | 20 | 1 | 14 | 2 | 14 | 15 | 10 | 1 | 11 | 7 | 218 |

Tabla 5.1.2 Comparación entre las particiones de *Linneo*$^+$ y *Klass*$^+$

Como una estrategia de trabajo, se aplicará la metodología haciendo el análisis para la variable DQO-AT (materia orgánica química en el agua tratada) y la partición de referencia dada por *Klass*$^+$, de tal forma que el método pueda seguirse de cerca. Esto se hace con el fin de ilustrar la ejecución de la metodología y permitir comentarios específicos en cada paso y posteriormente dar los resultados para el resto de las variables.

1. **Uso del boxplot múltiple como herramienta gráfica para la detección de variables caracterizadoras**

De acuerdo con el conjunto de datos obtenidos, lo primero que se realizó fue una descripción estadística, la cual permitió obtener información preliminar como: número de objetos pertenecientes a cada clase de la clasificación de referencia; la media, la mediana, la desviación estándar, los valores mínimos, máximos y atípicos (outliers) para cada variable, incluyendo DQO-AT, en cada una de las clases; un grupo de variables presento un 35 % aproximado de valores perdidos (NH4-AT, NO3-AT, IVF, CM y BIODIV-M), y cabe aclarar que estas son las que se miden con poca frecuencia en la clase, dado que en el resto de las variables se observó menos del 4 % de valores perdidos.

Con el boxplot múltiple, se visualiza la distribución de los valores de cada una de las variables por clases. En nuestro caso, el primer boxplot múltiple corresponde a la variable DQO-AT, es mostrado en la Figura 5.1.4, y consiste de una representación gráfica que muestra cómo los valores de la variable por clases se distribuyen. En cada boxplot por clase, los valores atípicos (outliers) son marcados por "*", y se despliega una caja desde Q1 (primer quartil) hasta el Q3 (tercer quartil) e incluye un 50 % de elementos de la clase; la mediana de los valores está marcada dentro de la clase con un signo horizontal y los "bigotes" se extienden hasta el mínimo y el máximo por clase.

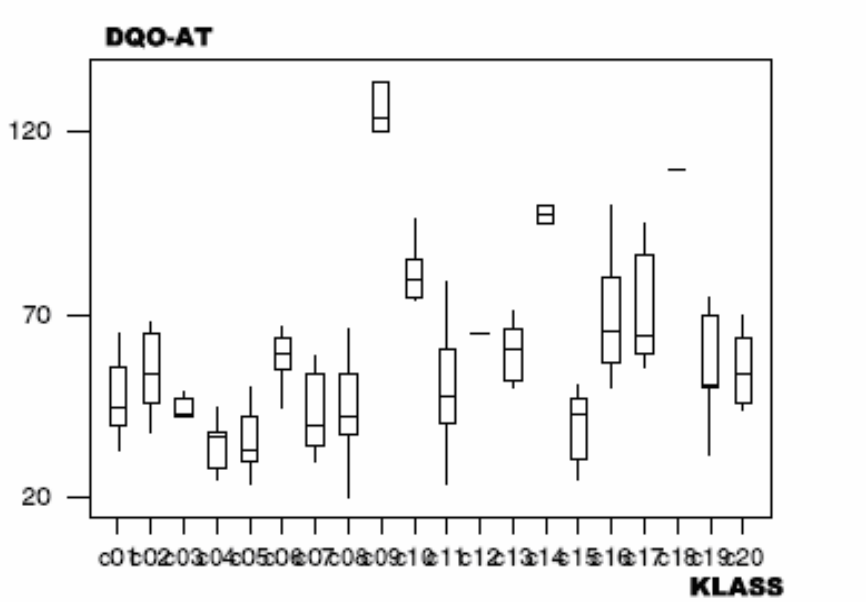



Figura 5.1.4 El Boxplot para la variable DQO-AT

A partir de la observación del boxplot se determinan los *valores caracterizadores,* en sentido estricto de las clases. Gráficamente, se puede apreciar si la proyección horizontal del boxplot de cierta clase no interseca con la de las demás; en un caso así, la variable es *totalmente caracterizadora* de esa clase.

Si observamos el boxplot de la Figura 5.1.4, cualquier valor en el intervalo (120,134] mg/l es *totalmente caracterizador* para la clase Ĉ09. Esto se debe a que en ese intervalo ningún otro boxplot interseca con el boxplot de la clase Ĉ09 y todos sus valores están comprendidos en dicho intervalo. Lo anterior significa que cualquier día en donde la variable DQO-AT (materia orgánica química) tome un valor en el intervalo (120,134] mg/l estará en la clase Ĉ09 y viceversa: todos los días que se encuentran en la clase Ĉ09 contienen un valor de DQO-AT entre (120,134] mg/l. Si se observa un valor en el intervalo (100, 110], se puede notar que ese valor es *totalmente caracterizador* para la clase Ĉ18 de la misma variable.

## 2. Determinación de la interacción entre clases para cada variable

Se observa en la Figura 5.1.4 que existen intervalos de la variable DQO-AT donde pueden coincidir Ĉ02, Ĉ06, Ĉ07, Ĉ11, Ĉ13, Ĉ16, y Ĉ17, como es el valor de DQO−AT = 57 mg/l, el cual se encuentra en el intervalo $I_{19}^{DQO-AT}$; u otros donde intersecan las clases Ĉ05, Ĉ07, Ĉ15 y Ĉ19, como es el caso para el valor de 31 mg/l localizado en el intervalo $I_6^{DQO-AT} = (30,32]$ mg/l. Queda claro que el poder informativo asociado al valor de la variable DQO-AT depende directamente de la cardinalidad del conjunto de clases que interseca.

## 3. Sistema de intervalos o ventanas de longitud variable

La generación de los intervalos o ventanas de diferentes longitudes, se realiza tomando los puntos de corte contiguos dos a dos del conjunto $Z^k$. Esto dio como resultado un sistema de intervalos abiertos por la izquierda y cerrados por la derecha, excepto el primer intervalo en cada variable, el que se considera cerrado por ambos lados. Esta forma de presentar los intervalos se debe a las características propias de la herramienta utilizada. Con ello se dispone de una variable categórica $I^{DQO-AT}$ asociada a DQO-AT, que indica todas las intersecciones entre clases. Para la variable DQO-AT, tenemos el siguiente sistema de intervalos:

$$
\begin{array}{rrr}
I^{DQO-AT} = & \{I_1^k = [20,24), & I_2^k = [24,24], \\
I_3^k = & (24,25], & I_4^k = (25,25], \quad I_5^k = (25,30], \\
& \ldots, & \ldots, \quad \ldots, \\
& \ldots, & \ldots, \quad \ldots, \\
I_{34}^k = (96,100], & I_{35}^k = (100,100], & I_{36}^k = (100,110], \\
I_{37}^k = (110,110], & I_{38}^k = (110,120], & I_{39}^k = (120,134]\}
\end{array}
$$

## 4. Construcción de la tabla de contingencia de Clases vs Intervalos

Una vez obtenido el sistema de intervalos, se construye la tabla de contingencia entre los intervalos y las clases. En los renglones marcamos los intervalos y en las columnas las clases; ya se dijo que las intersecciones de renglón con columna contienen el número de observaciones que hay en cada clase para cada intervalo. Tabla 5.1.3

Obsérvese que $I^{DQO-AT}$ representa una categorización *no arbitraria* de DQO-AT, que hace *emerger todos* los puntos donde cambian las intersecciones entre clases. Así, mientras el valor de $I^{DQO-AT}$, indica el intervalo [20, 24] ml/l de DQO-AT, también está indicando *la zona en la que se da la clase* Ĉ08. Por el modo como se ha construido $I^{DQO-AT}$ se puede notar que precisamente a partir de 24 mg/l de materia orgánica serán otras

las clases que se pueden dar simultáneamente.

| ***A*** | ***C₁*** | ***C₂*** | ***.....*** | ***C_c*** | ***.....*** | ***.....*** | ***C_ξ*** |
|---|---|---|---|---|---|---|---|
| $I_1^k$ | $n_{11}$ | $n_{12}$ | | $n_{1i}$ | | | $n_{1\xi}$ |
| $I_2^k$ | $n_{21}$ | $n_{22}$ | | $n_{2c}$ | | | $n_{2\xi}$ |
| ..... | .... | ..... | ..... | ..... | ..... | ..... | ..... |
| ..... | ..... | ..... | ..... | ..... | ..... | .... | .... |
| $I_s^k$ | $n_{s1}$ | $n_{s2}$ | | $n_{sc}$ | | | $n_{s\xi}$ |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| $I_{2\xi-1}^k$ | $n_{(2\xi-1)1}$ | $n_{(2\xi-1)2}$ | | | | | $n_{(2\xi-1)\xi}$ |

Número de elementos de $C$ cuyos valores de $X_k \in I_s^k$

Tabla 5.1.3 Tabla de contingencia de intervalos por clase

## 5. Construcción de la tabla de distribuciones condicionadas a los intervalos

Una vez obtenida la tabla de contingencia, se construye la tabla de distribuciones condicionadas a intervalos para la variable de estudio DQO-AT, ilustrada a continuación en la Tabla 5.1.4.

| ***B*** | ***C₁*** | ***C₂*** | ***.....*** | ***C_c*** | ***.....*** | ***.....*** | ***C_ξ*** |
|---|---|---|---|---|---|---|---|
| $I_1^k$ | $p_{11}$ | $p_{12}$ | | $p_{1c}$ | | | $p_{1\xi}$ |
| $I_2^k$ | $p_{21}$ | $p_{22}$ | | $p_{2c}$ | | | $p_{2\xi}$ |
| ..... | .... | ..... | ..... | ..... | ..... | ..... | ..... |
| ..... | ..... | ..... | ..... | ..... | ..... | .... | .... |
| $I_s^k$ | $p_{s1}$ | $p_{s2}$ | | $p_{sc}$ | | | $p_{s\xi}$ |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| $I_{2\xi-1}^k$ | $p_{(2\xi-1)1}$ | $p_{(2\xi-1)2}$ | | | | | $p_{(2\xi-1)\xi}$ |

Probabilidad de los objetos de valor $X_k \in I_s^k$ y que están en $C$

Tabla 5.1.4 Tabla de distribuciones condicionadas a intervalos

A partir de esta tabla, para cada valor $x_{iDQO-AT}$ se le asigna una probabilidad $p_{sc}$, $s = 1,.....,2\xi - 1,$ que representa el grado de pertenencia del individuo $i$ a la clase $C$ , de acuerdo con esta variable; y las probabilidades asociadas a la clase $C$ pueden verse como una distribución de posibilidades, que asigna a cada valor de la variable $I^k$ su grado de pertenencia a la clase $C$ y dicha distribución puede representarse por medio de un gráfico como se muestra en la Figura 5.1.5.

Para la variable de estudio DQO-AT se pueden reconocer los cuatro tipos de valores característicos:

- Valores propios *totalmente caracterizadores*:
  $I_{36}^{DQO-AT} = (100, 110]$ de $\hat{C}18$ y
  $I_{39}^{DQO-AT} = (120, 134]$ de $\hat{C}09$
- Valores propios *parcialmente caracterizadores* :
  $I_{28}^{DQO-AT} = (71, 74]$ de $\hat{C}10$
  $I_{33}^{DQO-AT} = (95, 96]$ de $\hat{C}10$
- Valores *No propios* son:
  $I_{20}^{DQO-AT} = (59, 65]$ de $\hat{C}12$
- Y, los valores *Genéricos*: El resto de los valores no nulos en la tabla de distribución son valores genéricos.

## 6. Generación del sistema de reglas $\Re(X_k, P)$

Interpretando $p_{sc}$ como una estimación de la probabilidad de que $i \in I^k(C \mid I^k)$, podemos obtener un sistema de reglas que represente los grados de pertenencia de un día a cada clase de acuerdo con una variable dada.

Considerando las distribuciones condicionadas a los intervalos como distribuciones de posibilidad (grado de imprecisión), podemos asociar a un objeto (día) cualquiera, su(s) grado(s) de pertenencia a la(s) clase(s) y obtener un sistema de reglas $\Re(X_k, P)$, que represente el grado de pertenencia a algunas de las clases. Esto da lugar a un gráfico de grados de pertenencia para cada clase y para cada variable, el cual representa una buena ilustración de lo que ocurre en una situación real. Así, si en la Figura 5.1.5 tomamos el valor de $x_{iDQO-AT}$ = 93 mg/l y trazamos una línea vertical sobre él, se obtienen los grados de pertenencia de este valor a cada una de las clases. Esta forma de representar gráficamente este sistema permite obtener conocimiento útil y comprensible para la *descripción conceptual* de las clases identificadas.

Con respecto a la variable DQO-AT, esta etapa del proceso genera un sistema de reglas $\Re(DQO-AT, P)$ de 123 reglas, una por cada celda no

nula en la matriz de distribuciones $B$ condicionadas a intervalos para dicha variable. Como la mayoría de los intervalos presenta diferentes grados de pertenencia a diferentes clases, esto genera un número de reglas con diferentes consecuentes dentro del mismo intervalo. Por ejemplo, si la

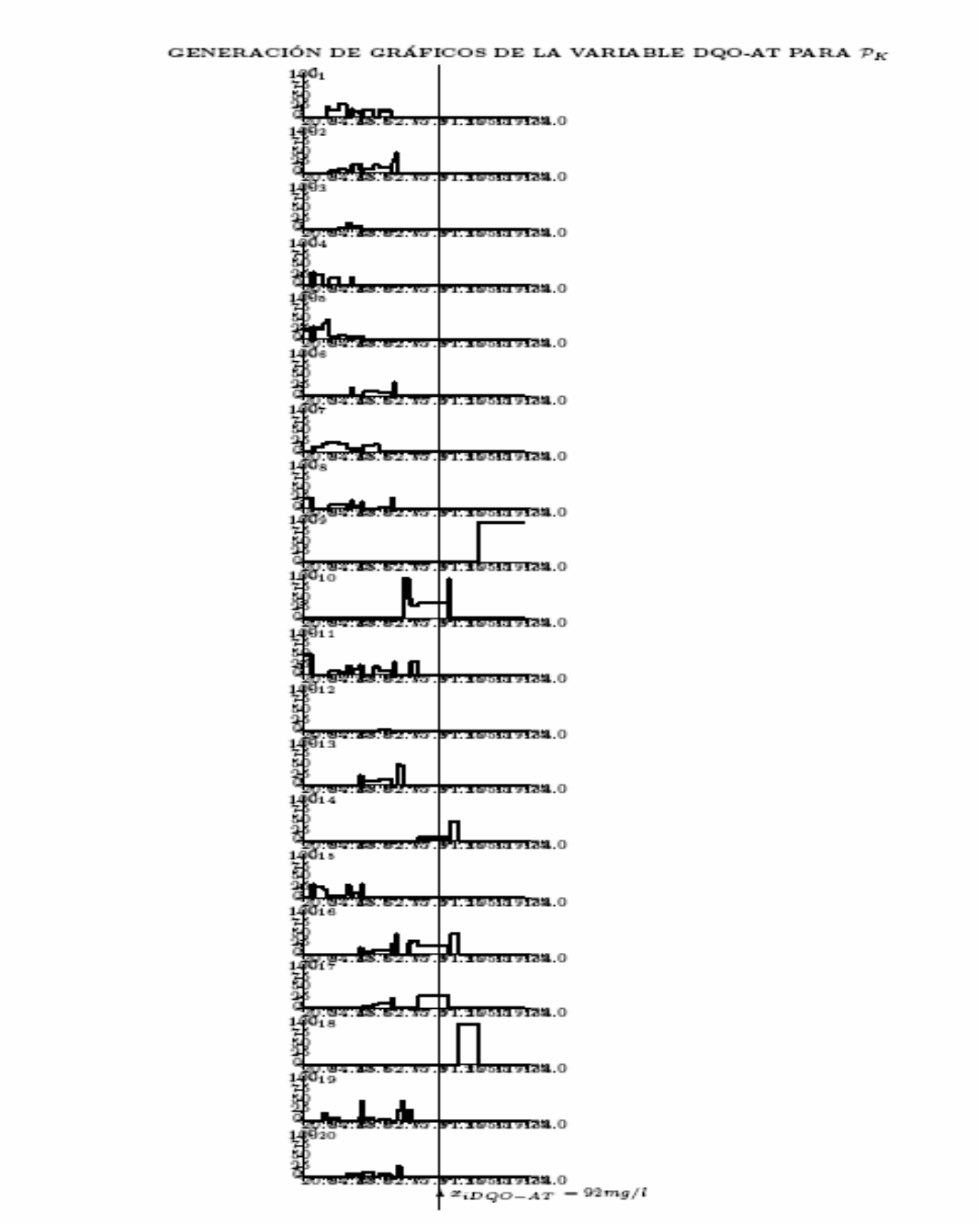



Figura 5.1.5 Gráfico de las funciones de pertenencia
para la variable DQO-AT

variable DQO-AT toma el valor de 92.6 mg/l, éste se localiza en el intervalo $I_{31}^{DQO-AT} = (79,95]$ y satisface cuatro reglas en el sistema global de reglas con

diferentes grados de pertenencia. En este caso particular, el grado de pertenencia a la clase Ĉ10 es 0.40, a la clase Ĉ14 es 0.10, a la clase Ĉ16 es 0.20, a la clase Ĉ17 es 0.30 y para el resto de las clases es 0.0; por lo tanto, hay cuatro reglas para asignar clases en este día, de acuerdo con el nivel de DQO-AT. En notación de cálculo de predicados de primer orden, las expresamos de la siguiente manera:

$$\text{Si } x_{ik} \in (79,95] \xrightarrow{0,40} i \in \hat{C}10, \qquad \text{Si } x_{ik} \in (79,95] \xrightarrow{0,10} i \in \hat{C}14$$
$$\text{Si } x_{ik} \in (79,95] \xrightarrow{0,20} i \in \hat{C}16, \qquad \text{Si} x_{ik} \in (79,95] \xrightarrow{0,30} i \in \hat{C}17$$

Esto presenta una situación ambigua, y la decisión de asignación de clase puede llevar a errores. Como una primera aproximación al proceso de tomar una decisión, proponemos reducir el conjunto de reglas de cada intervalo $I^k$ a sólo una regla, siguiendo el criterio del modelo clásico de razonamiento aproximado para sistemas de clasificación difusa, respecto a seleccionar la regla que presente probabilidad máxima en cada intervalo [130]. Esto corresponde a un criterio de agregación de información muy fuerte, que elimina la modelación difusa que tanto hemos defendido, con su consiguiente pérdida de información. Sobre esta decisión convendrá hacer un análisis a profundidad más adelante, pero de momento permite reducir la ambigüedad del sistema de reglas resultante, que llamaremos *Sistema Reducido de Reglas* $\Re^*(DQO-AT,P)$.

Evidentemente en este sistema de reglas hay como máximo una regla por intervalo, con lo que un conjunto de $card(\Re(X_k,P)) = (2\xi - 1)\xi,$ llega a ser de $card(\Re^*(X_k,P)) = (2\xi - 1).$

## 7. Descripciones conceptuales de las clases resultantes

Hemos mencionado que un método de apoyo a la generación de interpretaciones conceptuales es el uso de etiquetas lingüísticas [144] que nos permita dar el significado de las clases en forma natural. Según el gráfico generado para la variable DQO-AT, si tomamos, en la Figura 5.1.5, el valor de $x_{ik}^{DQO-AT} = 92.6 mg/l$ y trazamos una línea vertical sobre él, obtendremos el grado de pertenencia de este valor a cada clase. Para el ejemplo, obtenemos que el grado de pertenencia a la clase Ĉ10 es 0.40 %, el grado de pertenencia a la clase Ĉ14 es 0.10 %, el grado de pertenencia a la clase Ĉ16 es 0.20 %, el grado de pertenencia a la Ĉ17 es de 0.30 %, ya determinadas con el sistema de reglas del inciso 7, y para el resto de las clases es cero.

Además, observando el gráfico tenemos que los valores altos para la materia química orgánica (DQO-AT) a la salida de la planta se da en las clases Ĉ09, y Ĉ18, valores intermedios en las clases Ĉ02, Ĉ 05, Ĉ10, Ĉ13, Ĉ16, Ĉ17 y Ĉ19, valores bajos en el resto de las clases. Así, de esta forma podemos generar descripciones conceptuales de las clases.

- Si la variable materia química orgánica de salida toma valores altos entonces ese día se asocia a Ĉ09, donde el grado de pertenencia de la materia química orgánica concreto al concepto valores “altos” vendría dado por la función de pertenencia de la clase, digamos Ĉ09 en la Figura 5.1.5.

Como podemos darnos, cuenta la descripción conceptual por variable es un conocimiento parcial que poco nos ayuda, siendo más importante considerar la contribución de todas los demás variables para la caracterización e interpretación de clases para nuevos objetos. De esta forma terminamos la aplicación de la metodología por variable para considerar el análisis multivariable.

## 8. Validación del sistema de caracterización

En esta parte de la metodología consideraremos el conjunto de entrenamiento que ha sido previamente clasificado, tanto por *Linneo*$^{+}$ como por *Klass+*, para obtener el desempeño del sistema de reglas obtenido para la variable DQO-AT. El proceso se ha descrito en el capítulo 4.

En la Tabla 5.1.5 se compara la clase real de cada elemento con la asignada por las reglas, resumiendo el número de objetos que coinciden en ambas particiones. Los elementos asignados correctamente se ubican en la diagonal de esta tabla enmarcados en cursiva y el resto, representa errores de clasificación. Es importante hacer notar lo siguiente:

- Estamos tratando únicamente una variable con independencia de las demás
- La desambiguación es muy fuerte (grado máximo) y no toma en cuenta la probabilidad de las casillas “error”

| REF | $\hat{P}01$ | $\hat{P}02$ | $\hat{P}04$ | $\hat{P}05$ | $\hat{P}07$ | $\hat{P}09$ | $\hat{P}10$ | $\hat{P}11$ | $\hat{P}13$ | $\hat{P}15$ | $\hat{P}18$ | $\hat{P}19$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{C}01$ | *15* | 2 | | 1 | 4 | | | 1 | | | | |
| $\hat{C}02$ | 9 | *18* | | | 1 | | | 1 | | 1 | | |
| $\hat{C}03$ | 1 | 1 | | | | | | | | 2 | | |
| $\hat{C}04$ | 1 | | *2* | | 4 | | | | | | | |
| $\hat{C}05$ | 2 | 1 | 3 | *4* | 2 | | | | | 1 | | |
| $\hat{C}06$ | 6 | | | | | | | | | | | |
| $\hat{C}07$ | 7 | 3 | 1 | 2 | *5* | | | | | 1 | | |
| $\hat{C}08$ | 4 | | | | 2 | | | 1 | | | | 1 |
| $\hat{C}09$ | | | | | | *3* | | | | | | |
| $\hat{C}10$ | | | | | | | *9* | | | | | |
| $\hat{C}11$ | 4 | 3 | | | 2 | | | *10* | | 3 | | |
| $\hat{C}12$ | 1 | | | | | | | | | | | |
| $\hat{C}13$ | 7 | 1 | | | | | | 3 | *3* | | | |
| $\hat{C}14$ | | | | | | | 1 | | | | | |
| $\hat{C}15$ | 1 | | 3 | 1 | 1 | | | | | *5* | | 2 |
| $\hat{C}16$ | 4 | 3 | | | | | 4 | | | | | |
| $\hat{C}17$ | 5 | | | | | | 3 | | | | | |
| $\hat{C}18$ | | | | | | 1 | | | | | *1* | |
| $\hat{C}19$ | 2 | | | 1 | 1 | | 1 | | 2 | | | *3* |
| $\hat{C}20$ | 4 | 1 | | | | | | | 1 | 1 | | |

Tabla 5.1.5 Incidencia de los elementos (días) en la clase de referencia y la predicción para la variable DQO-AT

### 5.1.4 Análisis multivariable

En esta fase de la aplicación al caso de estudio consideraremos la contribución de información, que cada una de las variables en consideración tiene en la asignación de la clase para un objeto nuevo del conjunto de prueba.

Consideremos el análisis de todas las variables en forma conjunta. Por ejemplo, tomemos el objeto $i = 23$ (de $P_0$) que, de acuerdo con partición de *Klass*$^+$ se le asignó la clase Ĉ01; en el análisis por variable y tomando como criterio de agregación el de probabilidad máxima (*PM*), se tiene que para la variable DQO-AT su valor es $x_{iDQO-AT} = 55$, localizado en el intervalo $I_{17}^{DQO-AT}$, y la clase de predicción en el consecuente es Ĉ01 con una probabilidad de 0.18. Con respecto a la variable SST-AT, su valor es $x_{iSST-AT} = 5.6$, el cual se localiza en el intervalo $I_{11}^{DQO-AT}$, y de acuerdo con el sistema de reglas se le asigna la clase Ĉ11 con una probabilidad de 0.15.

Con respecto a la variable Q-AB, su valor es $x_{iQ-AB} = 9732$, en el intervalo $I_{17}^{Q-AB}$ y la parte derecha de la correspondiente regla al criterio de agregación ya establecido es la clase Ĉ01 con una probabilidad de 0.25. Para la variable Q-SP1, su valor $x_{iQ-SP1} = 9732$, en el intervalo $I_{27}^{Q-SP1}$, la clase asignada es Ĉ02 con una probabilidad de 0.60, y para el resto de las variables se procede de forma similar.

Este proceso puede realizarse sobre cada una de las variables. Sin embargo, genera un diagnóstico aparentemente inconsistente ya que esta primera aproximación no considera los grados de certeza de los demás variables. Consideremos el ejemplo anteriormente expuesto; la Tabla 5.1.6 resume lo que ocurre sobre el objeto (día) $i = 23$ del conjunto $P_0$ y cuatro variables de acuerdo con $\Re^*(X_k, P)$. La clase de referencia para $i = 23$ es Ĉ01, la cual es reconocida por tres de las cuatro variables consideradas.

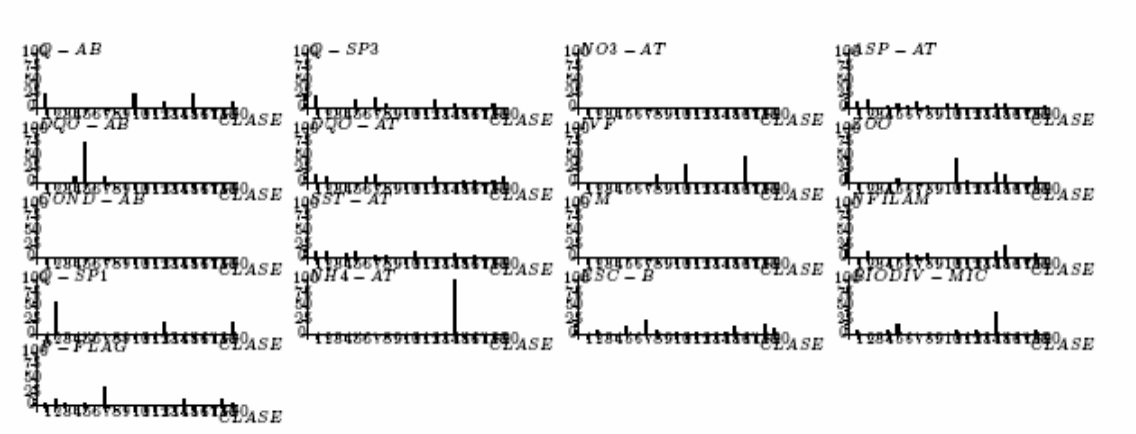



Figura 5.1.6 Gráfico de asignación de *clase* |*variable* para elemento $i = 23$ de prueba $P_0$

En el ejemplo se observa que diferentes reglas (probabilidad máxima) se disparan con consecuentes diferentes, nuevamente presentando el problema de ambigüedad de asignación de clases a nivel de variables.

Como se ha mencionado anteriormente, en este proceso sólo una regla por variable se satisface (grado máximo de pertenencia) usando el criterio para

desambiguar la confusión. La Figura 5.1.6 muestra la representación gráfica de la asignación $clase \mid \text{var}iable$ para el objeto $i = 23$ del conjunto de prueba $P_0$; estas representaciones gráficas de las reglas disparadas en todas las variables por individuo, permiten seleccionar criterios de agregación más adecuados.

Sin embargo, si trabajamos directamente con el sistema total de reglas $\Re(X_k, P)$ usando los nuevos criterios de agregación propuestos en este trabajo de tesis, en contra de lo esperado mejora la asignación de la clase. Nótese cómo se obtiene Ĉ01 en algunas de las reglas de las variables consideradas y estas reglas presentan grados de pertenencia razonables muy cercanos a la regla que resulta en $\Re^*(X_k, P)$.

| | | $\Re^*(X_k, P)$ | | $\Re(X_k, \mathcal{P})$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Atributo | Valor | $\hat{C}$ | $P$ | $\hat{C}$ | $P$ | $\hat{C}$ | $P$ | $\hat{C}$ | $P$ | $\hat{C}$ | $P$ |
| DQO-AT | 55 | $\hat{C}01$ | 0,18 | $\hat{C}02$ | 0,12 | $\hat{C}06$ | 0,12 | $\hat{C}07$ | 0,18 | $\hat{C}13$ | 0,12 ... |
| SST-AT | 5.6 | $\hat{C}11$ | 0,15 | $\hat{C}01$ | 0,13 | $\hat{C}02$ | 0,15 | $\hat{C}03$ | 0,02 | $\hat{C}04$ | 0,11 ... |
| Q-AB | 9732 | $\hat{C}01$ | 0,25 | $\hat{C}10$ | 0,25 | $\hat{C}13$ | 0,13 | $\hat{C}16$ | 0,25 | $\hat{C}20$ | 0,12 |
| Q-SP1 | 9732 | $\hat{C}02$ | 0,60 | $\hat{C}13$ | 0,20 | $\hat{C}20$ | 0,20 | | | | |

$\hat{C}$ clase de la $\mathcal{P}_K$ $P$ Probabilidad

Tabla 5.1.6 Asignación de clases para diferentes variables del elemento $i = 23$ para $P_0$.

Así, el proceso de validación del sistema total de reglas consiste en: a partir de un conjunto de prueba $P_0$ previamente clasificado, medir la estimación de exactitud de asignación de las clases (por ciento de objetos clasificados correctamente sobre el total de ellos) a los objetos nuevos, considerando un análisis en el cual un criterio de agregación de información de las variables se tome en forma conjunta. Esta es para predecir la clase de cada uno de los objetos nuevos del conjunto de prueba y estimación de exactitud de esta predicción. Del análisis de la Tabla 5.1.7, observamos que no todas las variables conducen a la misma clase de predicción. Por ejemplo, el objeto $i = 2$ tiene asignada la clase Ĉ07 con una probabilidad de 0.30 para la variable Q-AB y, también tiene asignada la clase Ĉ02 con una probabilidad de 0.20 para la variable DQO- AT, y la clase Ĉ07 con una probabilidad de 0.161, así sucesivamente. Nuevamente, tomaremos el criterio de probabilidad máxima para resolver el conflicto de asignación de clase para los objetos de $P_0$.

Comparando la clase real y la asignada por las reglas para cada uno de los elementos del conjunto de prueba $P_0$, se observa que un 55 % de días fueron bien clasificados.

De esta forma podemos hacer una estimación de la exactitud de la predicción. Así, tenemos que siete objetos de $P_0$: $i = 1$, 3, 8, 9, 10, 11, 14, 15, 17, 24, y 25 fueron mal clasificados con respecto a la partición de *Klass*$^+$, teniendo un error del 45 % y una estimación de la exactitud de la predicción del 55 %.

Para la partición de referencia de *Linneo*$^+$ se hizo un análisis similar, obteniendo un error global del 40 % y una estimación de la exactitud de la predicción del 60 %.

| $P_0$ | Q-AB | | DQO-AB | | COND-AB | | ... | ... | BIODIV | | P-FLAG | | Classes | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $i$ | $\hat{C}$ | P | $\hat{C}$ | P | $\hat{C}$ | P | $\hat{C}$ P | $\hat{C}$ P | $\hat{C}$ | P | $\hat{C}$ | P | $\hat{C}$K | $\hat{P}$PM |
| 1 | $\hat{C}$01 | .286 | $\hat{C}$5 | .20 | $\hat{C}$01 | .20 | .. .. | .. .. | * | * | $\hat{C}$01 | .129 | $\hat{C}$01 | $\hat{P}$15 |
| 2 | $\hat{C}$07 | .30 | $\hat{C}$02 | .20 | $\hat{C}$07 | .167 | .. .. | .. .. | $\hat{C}$20 | .154 | $\hat{C}$01 | .129 | $\hat{C}$07 | $\hat{P}$07 |
| 3 | $\hat{C}$16 | .325 | $\hat{C}$01 | .25 | $\hat{C}$08 | .20 | .. .. | .. .. | $\hat{C}$16 | .265 | $\hat{C}$16 | .27 | $\hat{C}$07 | $\hat{P}$11 |
| .. | .. | .. | .. | .. | .. | .. | .. .. | .. .. | .. | .. | .. | .. | .. | .. |
| .. | .. | .. | .. | .. | .. | .. | .. .. | .. .. | .. | .. | .. | .. | .. | .. |
| 24 | $\hat{C}$01 | .286 | $\hat{C}$05 | .75 | $\hat{C}$01 | .667 | .. .. | .. .. | $\hat{C}$15 | .417 | $\hat{C}$01 | .129 | $\hat{C}$01 | $\hat{P}$05 |
| 25 | $\hat{C}$06 | 0.50 | $\hat{C}$04 | .50 | $\hat{C}$02 | .50 | .. .. | .. .. | $\hat{C}$02 | .182 | $\hat{C}$01 | .129 | $\hat{C}$02 | $\hat{P}$15 |

$\hat{C}$K : clase de Klass  $\hat{P}$MP : Clase de predicción de probabilidad máxima

Tabla 5.1.7 Asignación de clase y probabilidad para cada variable e individuo del conjunto de prueba $P_0$

### 5.1.5 Criterios de agregación

Uno de los factores que inciden directamente en la asignación de clases es el criterio de agregación que se toma al hacer el análisis multivariable. Por lo tanto, como una de las contribuciones importantes de esta tesis, analizaremos la introducción de dos criterios que nos darán mejores resultados, dado que estos criterios, en principio, toman en cuenta la contribución de todas las variables.

Los nuevos criterios de agregación de información introducidos en la metodología de esta tesis, son:

- criterio de Votación (*Vot*), y
- criterio de Suma máxima (*Sum*).

Al igual que el criterio de agregación de probabilidad máxima, estos dos criterios tienen como entrada el conjunto de entrenamiento $T_0$, la partición de referencia, el conjunto de prueba $P_0$ y su partición correspondiente.

El criterio de agregación de información de votación (Vot) consiste en lo siguiente: para cada individuo $i$ del conjunto de prueba, leemos el valor $x_{ik}$ de la variable $X_k$, lo ubicamos en el intervalo correspondiente, digamos $I_s^k$, de la tabla de distribuciones e inicializamos un contador por variable para llevar el récord de cuántas variables con probabilidades distintas de cero se le asignan a Ĉ01, cuántas a Ĉ02 y así sucesivamente hasta determinar el número de variables que se le asignan a Ĉ20; acto seguido, nos fijamos en el número máximo de votos y al individuo $i$ le asignamos la clase correspondiente que tiene ese número.

El criterio de agregación de información de suma máxima (Sum) consiste en lo siguiente: para cada individuo *i* del conjunto de prueba, leemos el valor $x_{ik}$ de la variable $X_k$, lo ubicamos en el intervalo correspondiente, digamos $I_s^k$, de la tabla de distribuciones e inicializamos un sumador por variable para llevar la suma de las probabilidades de las variables que se les asigna la clase Ĉ02 y así sucesivamente hasta obtener la suma de probabilidades de las variables que se les asigna la clase Ĉ20; acto seguido, nos fijamos en la suma máxima y al

objeto $i$ le asignamos la clase correspondiente a esa suma máxima.

### 5.1.6 ¿Cómo evaluamos la estimación de la exactitud de predicción de un clasificador?

Para dar respuesta a esta pregunta nos planteamos la siguiente pregunta y tomamos en cuenta las siguientes consideraciones, analizando diferentes métodos para la estimación de exactitud de predicción.

1. ¿Qué tan predictivo es nuestro modelo que entrenamiento?.
2. El error en el entrenamiento de datos no es un buen indicador del rendimiento sobre futuros datos. Este error puede ser fácilmente reducida a 0, sin embargo necesitamos generalizaciones de nuestros datos.
3. Solución: dividir los datos en conjuntos de entrenamiento de de prueba.
4. Sin embargo, para crear un buen modelo necesitamos un gran conjunto de entrenamiento. Así que, lo que realmente necesitamos es una gran cantidad de datos preclasificados.
5. Necesitamos también confiabilidad estadística de las estimaciones de las diferencias en la exactitud de la predicción (pruebas de significación)
6. Medidas del desempeño:
    - Número clasificaciones correctas
    - Exactitud de las estimaciones de probabilidad
    - Costos asignados a tipos distintos de errores.
7. Para mejorar la precisión del clasificador podemos combinar múltiples modelos.
8. Podemos medir la exactitud de la predicción aplicando el Principio Descripción Longitud mínima (MDL).

**Entrenamiento y pruebas**

1. Tasa de error
    - Medida de rendimiento para problemas de clasificación.
    - *El éxito:* instancia de clase se predijo correctamente.
    - *Error:* instancia de clase se predice incorrectamente.
    - *(Observado) La tasa de error:* proporción de los errores cometidos sobre el conjunto total de instancias probadas.
    - *Resubstitución:* error sobre el conjunto de entrenamiento (medida demasiado optimista!).
    - *La verdadera* tasa de *error:* la tasa de error en la población (comúnmente estimada, porque en la mayoría de los casos la población no está disponible).

2. Pruebas
    - *Conjunto de prueba:* un conjunto de casos que no han sido utilizados en el proceso de entrenamiento.
    - Hipótesis: el conjunto de entrenamiento y de prueba son *representativos de las muestras* de una misma población
    - Algunos clasificadores trabajo en dos etapas (a menudo iterativamente):
        - Paso 1: aprendiendo la estructura básica.
        - *optimizando parámetros* que se utilizan en el aprendizaje.
    - El conjunto de prueba *no* deberá ser utilizado en *forma alguna en* el proceso de entrenamiento (incluso para los parámetros de sintonización, en el paso 2).
    - Puede haber otros conjuntos independientes de instancias para la optimización de parámetros (conjunto de validación). Es decir, dividimos el conjunto de datos conocidos en tres: conjunto de entrenamiento, conjunto de *validación* y el conjunto *de prueba.*
    - Procedimiento *Holdout*: el método de dividir los datos en conjunto de entrenamiento y de prueba..

      Dilema: el equilibrio entre la formación y de prueba.
3. Predicción del rendimiento (éxito verdadero/tasa de error).
    - Probando la estimación de la probabilidad de éxito sobre datos desconocidos (datos, no usados en entrenamiento ni pruebas).
    - ¿Qué tan buena es esta estimación? (¿Cuál es el éxito *verdadero /tasa de error?)* Necesitamos *intervalos de confianza* (una especie de razonamiento estadístico) para predecir esto.
    - Supongamos que el éxito y error son dos posibles resultados de un experimento estadístico (normalmente distribuidos en variable aleatoria).
    - P*roceso de Bernoulli:* Tenemos N experimentos y obtuvimos S éxitos Entonces, la tasa de éxito observada es *P = S / N.* ¿Cuál es la *verdadera* tasa de éxito?
    - Ejemplo:
        - *N = 100, S = 75.* Entonces, con un intervalo de confianza de un 80% *P* esta en *[0,691, 0,801].*
        - *N = 1000, S = 750.* Luego, con el intervalo de confianza de un 80% en *P* es *[0,732, 0,767].*
        -

**Estimación de la exactitud de predicción de un clasificador**

Consideremos los siguientes métodos:

1. Holdout
    - Reserva de una determinada cantidad para la realización de pruebas y usa el resto para el entrenamiento (en general, 1 / 3

para los conjuntos de pruebas y 2/3 para conjunto entrenamiento).

- Problema: las muestras podrían no ser representativas. Por ejemplo, algunas clases pueden ser representados con muy pocas instancias o, incluso sin la existencia de casos.
- Solución: *estratificación* –muestreo para el entrenamiento y pruebas dentro de las clases. Esto garantiza que cada clase se representa con proporciones aproximadamente iguales en ambos subconjuntos

2. Holdout Repetido. La estimación del éxito/error puede ser más confiable, repitiendo el proceso con diferentes muestras.
   - En cada iteración, una determinada proporción es seleccionada al azar para el entrenamiento (posiblemente con estratificación)
   - La tasa de error en las diferentes iteraciones son promedios y el resultado es la tasa de error.
   - Problema: Los diferentes conjuntos de pruebas pueden superponerse.¿Podemos evitar la superposición?
3. Cross-validación (CV). Evita superposición o solapamiento de los conjuntos de prueba.
   - *k-folds cross-validation*
     - Primer paso: los datos se divide en k subconjuntos de igual tamaño (por lo general en un muestreo aleatorio).
     - Segundo paso: cada subconjunto es usado para la realización de pruebas y el resto para el entrenamiento.
     - Las estimaciones del se promedian y el resultado es un promedio del error de estimación.
   - *Estratificación cross-validation:* son subconjuntos estratificados antes que la validación cruzada se realiza.
   - *Estratificado 10-folds cross-validation*
     - Método estándar para la evaluación. Extensos experimentos han demostrado que esta es *la mejor opción* para obtener una estimación exacta. También hay algunas pruebas teóricas para la anterior afirmación.
     - Estratificación reduce la estimación de la varianza.
     - Repetidas estratificaciones *10-folds cross-validation* es aún mejor. *10-folds cross-validation* se repite diez veces y los resultados se promedian.
4. Leave-One-Out Cross-Validation (LOO CV)
   - LOO CV es una *n-veces* validación cruzada, donde *n* es el número de instancias de capacitación. Esto es, *n* clasificadores se construyen para todos los subconjuntos de *(n-1)*-elementos posibles del conjunto entrenamiento y entonces se prueban sobre la instancia única restante
   - LOO CV hace máximo uso de los datos.

- No se considera muestreo al azar.
- Problemas
  - LOO CV es muy costoso computacionalmente.
  - Estratificación no es posible. En realidad este método no garantiza una muestra estratificada (sólo hay un caso en la prueba).
  - El peor de los casos ejemplo: asumir un conjunto de datos *completamente al azar,* con dos clases cada uno representado por el 50% de los casos. La mejor clasificación de estos datos es el mayor predictor. LOO CV predecir el 100% de error (!) Para este tipo de clasificador.

5. Bootstrapping
   - CV utiliza el muestreo sin reposición. Es decir, la misma instancia, una vez seleccionada, no puede ser seleccionado de nuevo para un entrenamiento/conjunto de prueba.
   - El bootstrap es un método de estimación que utiliza el muestreo con remplazamiento para formar el conjunto de entrenamiento.
   - El conjunto de entrenamiento: un conjunto de datos de *n* instancias es muestreado con sustituciones *n* veces para formar el conjunto de *n* ejemplos (posiblemente con repeticiones).
   - Conjunto de prueba: las instancias del conjunto de datos original que no ocurren en el conjunto de entrenamiento.
   - 0,632 bootstrap:
     - Una instancia particular tiene una probabilidad de (1-1/n) de no ser seleccionado para el conjunto entrenamiento. Así, una instancia caerá en el conjunto de prueba con probabilidad $(1-1/n)^n$ = (para n grande) = 1 / e = 0,368.
     - Esto significa que el conjunto de entrenamiento de datos contendrá aproximadamente el 63,2% de las instancias y, en consecuencia, obtendremos un error de estimación muy pesimista.
   - Bootstrapping es el mejor estimador de error para los pequeños conjuntos de datos.
6. Cálculo el costo
   - Diferentes tipos de errores de clasificación suelen incurrir en costos diferentes.
   - Ejemplo: predecir el cáncer. Compare el costo de la predicción "no" cuando la clasificación es "sí" y la predicción de "sí" cuando la clasificación es "no". Es evidente que el primer error es mucho más costoso.

- Matriz de confusión

| Actual \clase predicha | Sí | No |
|---|---|---|
| Sí | Verdadero positivo (TP) | Falsos negativos (FN) |
| No | Falsos positivos (FP) | Verdadero negativo (TN) |

- Error total = (FP + FN) / (VP + FP + TN + FN)

Así de la revisión hecha de los métodos de estimación de la precisión más comunes como: *Holdout, Cross-Validation, Leave-One-Out Cross-Validation (LOO CV) y BootStrapping*, los ejemplos mostraron donde cada uno deja de producir una buena estimación. Hemos comparado los enfoques sobre una gran variedad de bases de datos reales con diferentes características.

| Método | Desventaja | Ventaja |
|---|---|---|
| Test Set | Variancia: Estimaciones poco confiable | Barato |
| Leave-One_Out | Caro. Tiene algunos comportamientos extraños | Casi no desecha datos |
| 10-Folds | Desecha 10% de los datos. 10 veces más caros que los Test Set | Sólo desecha el 10% de los datos. Sólo 10 veces más caros en lugar de R veces |
| 3-Folds | Desperdicia más que Ten Folds. Más caro que Test Set | Ligeramente mejor que Test Set |
| R-Folds | Idéntico a Leave-One-Out | Idéntico a Leave-One-Out |

Tabla 5.1.8 Muestra las ventajas y desventajas del método k-fols cross-validation

En base a las pruebas se muestra que si los algoritmos de inducción son estables para un conjunto de datos dado, la varianza de las estimaciones de Cross-Validation deberá ser aproximadamente la misma, independiente del número de folders. Aunque los algoritmos de inducción no son estables, las estimaciones son aproximadamente estables. El K-folds Cross-Validation con valores moderados de K (10-20) se reduce la varianza al tiempo que aumenta el sesgo (predisposición). Cuando k disminuye (2-5) y el tamaño de las muestras más pequeñas, hay variancia en las estimaciones debido a la inestabilidad del entrenamiento, considerando el ten-folds croos-validation como el mejor método

para medir la estimación de exactitud de predicción de un clasificador. La Tabla 5.1.8 muestra las ventajas y desventajas para diferentes valores de k, siendo k=10 la mejor opción para la estimación de predicción.

### 5.1.7 Comparación de métodos inductivos usados en el descubrimiento de conocimiento de una planta de aguas residuales

En esta parte se presenta el estudio comparativo de diferentes métodos (estadísticos y de aprendizaje automático) para el reconocimiento de patrones de conocimiento del conjunto de datos provenientes de una planta depuradora de aguas residuales, que se discute en [44]. En el artículo se analiza el desempeño cuantitativo, en términos de la exactitud de la predicción sobre ejemplos no vistos, número de variables, ejemplos usados y el desempeño cualitativo en términos de la interpretación del significado para los expertos del dominio. Los métodos usados fueron: El método C4.5 de árboles de inducción, la técnica de inducción de reglas CN2 y su extensión J48, el antecedente de CIADEC: la técnica de inducción de reglas BPRI (Boxplot Rule Induction) y Opencase2, un método basado en memoria. Los resultados iniciales del estudio comparativo fueron publicados en la revista AI Communications en el año 2001, edición de marzo [44].

**Pruebas Experimentales**

Las pruebas experimentales y el procedimiento llevado a cabo para comparar los diferentes métodos se explican a continuación:

- Cinco diferentes técnicas de aprendizaje automático fueron usadas para obtener descubrimiento los patrones de los datos históricos que podrían ser de utilidad para alimentar a un sistema basado en conocimiento y así mejorar la supervisión de una planta depuradora de aguas residuales (ver [27] y [28]).
- Con el fin de probar el desempeño cuantitativo de las diferentes técnicas usadas en la predicción de casos no vistos, la estimación de la exactitud de predicción se puso a prueba: (i) sobre el conjunto de datos (usando los 243 individuos para ambos conjunto de entrenamiento y de prueba y, (ii) usando el método de estimación de exactitud de predicción ten-folds cross-validation.
- En el ten-Folds cross-validation, el conjunto total de 243 ejemplos se dividió en 10 conjunto de 24/25 ejemplares cada uno: estos a su vez fueron utilizados como de prueba, mientras que el resto de 219/218 ejemplos se utilizaron para el entrenamiento.
- Algunos algoritmos de bagging (empaquetamiento) y boosting (incremento) fueron aplicados al método J48, similar a C4.5, para mejorar su precisión de la predicción.
- El papel del especialista en los experimentos fue para seleccionar las variables más relevantes e interpretarlas, validar y determinar la utilidad de los árboles inducidos con C4.5 y Opencase, y las reglas del CN2,

BPRI derivadas de los 243 individuos de entrenamiento y posteriormente de CIADEC. El proceso de validación incluye comentarios sobre cuál de las ramas del árbol o reglas tienen sentido y cuáles no desde el punto de vista del experto: es decir, la forma en la cual fueron usadas las herramientas de inducción y marcar la importancia de los atributos. También, decidir sobre el significado y utilidad de los patrones de conocimiento descubiertos. La interpretación también implica un análisis de las razones por las cuales la exactitud de precisión y el significado cambia cuando los atributos usados también cambian. Finalmente, la interpretación también incluye observar el nuevo conocimiento en las nuevas reglas y árboles inducidos.

- Se realizaron 10 veces los experimentos para cada método determinando el promedio

### 5.1.8 Resultados

La comparación de los diferentes métodos se resume en la Tabla 5.1.9. Los parámetros utilizados en este experimento fueron: la exactitud de predicción, el número de atributos, ejemplos utilizados y el significado de la interpretación de los patrones inducido por el experto. Los patrones de los conocimientos descubiertos por C4.5, CN2, BPRI, Opencase y CIADEC puede ser codificado como reglas de decisión, o en una representación similar, y se puede ser añadidos a la base de conocimientos de un sistema basado en el conocimiento.

| Método | Mejor promedio obtenido en la predicción (hasta décimas de punto porcentual) |
|---|---|
| CN2 | 65.4% |
| C4.5 | 65.1% |
| J48 | 64.4% |
| Opencase2 | 62.5% |
| BPRI | 58.9% |
| **CIADEC** | **65.5%** |

Tabla 5.1.9 Comparación de métodos inductivos para predicción.

Con los avances logrados en la formalización, las mejoras al método introducidas por el análisis multivariable y los criterios de agregación, aunados a la automatización de la metodología BPRI la cual ha evolucionado hasta convertirse en CIADEC, los resultados más recientes son prometedores. La Tabla 5.1.9 muestra los mejores resultados obtenidos por cada uno de los métodos comparados, en una serie de experimentos realizados después de 2001, y validados con la técnica 10-fold *cross-validation*, con motivo de este trabajo de tesis.

Nótese la mejoría en rendimiento que exhibe CIADEC respecto de su antecesor, el BPRI.

# CAPÍTULO 6.

## CONCLUSIONES Y TRABAJO FUTURO

En este capítulo se presentan las conclusiones derivadas de los resultados obtenidos en el proceso de este trabajo de tesis; además, se proponen algunos de los posibles trabajos que se podrían realizar con el objeto de continuar con las ideas propuestas aquí, dando la pauta a futuros investigadores sobre los puntos no cubiertos, pero que pudieran ser afrontados en otros trabajos de investigación.

### 6.1 Conclusiones

**A nivel teórico-conceptual:**

1. La aportación principal es el diseño de una nueva Familia de Algoritmos para la Caracterización de Clases con aprendizaje inductivo para clasificación supervisada, en todo algoritmo de clasificación hay aprendizaje, los algoritmos de aprendizaje inductivo son especialmente relevantes porque han demostrado especialmente útiles.
2. Este modelo representa una nueva forma de extraer conocimiento relacionado con las diferentes situaciones que se presentan en el proceso de caracterización de clases.
3. En este trabajo hemos determinado la existencia de cuatro tipos de valores: propios ó no, parcialmente y totalmente caracterizadores, que juegan importantes papeles en el sistema de reglas.
4. El modelo propuesto es la base de una herramienta para la identificación de variables caracterizadoras.
5. Se obtuvo una metodología formal para la caracterización e interpretación de clases
6. Se ha creado un método de inducción de reglas basado en intervalos de longitud variable, los cuales resultan de la utilización novedosa de la herramienta estadística denominada boxplot

**A nivel metodológico:**

1. No es muy común y ha resultado interesante combinar algunos conceptos, técnicas y/o métodos de los dos enfoques tan distintos como son el Estadístico y el Lógico Combinatorio, ya que cada uno por separado tiene sus propias buenas aplicaciones en áreas distintas pero resultó absolutamente novedoso integrarlas en un Sistema Híbrido para la Caracterización de Clases y, además podemos asegurar por los resultados obtenidos que esta investigación abre una nueva Línea de Investigación que todavía da para más.
2. La interpretación de las clases para el caso de estudio se representó gráficamente, sobre las funciones de pertenencia para cada clase y cada variable cuantitativa; esta representación gráfica hace patente que el

paradigma difuso constituye un excelente soporte al proceso de interpretación.
3. Se utilizó el paradigma difuso como soporte en la interpretación de clases, a través de las etiquetas lingüísticas.
4. Se usaron algunos criterios de agregación para realizar la clasificación, a saber: votación, probabilidad máxima y suma de probabilidades.
5. A la vista de los resultados obtenidos en el proceso de predicción de la metodología, el criterio de agregación que mejor desempeño mostró fue el criterio de votación.
6. La metodología desarrollada muestra un buen desempeño en cuanto a tareas de predicción y diagnóstico.
7. Según los resultados presentados en el capítulo 5, sección 5.1, apartado 5.1.8, podemos decir que el modelo obtenido de inducción de reglas es mejor, en desempeño, con el método de clustering *C4.5* y su similar, el método de inducción de reglas *CN2*, en cuanto a tareas de predicción; su desempeño es mucho mejor en la interpretación de resultados.
8. El costo computacional de la metodología implementada es bajo en relación a la información que proporciona, puesto que se resuelve un análisis de intersecciones de grado $\xi$ (en el caso de estudio 20), calculando solamente $\xi$ máximos y $\xi$ mínimos, y ordenándolos.

Por lo anterior, se puede afirmar que la metodología original de esta tesis es adecuada, y ha empezado a dar buenos resultados; además, continúa en proceso de mejoramiento, con miras a aumentar su capacidad de caracterización e interpretación de las clases resultantes de una partición de referencia, así como en tareas de predicción,

## 6.2 Trabajo futuro

Como trabajo futuro, se plantea un desarrollo más profundo sobre algunos temas que hay necesidad de estudiar con más profundidad para consolidar este modelo, como son:

1. Desde el punto de vista de la aplicación, revisar la clasificación de referencia utilizando un algoritmo de clasificación basado en reglas, que de acuerdo con el experto de cada dominio mejore la calidad de la clasificación de referencia. Los resultados de este trabajo podrían mejorarse con el desarrollo de un módulo de clasificación que incluya diversos métodos inductivos al sistema CIADEC [15].
2. Estudiar la conveniencia, o no, de optimizar los $I^k$ para que no contengan modalidades vacías, que resultan al combinar extremos de clases distintas de igual valor [50].
3. El método aplicado como hasta ahora para todas las variables produce poca cobertura (los $\varepsilon$ son pequeños). Se propone introducir las modificaciones necesarias al modelo, para cubrir la mayor parte de casos con reglas [50].

4. Establecer, si es posible, una relación entre la $p_{sc}$ (frecuencia relativa de individuos en $I_s^c$ y que pertenecen a la clase $C \in P$ ) y $\varepsilon$ (grado de caracterización a la clase $C$ ) [50, 51].
5. Con el uso de algoritmos y heurísticas eficientes, concatenar la metodología con un método de razonamiento difuso, que permita generar, en forma automática, la asignación de clases y las descripciones conceptuales de éstas, elementos esenciales en la interpretación de resultados en los así llamados dominios poco estructurados [50, 51]
6. Suavizar las funciones de pertenencia de los gráficos por variable, con el fin de obtener funciones de pertenencia difusas que mejoren la interpretación de clases y, en consecuencia, la visualización de resultados [50, 51]
7. Seleccionar un nuevo criterio de agregación de información que nos permita considerar todas las reglas que se "disparan", tanto a nivel de variable como de combinación de éstas, para mejorar la eficiencia del sistema de reglas obtenido [51, 152]
8. Aplicar el modelo en otros dominios, como son: Atmosférico, Médico, Crediticio, Educativo, Industrial, entre otros.

### 6.3 Publicaciones

1. Vázquez, F., Gómez, P. *Automatic Construction of Fuzzy Rules for Modelling and Prediction of the Central Nervous System.* IbPRIA 2007, 3rd Iberian Conference on Pattern Recognition and Image Analysis. Universitat de Girona, Spain, June 6-8, 2007. In preparation process in *Part I, LNCS 4477-0443*. 2007.
2. Rodas, J., Alvarado, G. *and* Vázquez, F. *KDSM: Effectiveness Detection on Government Programs*. Universidad de Puerto Rico, Puerto Rico, USA.. Latin American and Caribbean *Journal* of Engineering Educations. LACJEE. 2007.
3. Rodas, J., Alvarado, G. *and* Vázquez, F. *KDSM: Detección de efectividad en programas gubernamentales*. Latin American and Caribbean Consortium of Engineering Institutions. Universidad de Puerto Rico at Mayagüez, Puerto Rico, USA. June 18th. LACCEI .2006
4. Rodas, J. & Vázquez F. Alvarado, G. *Using the KDSM methodology for knowledge discovery from a labour domain*. Sixth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. Towson University, Towson, Maryland, USA. May 2005
5. Rodas, J. & Vázquez F. Alvarado, G. *Using the KDSM methodology for knowledge discovery from a labour domain*. Journal: SNPD-SAWN, volume = 00, isbn = 0-7695-2294-7, pages: 64--69, Towson University. Maryland, USA, http://doi.ieeecomputersociety.org/10.1109/SNPD-SAWN.2005.79, Los Alamitos, CA, USA, IEEE Computer Society, may 2005.

6. Vázquez F. & Díaz de León J.L. *Characterization and Interpretation of Classes Based on Fuzzy Rules in ill-Structured Domains*. Fourth Mexican International Conference on Artificial Intelligence. MICAI-2005, Monterrey, N.L. México. Nov, 2005.
7. Vázquez F. & Gómez P. *Caracterización e interpretación automática de descripciones conceptuales en dominios poco estructurados*. CN y CIIC 2003, ISBN 970-36-0102-2, Zacatecas, México. Octubre 2003.
8. Rodas, J., Alvarado, G. & Vázquez, F. *Applying KDSM to an specific domain where very short and repeated serial measures with a blocking factor are presented*. Research LSI-02-53-R, Technical University of Catalonia, Barcelona. Spain, January 2002. http://www.lsi.upc.es/dept/techreps/html/R02-28.html.
9. Vázquez F. & Gibert K. *Implementation of the methodology "Automatic Characterization and Interpretation of Conceptual Descriptions in ill-Structured Domains*. Research LSI-02-28-R, Technical University of Catalonia, Barcelona. España, Enero 2002. http://www.lsi.upc.es/dept/techreps/html/R02-28.html.
10. Vázquez, F., Gibert, K. Robustness of class prediction depending on references partition in-Ill-Structured Domains. 8th. Iberoamerican Conference on Artificial Intelligence. Sevilla, España. 2002.
11. Vázquez F. & Gibert K. *Fundamentos de la Teoría de los Conjuntos Borrosos y la Lógica Borrosa*. Research LSI-02-3-T, Technical University of Catalonia, Barcelona, Spain, March 2002. http://www.lsi.upc.es/dept/techreps/html/R02-28.html.
12. Vázquez F. & Gibert K. *Generación Automática de Reglas Difusas en Dominios Poco Estructurados con Variables Numéricas.* En Actas de la Conferencia de la Asociación Española para la Inteligencia Artificial, volumen 1, Págs. 143-152, España, CAEPIA 01. Nov 2001.
13. Vázquez F. & Gibert K.. *Automatic generation of fuzzy rules in ill structures domains with numerical variables*. Research LSI-01-51-R, Technical University of Catalonia, Barcelona, Spain, http://www.lsi.upc.es/dept/techreps/html/R01-51.html. December 2001.

# REFERENCIAS


[1] Pozo, J. I. Adquisición del Conocimiento. 271 pp. ISBN: 84-7112-489-0. Madrid: Ediciones Morata. 2005.

[2] Carrión, J. (n/d). Diferencia entre dato, información y conocimiento. http://www.gestiondelconocimiento.com. 2004 Carrión, J. (n/d). Diferencia entre dato, información y conocimiento. http://www.gestiondelconocimiento.com. 2004

[3] Poole, D. Mackworth, A. & Goebel, R. Computational Intelligence: A Logical Approach. Oxford University Press, 1998.

[4] Shortlife E.H. MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection. PhD thesis, Stanford University, USA, 1976.

[5] Clancey W.J., & Shortliffe E.H.. "Readings in Medical Artificial Intelligence". Addison-Wesley, 1984.

[6] Szolovits P., & Pauker S.G. "Categorical and probabilistic reasoning in medical diagnosis". Artificial Intelligence, Vol. 11, pp. 115-144, 1978.

[7] Michalski R. & Steep R.E. “A Theory and Methodology of Inductive Learning”. In J. Carbonell, editor, Machine learning: A Artificial Intelligence Approach”, Chapter 11, pages 331-363. Ed. Tioga, Palo Alto, California, 1984.

[8] Quinlan, J.R. Discovering Rules by Induction from Large Collection of Examples. In Michele, D (Ed.) Expert System in The Micro-electronics Age. Edinburgh University Press, 1979.

[9] Núñez, G., et al. About the attribute relevance's nature. En Proceedings of TEC. COM. 91, Approaches to non-conventional computing: towards intelligent systems. México, 1991.

[10] Gibert K. L’us de la Informaciò Simbòlica en l’Automatitizaciò del Tractament Estadístic de Deominis Poc Estructurats. In the Statistics and operations research Phd. Thesis., Universitat Politécnica de Catalunya, Barcelona, Spain, 1994.

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

[12] Gibert K. Técnicas híbridas de Inteligencia Artificial y Estadística para el descubrimiento de conocimiento y la minería de datos. Departamento de Estadística e Investigación Operativa, Universitat Politècnica de Catalunya. Ediciones UPC. 2004.

[13] Bisquerra. Introducción conceptual al análisis multivariable. Un enfoque informático con los paquetes SPSS-X, BMDP y SPAD. Volumen III. McGraw Hill, España. 1989.

[14] Aluja T. Análisis Factoriales Descriptivos con SPAD-N. UPC. España. 1996.

[15] Sànchez-Marrè M., Cortés U., Lafuente J., & Poch M. Concept formation in WWTP by means of classification techniques: A compared study. Applied Intelligence. 7:147-166., 1997.

[16] Roda-I. Cortés U. Gibert, K. & Sàchez-Marrè. Identifying characteristic situations in wastewater treatment plants. Workshop in Binding Environmental Sciences and Artificial Intelligence, 1:1-9, EDAI, 2000.

[17] Rodríguez D. Análisis de los datos de una planta depuradora de aguas utilizando la clasificación basada en reglas, 1999.

[18] Sànchez-Marrè M. An Integrated Supervisory Multi-level Architecture for WasteWater Treatment Plants. PhD thesis, UPC, 1995.

[19] Gibert K. & A. Salvador. Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales. In X Congreso Español sobre tecnologías y lógica fuzzy, pages 497-502, España, sep 2000. ESTYLF 2000.

[20] Bayona, S. Descriptiva de dades y de classes. PFC Facultad d' Informática, UPC, jul 2000.

[21] Aluja, T. & Morineau, A. Aprender de los Datos: El Análisis de Componentes Principales. Una aproximación desde el Data Mining. Ed: EUB S.L. 1999.

[22] Gibert K. & Cortés U. Combining Knowledge bases system with a clustering method for an inductive construction of models. In Proc. 4th In Work. On AI and Stats. Florida, USA,1993.

[23] Gibert K. & Cortés U. On the uses of the expert Knowledge for automatic biasing of a clustering method. In ITI 93. Proceedings of the International Conference on Information Technology Interfaces, pages 219-224, ISSN 1330-1012, Croatia, 1993.

[24] Gibert K. Klass. Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Master's thesis, UPC 1991.

[25] Gibert K. & Cortés U. Combining a knowledge-based system and a clustering method for a construction of models in ell-structured domains. In Artificial Intelligence and Statistics IV, volume 89 of Lecture Notes in Statistics, pages 351-360, Springer-Verlang, New York, N.Y. US., 1994.

[26] Gibert K. & Cortés U. KLASS: Una herramienta estadística para la creación de prototipos en dominios poco estructurados. Proa. IBERAMIA-92., pages 483-497, Noriega Eds. México, 1992.

[27] Gibert K., Hernández, & Cortés U. Classification based on rules: an application to Astronomy. In Ed. Tokio. Japón, editor, Proceedings of 5. Conference of International Federation of Classification Societies, pag. 69-72, Mar 1996.

[28] Gibert K. & Cortés U. Clustering based on rules and knowledge discovery in ill-structured domains. Computación y Sistemas., 1(4): 213-227, ISSN 1405-5546. Impreso en México, 1998.

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

[30] Rodas J., Gibert K., & Rojo J. Electroshock Effects Identification Using Classification Techniques. Springer's Lecture Notes of

Computer Science Series; Crespo, Maojo and Martin (Eds.):238-244, Second International Symposium, ISMDA 2001.

[31] Gibert K., Alhuja T., & Cortés U. 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, Springer-Verlang. Interpreting Results. Nantes, 1998.

[32] Fayyad U. From Data Mining to Knowledge Discovery: An overview. ISBN 0-262-56097-6. USA, 1996.

[33] Fayyad U., Piatetsky-Shapiro G., Smyth P., & Uthursamy R. Advances in Knowledge Discovery and Data Mining. AAAI Press. 1996.

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

[35] Gibert K. & Alhuja T. A computational technique for comparing classifications and its relationship with knowledge discovery. In International Seminar on New Techniques and Technologies for Statistics, pages 193-198. Italy, Nov. 1998.

[36] Diday E. & Gowda K.C. Symbolic clustering using a new similarity measure. In IEEE Trans. On systems, man., and cib., volume 22, pages 368-378, 1992.

[37] Gibert K. & Cortés U. Weighing quantitative and qualitative variables in clustering methods. Math ware and Soft Computing, 4(3):251 – 266, 1997.

[38] Gibert K. On the uses and costs of rules-based classification. In A. Prat. Physical-Verlang, editor, Proceedings of Computational Statistics, pages 265-270, march 1996.

[39] Castillejo X.. Un entorn de treball per a Klass. PFC Facultad d' Informática UPC, julio, 1996.

[40] Márquez J. & Martín J.C. La clasificación automática en las ciencias de la salud. PFC, Facultat de Matemàtiques i Estadística, UPC, Octubre, 1997.

[41] Gibert K. & Sonicki Z. Classification Based on Rules and Medical Research. In Rocco Curto, editor, VIII International Symposium on Applied Stochastic Models and Data Analysis, pages 181-186, ASMDA97, Italy, 1997.

[42] Tubau X. Sobre el comportement de les mètriques mextes en algorismes de Clustering. PFC, Facultat d´Informàtica, UPC Octubre 1999.

[43] 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.

[44] Comas J., S. Dzeroski S., Gilbert K., Rodas I., & Sánchez-Marré M. Knowledge discovery by means of inductive methods in wastewater treatment plant data. AI communications. The European journal on artificial intelligence, 14 (1):45-62, march 2001.

[45] Gómez B. Herramientas de muestreo y de clasificación basada en bootstrap. PFC, Facultat de Matemàtiques i Estadística, UPC. Octubre 2000.

[46] Nieto M. A. Compilación de técnicas de minería de datos y de descubrimiento de conocimiento. PFC, Facultat de Matemàtiques i Estadística, UPC. Octubre 2000.

[47] Gower J.C.. A. General coefficient if similarity and some of its properties. Biometrics, 27:857-874, 1971.

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

[49] Rodas J. Metodología para el descubrimiento de conocimiento en medidas seriadas muy cortas y repetidas con factor de bloque. Phd. Thesis., Universitat Politécnica de Catalunya, Barcelona, España, 2003.

[50] Vázquez F. & Gibert K.. Automatic generation of fuzzy rules in ill structures domains with numerical variables. Research LSI-01-51-R, Technical University of Catalonia, Barcelona, Spain, http://www.lsi.upc.es/dept/techreps/html/r01-51.html. December 2001.

[51] Vázquez F. & Gibert K. Generación Automática de Reglas Difusas en Dominios Poco Estructurados con Variables Numéricas. En Actas de la Conferencia de la Asociación Española para la Inteligencia Artificial, volumen 1, Págs. 143-152, España, CAEPIA 01. Nov 2001.

[52] Vázquez F. & Gibert K. Implementation of the methodology "Automatic Characterization and Interpretation of Conceptual Descriptions in ill-Structured Domains. Research LSI-02-28-R, Technical University of Catalonia, Barcelona. España, Enero 2002. http://www.lsi.upc.es/dept/techreps/html/R02-28.html.

[53] Rodas, J., Alvarado, G. & Vázquez, F., Applying KDSM to an specific domain where very short and repeated serial measures with a blocking factor are presented. Research LSI-02-53-R, Technical University of Catalonia, Barcelona. Spain, January 2002. http://www.lsi.upc.es/dept/techreps/html/R02-28.html.

[54] Vázquez F. & Gómez P. Caracterización e interpretación automática de descripciones conceptuales en dominios poco estructurados. CN y CIIC 2003, ISBN 970-36-0102-2, Zacatecas, México. Octubre 2003.

[55] Rodas, J. & Vázquez F. Using the KDSM methodology for knowledge discovery from a labour domain. Sixth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. Towson University, Towson, Maryland, USA. May 2005

[56] Vázquez F. & Díaz de León J.L. Characterization and Interpretation of Classes Based on Fuzzy Rules in ill-Structured Domains. Mexican International Conference on Artificial Intelligence. MICAI-2005, Monterrey, N.L. México. Nov, 2005.

[57] Michalski R.S. Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data. IJPAIS, 4:219-243., 1980.
[58] Stephen M. Stigler. Statistics on the Table: The History of Statistical Concepts and Methods, Harvard University Press; Reprint edition, august, 2002.
[59] Pearson K., The Life, Letters, and Labours of Francis Galton. London, 1914-30. 1967.
[60] Fisher, R. A. Biologist, statistician. Published much seminal work in the field of population genetics. Author of "Design of Experiments" (1935), "Genetical Theory of Natural Selection" (1930), and "Statistical Methods and Scientific Inference, 1956.
[61] Ashok Rudra. Prasanta Chandra Mahalanobis. A Biography. Oxford University Press. Dec 1997.
[62] P.H. Sneath & R.R. Sokal. Numerical Taxonomy - The principles and practice of numerical classification. W. H. Freeman, San Francisco, USA, 1973.
[63] Belzer J., Holzman A. G., & Kent A. Encyclopedia of Computer Science and Technology. Marcel Dekker, Inc. USA, 1980.
[64] Partridge D. & Alexander Y. Wilks. The Foundations of Artificial Intelligence: A Sourcebook. Cambridge University Press, G.B. 1990.
[65] Buchanan D. R. & Shortliffe,E.H., "Production Systems as a Representation for a Knowledge-based Consultation program", Artificial Intelligence, 8, (1), pp. 15-45, 1977.
[66] C.Lau, "Neural Networks, Theoretical Foundations and Analysis", IEEE Press, 1991.
[67] Coello C. La Computación Evolutiva en el Contexto de la Inteligencia Artificial. LANIA, A.C., México, 2000.
[68] Winston, Patrick H., Inteligencia Artificial. Addison-Wesley Iberoamericana, 3ª ed., 1994.
[69] Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, G. Weiss (ed.) The MIT Press, Cambridge, 1999.
[70] Knapik, M. & Johnson J., Developing Intelligent Agents for Distributed Systems: Exploring Architecture, Technologies & Applications, McGraw-Hill, 1998.
[71] SPSS Inc's. Clementine 10.0, with access data collected using Dimensions™ family of survey research products. 2005
[72] M. R. Anderberg. Cluster Analysis for aplications. Academic Press, 1973.
[73] M. Volle. Analyse des données, 1985. Ed. Económica, Paris, France.
[74] L. Lebart. Traitement statistique des données. Ed. Económica, Paris, 1990.
[75] Ralambondrainy H. A conceptual version of the K-means algorithm. Lifetime Learning Publications, Belmont, California, 1995.

[76] Ruiz-Shulcloper J. et al. Data analysis between sets of objects. In 8th ICSRIC, volume III, pages 85-81, Baden Baden, august 1996.
[77] Tukey J.W. Exploratory Data Analysis. Addison-Wesley, 1977.
[78] Ruiz-Shulcloper J. et. al Introducción al Reconocimiento de Patrones. Serie Verde No. 51. Editorial CINVESTAD-IPN
[79] Cheremesina E.N., J. Ruiz-Shulcloper (1992). Cuestiones metodológicas de la aplicación de modelos matemáticos de Reconocimiento de Patrones en zonas del conocimiento poco formalizadas. Revista Ciencias Matemáticas, vol. 13, No.2, pp. 93-108, Cuba.
[80] Springer-Verlang. Artificial Intelligence and Statistical IV, volumen 89, USA, 1994.
[81] Brachman R. J. & T. Anand. The process of knowledge discovery in databases: A human centered approach. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, chapter 2, pages 37-57. AAAI/MIT Press, 1996.
[82] S. Watanabe. Pattern Recognition: Human and Mechanical. Wiley, 1985.
[83] Fu K. S. A step toward unification of syntactic and statistical pattern recognition. IEEE Trnas. Pattern Analysis and Machine Intelligence. 5(2):200-205, 1983.
[84] Bajcsy R. & Kovacic S. Multiresolution elastic matching. Computation Vision Graphics Image Process., 46:1 -21, 1989.
[85] Grenander U.. General Pattern Theory. Oxford University Press. First Edition. 1993.
[86] Devroye L., Gyorfi L. & Lugosi G. A Probabilistic Theory of Pattern Recognition. Springer-Verlang, Berlin, first edition, 1996.
[87] Duda R.O. & Hart P.E.. Pattern Classification and Scene Analysis. Wiley and Sons., New York, 1973.
[88] Fu K.S. Syntactic Pattern Recognition and Applications. Prentice-Hall, Englewood Cliffs. 1982.
[89] Pavlidis T. Structural Pattern Recognition. Springer-Verlag., New York. 1977.
[90] Perlovsy L.I. Conundrum of combinatorial complexity. IEEE Trans. Pattern Analysis and Machine Intelligence. 20:666-670,1998.
[91] Jain A.K., Dubes R.C. & Chen C.C. Bootstrap Techniques for error estimation. IEEE Trans. Pattern Analysis and Machine Intelligence. 9:628-633, 1987.
[92 Kohonen T. Self-Organizing Maps. Springer Series in Information Sciences, 30, USA, 1995.
[93] Yañez Márquez C. & Díaz de León J.L. "Lernmatrix de Steinbuch", IT 48 Serie Verde, CIC-IPN, México, 2001.
[94] Yañez Márquez C. & Díaz de León J.L. "Linear Associator de Anderson-Kohonen", IT 50 Serie Verde, CIC-IPN, México, 2001.
[95] Castellanos Sánchez C. & Díaz de León J.L. y Sánchez López A. "El Paradigma de las Redes Neuronales Morfológicas", México, 1999.

[96] Yañez Márquez C. “Memorias Asociativas Basadas en Relaciones de Orden y Operadores Binarios”. Tesis doctoral. CIC-IPN, México, 2003.
[97] Santiago Montero R. “Clasificador híbrido de patrones basados en la Lernmatrix de Steinbuch y Linear Associator de Anderson-Kohonen”. Tesis de Maestría. CIC-IPN. 2003.
[98] Fukunaga K. Introduction to Statistical Pattern Recognition. Academic Press. USA, 1990.
[99] Devijver P.A. & Kittler J. Pattern Recognition: A Statistical Approach. Prentice Hall. London first edition, 1982.
[100] Bishop C.M. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, firs edition, 1995.
[101] Ripley B. Pattern Recognition and Neural Networks. Cambridge University Press., Cambridge, first edition, 1996.
[102] Schuhfried G. Wiener Test system. Vienna Reaction Unit, Basic Program. Development and production of scientific equipment. Molding, Austria, 1992.
[103] McLachlan G. Discriminate Analysis and Statistical Pattern Recognition. Wiley and Sons, New York, first edition, 1992.
[104] Nagy G. State of the art in pattern recognition. Proc. IEEE., 56:836-862, USA, 1968.
[105] Kantrowitz M. Milestones in the Development of Artificial Intelligence 1994. Web, 1994.
[106] Ruiz-Shulcloper J. y Lazo M. (1990). Modelos matemáticos para el Reconocimiento de Patrones. Editorial UCLV, Santa Clara, Cuba.
[107] Martínez-Trinidad, J. Fco., Ruiz-Shulcloper J. y Lazo M. “Structuralization of universes”. Fuzzy Sets & Systems 112/3, 2000b, pp 485-500.
[108] Martínez-Trinidad, J. Fco., Guzman-Arenas, A. The logical combinatorial approach to pattern recognition an overview through selected works, Pattern Recognition, 2001, 34/4 1-11.
[109] Ruiz-Shulcloper J., Guzman-Arenas, A., Martínez-Trinidad, J. Fco. Enfoque Lógico Combinatorio al Reconocimiento de Patrones, Cinvestav-IPN, 1999.
[110] Everitt B. Cluster Analysys. Heinemann, London, 1981.
[111] Dmitriev A.N., Zhuravliov, Yu I., Krendelev F.P. Acerca de los principios matemáticos de la clasificación de objetos y fenómenos. Sbornik Diskrtnii Analisis, Tomo 7, pp 3-15, Novosibirsk. 1966.
[112] Mamdani E.H. & Gaines G.R. Fuzzy reasoning and its Applications. Mamdani-Gains eds., USA, 1981.
[113] Hughes G.E. & Creswell M.J. An Introduction to Modal Logic. London, England, eds., 1968.
[114] McDermott J. R1: A rule-based configured of computer systems. USA, 1982.
[115] Brachman R. & Anand T. The Process of Knowledge Discovery in Databases: A Human-Centred Approach. In Advances in Knowledge discovery and Data Mining, pages 65-78, Ed. U.

Fayyad, G Piatetsky-Shapiro, P. Smyt, and R. Uthurusamy, AAAI/MIT Press, 1996.

[116] Zadeh L.A. Fuzzy Sets. Information and Control, pages 338-353, USA, 1965.

[117] Vázquez F. & Gibert K. Fundamentos de la Teoría de los Conjuntos Borrosos y la Lógica Borrosa. Research LSI-02-3-T, Technical University of Catalonia, Barcelona, Spain, March 2002. http://www.lsi.upc.es/dept/techreps/html/R02-28.html.

[118] Mizumoto M. & Zimmermann H.J. Comparison of fuzzy reasoning methods, Great Britain, first edition, 1982.

[119] Alsina C. & Trillas E., and Valverde L. On some logical connectives for fuzzy set theory. Math. Anal. Appl., 93:149-163, 1997.

[120] Aguilar J. & Gibert K. Sobre variables lingüísticas, difusas, paradigmas parmenidianos y lógicas multivaluadas. ESTYLF, 1:185-192, 1991.

[121] Aguilar J., Gilbert K. & Rodriguez. Fuzzy semantic in expert process control. LNAI, 1993.

[122] Dubois D., Prade H., & Bezdek J. Fuzzy sets in approximate reasoning and information system, volume 1. Kluwer Academic Publishers, 1999.

[123] Pedrycz W. & Gomide F. An Introduction to Fuzzy Sets. The MIT, Press. 1998.

[124] Klir, G.J. & Folger, T.A. Fuzzy Sets, Uncertainty and Information. Englewood Cliffs, NJ: Prentice Hall. 1988.

[125] Zadeh, L.A. Possibility theory and Soft data analysis. In Mathematical Frontiers of Social and Policy Sciences, ed. L Cobb and R. Thrall, 69-129. Boulder, Co: Westview Press. 1981.

[126] Dubois D. & Prade H. A Review of fuzzy set aggregation connectives. Information Sciences, 36:85-121, 1985.

[127] Zadeh L.A. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. System. Man Cabernet, pages 28-44, 1973.

[128] Zimmermann, H.J. Fuzzy Set Theory and its Applications, Boston: Kluwer Academic Publishers, cop.1996.

[129] Zadeh L.A. The role of fuzzy logic and soft computing in the conception and design of intelligent systems. 8th Austrian Artificial Intelligence Conference, LNAI 695. 695:1-5, 1993.

[130] Cordón, O., M.J. Del Jesús, y F. Herrera. A proposal on reasoning methods in fuzzy rule-based classification system. 1998.

[131] Cordon F. Herrera, and A. Applicability of the fuzzy operators in the design of fuzzy logic. Controllers. 1997.

[132] Abe S. & Thawonmas R. A fuzzy classifier with ellipsoidal regions. IEEE Trans. on Fuzzy Systems,, pages 358-368, 1997.

[133] Cordón O., De Jesús M.J., & Herrera F. Completeness and consistency conditions for learning fuzzy rules, USA, 1999.

[134] Ishibuchi H., Nozaki K., & H. Tanaka. "Distributed representation of fuzzy rules and its applications to pattern classification". Fuzzy Sets Syst. Vol 52, pp 21-32, 1992.

[135] Mandal D.P., Murthy C.A. & S.K. Pal, "Formulation of a Multivalued Recognition System", IEEE Trans. Syst., Man and Cyberns., vol. 22, pp. 607-620, 1992.

[136] Cordón O., del Jesús M.J., Herrera F. Métodos de Razonamiento Aproximado Basados en el Concepto de Mayoría Difusa para Sistemas de Clasificación. VIII Congreso Español sobre Tecnologías y Lógica Fuzzy. Pamplona (Spain), 1998, pp. 399-404.

[137] Cordón O., del Jesus M.J., Herrera F. A Proposal on Reasoning Methods in Fuzzy Rule-Based Classification Systems. International Journal of Approximate Reasoning. Vol. 20 (1999), pp. 21-45. (22 pages).

[138] Chi Z., Yan H. & Pham T. Fuzzy algorithms whit applications to image processing and pattern recognition. World Scientific, pages 101, 105, 1996.

[139] Yager R.R. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans. On Systems, Man and Cybernetics. 18:183-190, 1988.

[140] Cordón O., del Jesús M.J. & Herrera F. Analyzing the Reasoning Mechanisms in Fuzzy Rule-Based Classification Systems. Mathware & Soft Computing. Vol. 5: 2-3 (1998), pp. 321-332.

[141] Yager R.R. Families of OWA operators. Fuzzy Sets and Systems. 59:125-148, 1993.

[142] López de Mántaras. Approximate reasoning models. Ellis Horwood series in AI, 1990.

[143] Font, J.M. & Hájek, P. On Lukasiewicz's four-valued modal logic. Studia Logica. 70. 157–182, 2002.

[144] Zadeh, L.A. The Concept of a Linguistic and its Application to Approximate Reasoning, Memorandum Erl-M 411, Berkeley, October 1973.

[145] Zadeh, L.A. From Computing with Numbers to Computing with Words from Manipulation of Measurements to Manipulation of Perceptions, IEEE Trans. On Circuits and System 1: Fundamental Theory and Applications, 1999.

[146] Zadeh, L.A. Toward a Perception Based Theory of Probabilistic Reasoning with Imprecise Probabilities. Journal of Statistical Planning and Inference (105), 2002.

[147] Zadeh, L.A. Toward a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic. Fuzzy Sets and Systems, Vol. 90, 1997

[148] Roda, R., Poch, M., y Sánchez-Marrè, M. Tratamiento de Aguas Residuales. Barcelona. 1995.

[149] Peña, D. Estadística, Modelos y Métodos. Modelos lineales y series temporales, volumen II. Alianza, Madrid, segunda edición, 1989.

[150] Walpole, R., Myers, R. y Myers S. Probability and Statistics for Engineers and Scientists, volume I. Prentice Hall, sixth edition, 1998.

[151] Vázquez, F., Gibert, K. Robustness of class prediction depending on references partition in-Ill-Structured Domains. 8th. Iberoamerican Conference on Artificial Intelligence. Sevilla, España. 2002.

[152] Vázquez F. & Gibert K. Caracterización e Interpretación Automática de Descripciones Conceptuales en Dominios poco Estructurados usando variables numéricas. Research LSI-02-51-R, Technical University of Catalonia, Barcelona, Spain, http://www.lsi.upc.es/dept/techreps/html/02-51-R.html. Mayo 2002.

[153] Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[154] Vázquez, F., Gómez, P. Automatic Construction of Fuzzy Rules for Modelling and Prediction of the Central Nervous System. IbPRIA 2007, 3rd Iberian Conference on Pattern Recognition and Image Analysis. Universitat de Girona, Spain, June 6-8, 2007. Part I, LNCS 4477-0443. 2007.

[155] Rodas, J., Alvarado, G. and Vázquez, F. KDSM: Effectiveness Detection on Government Programs. Universidad de Puerto Rico, Puerto Rico, USA.. Latin American and Caribbean Journal of Engineering Educations. LACJEE. 2007.

[156] Rodas, J., Alvarado, G. and Vázquez, F. KDSM: Detección de efectividad en programas gubernamentales. Latin American and Caribbean Consortium of Engineering Institutions. Universidad de Puerto Rico at Mayagüez, Puerto Rico, USA. June 18th. LACCEI .2006.

[157] Rodas, J. & Vázquez F. Alvarado, G. Using the KDSM methodology for knowledge discovery from a labour domain. Sixth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. Towson University, Towson, Maryland, USA. May 2005.

[158] Rodas, J. & Vázquez F. Alvarado, G. Using the KDSM methodology for knowledge discovery from a labour domain. JOURNAL: SNPD-SAWN, volume =1 00, isbn = 10-7695-2294-7, pages: 64--69, Towson University. Maryland, USA, http://doi.ieeecomputersociety.org/10.1109/SNPD-SAWN.2005.79, Los Alamitos, CA, USA, IEEE Computer Society, may 2005.

[159] Béjar Alonso, J. Adquisición Automática de Conocimiento en Dominios poco Estructurados. Tesis doctoral. Departament de Llenguatges i Sistemes Informàtics . UPC, 1995.

[160] Bongard M.N. et al., Solución de problemas geológicos con la ayuda de programas de reconocimiento. Sov Geología, No. 6, Moscú. 1963.

[161] Voronin Yu. A. Introducción a la Teoría de la Clasificación. Novosibirsk. 1982.

[162] Dmitriev A.N., Zhuravliov, Yu. I. & Krendelev F. P. Acerca de los principios matemáticos de la clasificación de objetos y fenómenos. Sbornik Diskrtnii Análisis, Tomo 7, pp. 3-15, Novosibirsk. 1966.
[163] Ruiz-Shulcloper J. et al. Tópicos acerca de la Teoría de Testores. Colección Amarilla, No. 134, pp. 1-51. CINVESTAV-IPN, México. 1994.
[164] Godoy Calderón S. Generalización de los Conceptos de Testor y Testor Típico para Entornos Difusos. Tesis en opción al grado de Maestro en Ciencias. Depto. Ing. Eléctrica. Cinvestav-IPN. México.1994.
[165] Godoy Calderón S., Lazo Cortés M. Delta-Testores. Una Generalización del concepto de Testor para entornos difusos. II Taller Iberoamericano de Reconocimiento de Patrones, La Habana, Cuba, 1997.
[166] Godoy Calderón S., Martínez-Trinidad J. Fco, Lazo Cortés M. Proposal for a Unified Methodology for Evaluating Supervised and Non-Supervised Classification Alghoritms. Progress in Pattern Recognition, Image Analysis and Applications, 11th Iberamerican Congress in Pattern Recognition, CIARP, 2006.
[167] Godoy Calderón S. Evaluación de Algoritmos de Clasificación Basada en el Modelo Estructural de Cubrimientos. Tesis grado de doctor. CIC-IPN, México, 2006.
[168] Zhuravliov, Yu. I. Acerca del enfoque algebraico para la solución de problemas de reconocimiento o clasificación. Revista Problema Kibernetiki, 33, pp. 5-68, Moscú (1978).
[169] Zhuravliov, Yu. I. Algebras correctas sobre conjuntos de algoritmos incorrectos (heurísticos) I. Revista Cibernética, 4, pp. 14-21, (1977).
[170] Zhuravliov, Yu. I. Algebras correctas sobre conjuntos de algoritmos incorrectos (heurísticos) II. Revista Cibernética, 6, pp. 21-27, (1977).
[171] Zhuravliov, Yu. I. Algebras correctas sobre conjuntos de algoritmos incorrectos (heurísticos) III. Revista Cibernética, 1, pp. 35-43, (1977).
[172] Turkey, Jhon. Analisis exploratorio de datos. Adison Wesley, Reading MA. New York, USA, 1977.
[173] Yoav Benjamini. “Apertura de la caja de un Boxplot”. La Estadística de América. Vol. 42 (4), pp. 257-262, USA, 1988.

## SITIOS WEB

[174] WWW.CS.WAIKATO.AC.NZ/ML/WEKA
[175] WWW.SPSS.COM/CLEMENTINE
[176] WWW.ATTAR.COM

[177] HTTP://SMRL.OTAGO.AC.NZ/

# APENDICE A

# 1. SISTEMA CIADEC

En esta segunda parte del proyecto de tesis, se hace una breve introducción al sistema *CIADEC,* se describe su estructura y sus funcionalidades.

## 1.1 INTRODUCCIÓN

El sistema *CIADEC* implementa el modelo que se propone en este trabajo de tesis definido en el capítulo 4, cuyo título es "Caracterización e interpretación Automática de Descripciones Conceptuales en Dominios poco Estructurados" [50], el que permite caracterizar las diferentes clases a partir de una clasificación previamente establecida, en dominios poco estructurados y obtener automáticamente interpretaciones conceptuales de éstas, con respecto a variables cuantitativas.

## 1.2 DISEÑO MODULAR DEL SISTEMA *CIADEC*

El sistema *CIADEC*, surge de la necesidad de automatizar la caracterización e interpretación de clases en dominios poco estructurados previamente particionados combinando conceptos, técnicas de inteligencia artificial, estadística y lógica difusa. Mediante la automatización se persigue reducir el tiempo necesario para llevar a cabo esta tarea, agilizando tanto las actividades asociadas al análisis de datos como a la obtención de información relevante que posteriormente sea útil en la gestión y/o toma de decisiones en esos dominios.

### 1.2.1 Arquitectura del sistema CIADEC.

La entrada del sistema es la matriz de datos $X$ y la partición de referencia $P$, teniendo como salidas, según la opción del usuario:

- La asignación de clases a un conjunto de objetos nuevos
- La calidad de asignación del sistema de reglas

La representación gráfica de las funciones de pertenencia por variable. Dichos gráficos son generados en código *LaTeX* y se pueden exportar a cualquier documento o bien ser visualizados en pantalla conectando con el visualizador de *LaTeX*. Este tratamiento se adecua a la filosofía de otras herramientas que se comparten en el mismo equipo de trabajo y que, en un futuro, se han de integrar en una herramienta de minería de datos.

A nivel conceptual el sistema *CIADEC* esta formado por los siguientes cinco módulos:

- *Módulo I*. Generación de Intervalos de Longitud Variable (GILOVA).
- *Módulo II*. Generador de Tablas de Distribuciones (funciones de pertenencia) Condicionadas a Clases (GETADI).
- *Módulo III*. Generador de Sistemas de Reglas (GESIRE).
- *Módulo IV.* Generador de Gráficos de Funciones de Pertenencia de $X_k | C$ (GEGRALA).
- *Módulo V.* Validación (VALIDA).
- *Módulo VI.* Interpretación (GETAIN)

La integración de los módulos anteriormente mencionados, tiene una amplia interrelación para la validación del sistema de reglas y la predicción de clases para nuevos individuos que no poseen esta última. En la Figura 1 se muestra la integración de funcionalidades de los módulos que forman la arquitectura del sistema *CIADEC*.

**Módulo I. Generación de intervalos de longitud variable (GILOVA).**

Dada la variable de estudio $X_k$ este módulo tiene la prioridad de generar un sistema de intervalos de longitud variable calculando los valores $I^k$. La llamada a la función principal del módulo es:

$$\text{TablaInterv } (X, X_k, P, \text{int } nclases)$$

**Entrada**: Los parámetros de entrada en este módulo son: la variable a seleccionar $X_k$, la matriz de datos $X$ y en su caso la partición $P$ del conjunto de datos y el número de clases *nclases* de la partición.

**Salida:** Para cada atributo seleccionado $X_k$, como salida un vector que representa el sistema de intervalos de longitud variable y cuya estructura es la de un fichero con extensión iks, digamos el nombre de la variable en estudio y representado por < $X_k$ .*iks* >.

**Descripción:** En este módulo se realizan las siguientes funciones al hacer la llamada a la función principal y son: Construir un vector con el mínimo y máximo de $X_k$ en cada clase, ordenar los valores de ese vector de menor a mayor y generar el fichero *.iks*. La función principal se encuentra programada en la clase Gilova.java en relación con TablaInterv.java que generan tablas de intervalos de acuerdo a los máximos y mínimos de cada variable en cada clase.

**Módulo II. Generación de la tabla de distribuciones condicionadas a intervalos (GETADI).**

El módulo II GETADI tiene como principal objetivo la generación de la tabla de distribuciones condicionadas a intervalos (o funciones de pertenencia). La llamada a la función principal es:

TablaFrec $(X, X_k, P, short\ nClases, short\ nInterv)$

**Entrada:** Este módulo tiene como entrada el conjunto de datos $X$, la variable seleccionada $X_k$, la partición $P$, el número de clases *nClases* y el número de intervalos *nInterv*.

**Salida:** La salida en este módulo es una tabla de distribuciones condicionada a intervalos de la forma $P|I^k$ representada por medio de un archivo, el nombre de la variable y con extensión *dci* (< $X_k$ .*dci* >).

**Descripción:** En esta parte se lee nuevamente la columna $X_k$ de la matriz de datos $X$, ubicando cada valor $x_{ik}$ en el intervalo $I_s^k$ y clase *C* correspondiente, contando el número de elementos en cada una de las casillas de la tabla $P|I^k$ desde $k = 1,2,3,....,2\xi - 1$ y posteriormente sumando columnas obtenemos los $n_{I_s^k}$ para dividir cada casilla por su $n_{I_s^k}$ y determinar las probabilidades $p_{sc}$ correspondientes. La función principal se encuentra programada en la clase Getadi.java en relación con TablaFrec.java que generan tablas de frecuencias.

**Módulo III. Generación de sistema de reglas (GESIRE).**

Este módulo llamado GESIRE genera primero, un sistema de reglas difusas de inducción basado en la matriz de distribuciones condicionada a intervalos por variable seleccionada $X_k$, $\Re(X_k, P)$ en un archivo de tipo *.srg* para el sistema de reglas completo y una vez que hemos elegido un cierto criterio de agregación para reducir la ambigüedad inherente al sistema de reglas obtenemos un sistema de reglas reducido $\Re^*(X_k, P)$ que se guarda en un archivo tipo *.srr*. La llamada a la función principal del módulo es:

GeneradorReglas ( $P \mid I^k$, $X_k$, *short nClases*, *short nInterv*)

**Entrada:** La entrada es un archivo con extensión *.dci* que representa la tabla $P \mid \mathrm{I}^k$ de distribuciones condicionada a intervalos de la variable $X_k$ en estudio y sus dimensiones *nClases* y *nInterv*.

**Salida:** La salida es a dos niveles dependiendo del interés del usuario: si se desea el conjunto global de reglas es un archivo con extensión *src*, por otro lado si se desea elegir algún criterio de agregación la salida será un archivo con extensión *srr*.

**Descripción:** En este módulo se construye el sistema global de reglas $\Re(X_k, P)$ por variable seleccionada $X_k$ a partir de la tabla de distribuciones

condicionada a intervalos y considerando un criterio de agregación (pj., el de probabilidad máxima) obtenemos un sistema reducido de reglas $\Re^{*}(X_k, P)$.

**Módulo IV. Generación de Gráficos en LATEX (GEGRALA).**

Este módulo denominado GEGRALA tiene como objetivo generar gráficos para las funciones de pertenencia $f(X_k \mid \mathrm{C})$. La llamada a la función principal del módulo es:

$$\text{GeneradorGrafico } (< X_k.dci >, X_k.tex, X_k)$$

**Entrada:** El fichero de entrada para la generación de gráficos *LATEX* es la tabla $P \mid \mathrm{I}^{\mathrm{k}}$ representadas por un archivo con extensión *dci* $(< X_k.dci >)$.

**Salida:** La salida es de dos tipos:

- Como archivo con extensión *.tex*, con el nombre de la variable de estudio $(< nombre\_\mathrm{var}iable.tex >)$.
- Y visualización en pantalla del gráfico.

La estructura de los archivos *tex* es la que sigue un formato de archivo en *LATEX* (ver tabla 5) que dibuje el gráfico.

**Descripción:** En este módulo la generación de gráficos de las tablas $P \mid \mathrm{I}^{\mathrm{k}}$ de distribuciones se hace a partir de la tabla de distribuciones generada en el módulo II y considerando las operaciones de transformación sobre la generación de gráficos en *LATEX*

**Módulo V. Validación del Sistema de Reglas (VALIDA)**

Este módulo representa la etapa final del proceso y es donde se realiza la validación del sistema de reglas que hemos obtenido cuando existe un conjunto de validación $P_0$, primero para cada una de los atributos $X_k$ y luego una vez hecha la asignación de la clase correspondiente *C* a cada uno de los individuos $i$ del conjunto de prueba $P_0$, comparándola con la clase de referencia asociada a cada uno de éstos, obtener la calidad de asignación del sistema. La llamada a la función principal del módulo es:

$$Valida\mathrm{Re}glas\ (FileDate\ P_0.dat, short\ nInterv)$$

**Entrada:** Recibe como entrada el conjunto de prueba $P_0$ donde cada individuo tiene asignada la clase $(C)$ que le corresponde de acuerdo a la clasificación de referencia y por otro lado, la tabla $I^k \mid \mathrm{P}\ \ (< X_k.dci >)$. Con esta tabla y el valor $x_{ik}$ se puede calcular la clase de predicción (Ĉ) de cada

individuo y aplicando el criterio de probabilidad máxima obtener el sistema reducido de reglas.

**Salida:** La salida es un fichero con extensión *tcp* que determina el número de coincidencias entre las clases de referencia ( $C$ ) y de predicción ($\hat{C}$) para los elementos de $P_0$.

**Descripción:** En este módulo se hace una comparación cruzada entre la clase asignada (la de referencia) y la de predicción (la asignada debido al sistema de reglas reducido) para determinar el grado de confiabilidad de nuestro sistema de reglas. El error se calcula en %.

La Figura .1 muestra la arquitectura modular del sistema CIADEC 2.0.

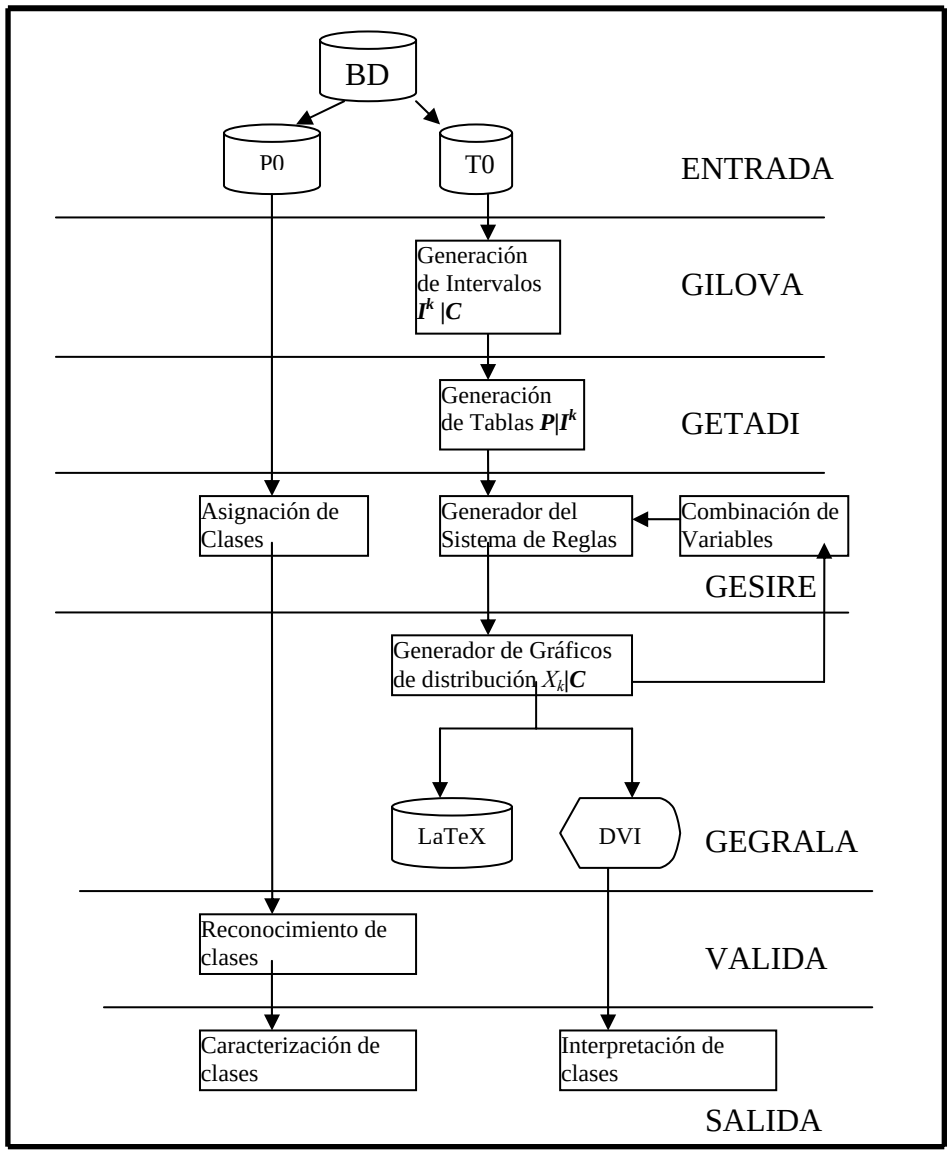



Figura 1. Integración conceptual de los módulos del sistema *CIADEC* 2.0

### 1.2.2 Generación de Gráficos en *LaTeX*

En esta parte se propone una forma de representar gráficamente este sistema de reglas que permite obtener conocimiento útil y comprensible para la interpretación conceptual de las clases identificadas.

A nivel diseño este módulo se hizo en forma imbricada a dos niveles para la reutilización del código LaTeX de estos gráficos en documentos posteriores.

1. **A nivel de figura principal**. La generación del paquete de grafos de interpretación para la partición P de un cierto dominio. Para hacer esto, se genera una figura principal en la que están imbricados todos los grafos $Gc$, 1 ≤ c ≤ ξ, donde ξ es el número de clases de la partición P (20 en nuestro caso). En este nivel generamos una figura principal con las siguientes convenciones e instrucciones de *LaTeX*.

   Definimos: $w = 540$ el ancho de la figura grande, $h = 700$ la altura de la figura principal (paquete), $(x_0, y_0)$ = $50,-175)$ el origen de la figura (esquina inferior izquierda). Las instrucciones en *LaTeX* a este nivel por página son:

   - % Contenido de la figura principal para la variable $X_k$.
   - *\begin {figura} {\setlength {\unitlength{1pt}}}*
   - *\begin{picture}(540, 700)(50, −175)*
     Para cada clase $C_i$ el origen de cada grafo se coloca en la posición (0,140·c), donde c varía de 0 a 3 en cada hoja tamaño a4, utilizando la instrucción:
   - *\put(0, 140·c){* $Gc_i$*,}*
     donde $Gc_i$, es el código *LaTeX* del grafo para la clase $C_i$, con 1 ≤ $C_i$ ≤ ξ.
   - *\end{picture}*
   - *\end{figure}*

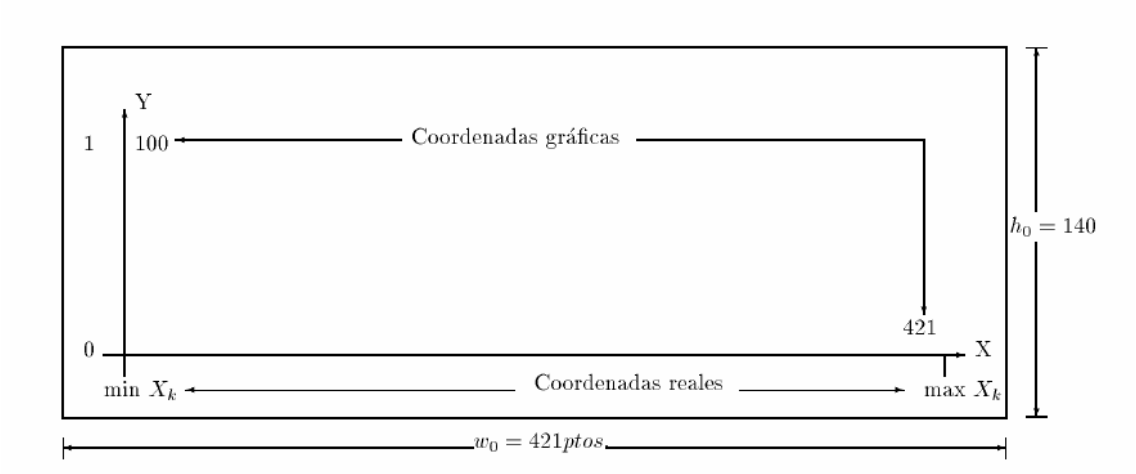



Figura 2. Diagrama de la figura principal

2. **A nivel de grafos.** La generación para cada una de las clases de la partición $P$ de los grafos $\{Gc_i\}$, se hace tomando en cuenta los siguientes elementos.

    1) Marco de cada grafo $\{Gc_i\}$.
    2) Etiqueta del grafo correspondiente a la clase $C_i$.
    3) Trazo, graduación y etiquetas del eje de las X's, así como las marcas de límites de los intervalos sobre este eje.
    4) Trazo, graduación y etiquetas del eje de las Y's, así como las marcas de las probabilidades sobre este eje.
    5) Función de pertenencia correspondiente al grafo $Gc_i$.

    El Marco de cada grafo $\{Gc_i\}$. A este nivel de grafo para cada clase $C_i$ se deberán tener los siguientes datos: Siendo, $w_0 = 421$ el ancho, $h_0 = 140$ la altura del grafo, $(x_0, y_0) = (50, -175)$ la esquina inferior izquierda del marco de cada grafo, $C_i$ el índice de las clases, con $C1 \le C_i \le C\xi$, donde en este caso ξ es el número de clases (en nuestro caso 20), este marco se obtiene con las siguientes instrucciones en *LATEX*:

    ```
    \begin{picture}(421, 130) (0, 0)
    {Elementos del grafo}
    \end{picture}
    ```

3. **Los elementos del grafo son:**

a) Etiqueta del grafo correspondiente a la clase $C_i$.
   `\put(5, 100){${\cal C }_{c}$}`

b) Con respecto al eje de las X's

- Trazo del eje de las *X's*.
- `\put(0, 0){\line(1, 0){421}}`
- Graduación sobre el eje *X*. Representar las marcas graduales de longitud 10 sobre el eje *X*, dividimos la longitud $l_x$ del eje *X* entre 8 y utilizando la siguiente instrucción:
  Para $i \leftarrow i+1,\ 0 \le i \le 8$
  `\put(`$(l_x/8)\cdot i$`, 0){\line(0, −1){10}}`
- Etiquetas sobre el eje *X*. Representar las etiquetas $\{e_i\}$ de las marcas sobre el eje *X* cada $\dfrac{M^k - m^k}{8}$ unidades a partir de la primera marca vertical, que indiquen la abcisa de la variable $X_k$

que se está representando. La marca *0* coincide con el mínimo de la variable $X_k$, en general $e_i = m^k + \frac{M^k - m^k}{8} \cdot i,$ con $0 \le i \le 8$.

Estas etiquetas se situarán exactamente debajo de cada marca con lo que sus coordenadas en *X* serán las mismas que para las marcas y las de *Y* serán constantes a -20 ptos., considerando que 10 puntos por debajo del eje *X* están ocupados por la propia marca y reservamos 10 puntos para la etiqueta.

Para $i \leftarrow i,\ 0 \le i \le 8$

$\backslash put((\frac{l_x}{8}) \cdot i, -20)\ \backslash \mathrm{mbox[c]\{e_i\}})$

- Marcas de límites de los intervalos sobre el eje *X*. Son marcas de longitud 5 sobre el eje X, que representan los límites de los intervalos $I_s$ de la variable $I^k$ sobre el eje *X* y convenientemente reescalada sobre el gráfico, con un factor de escala $T_x$ entre el rango de la variable $X_k$ y la longitud $l_x$ del eje *X*. Así, si el rango de la variable $X_k$ es $[m^k, M^k]$, donde: $M^k$ es el máximo y $m^k$ es el mínimo para la variable $X_k$ y la longitud del eje *X* es $l_x$ puntos entonces un valor $x$ cualquiera estará posicionado en el grafo en la posición dada por la siguiente transformación:

$$x' = \frac{x - m^k}{M^k - m^k} \cdot l_x = (x - m^k) \cdot \frac{l_x}{M^k - m^k} = (x - m^k) \cdot T_x$$

Las coordenadas de los intervalos están distanciadas $|\mathrm{I}_1|, |\mathrm{I}_2|, |\mathrm{I}_3|, ....., |\mathrm{I}_{2\xi\text{-}1}|$ respectivamente. Sus posiciones sobre el eje *X* son:

$$m^k + |\mathrm{I}_1|, \mathrm{m}^k + |\mathrm{I}_1| + |\mathrm{I}_2|, ....., \mathrm{m}^k + |\mathrm{I}_1| + |\mathrm{I}_2| + ..... + |\mathrm{I}_{2\xi\text{-}1}|$$

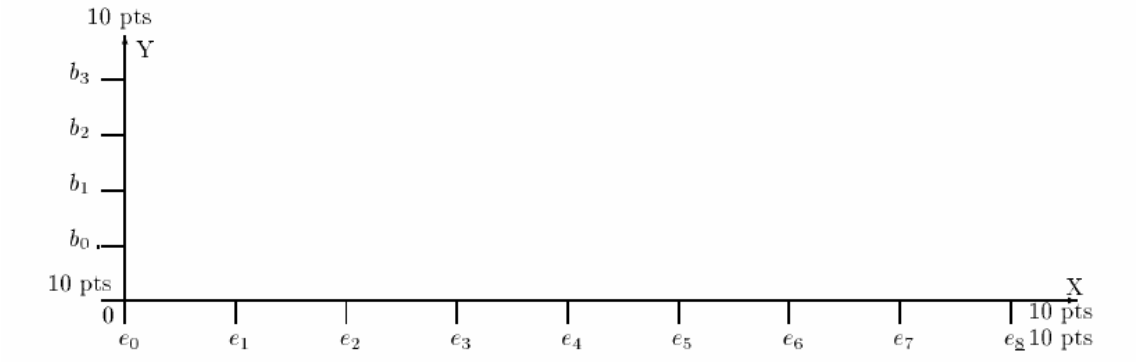



Figura 3. Posiciones de las etiquetas

Con lo cual los límites del intervalo $I_s$ serán las posiciones

$$[\min I_s = \max I_{s-1}, \max I_s]$$

Y en términos de las magnitudes de los intervalos $I^k$ igual a:

$$[\mathrm{m^k} + |\mathrm{I_1}| + |\mathrm{I_2}| + .....+ |\mathrm{I_{\xi\text{-}1}}|,\ \mathrm{m^k} + |\mathrm{I_1}| + |\mathrm{I_2}| + .....+ |\mathrm{I_{\xi}}|]$$

Así, transformando la marca del límite superior del intervalo $I_j$ sobre el eje *X* está dada por la siguiente instrucción:
Para j ← j+1, $0 \le j \le 2\xi - 2$:

$$\backslash put(\textstyle\sum_{s=1}^{j} |\mathrm{I_s}| \cdot \mathrm{T_s}, 0)\{\backslash \mathrm{line}(0,-1)\{5\}\},$$

donde, $|\mathrm{I_s}| = |\limsup \mathrm{I_s} - \liminf I_s|$ representa la longitud del intervalo $I_s$.

c) Con respecto al eje de las Y's

- Eje de las *Y's*.
  $\backslash put(0,0)\{\backslash \mathrm{line}(0,1)\{100\}\}$

- Graduación sobre el eje *Y*. Representar las marcas graduales de longitud 10 sobre el eje *Y*, dividimos la longitud $l_y$ del eje *Y* entre 4 y utilizando la siguiente instrucción:
  Para i ← i + 1, *0: i: 3*
  $\backslash put(0,\ (\frac{l_y}{4} \cdot i))\{\backslash \mathrm{line}(-1,0)\{10\}\}$

- Etiquetas sobre el eje *Y*. Representar las etiquetas $\{b_i\}$ de las marcas sobre el eje *Y* cada $\frac{l_y}{4}$ unidades a partir de la primera marca horizontal, que indiquen la probabilidad que se está representando. La marca *0* coincide con la probabilidad 0 y en general se tiene que:
  Para $i \leftarrow i+1,\ 0 \le i \le 3,\quad b_i = \frac{l_y}{4} \cdot i$ y su ubicación utilizando la siguiente instrucción:

  $\backslash put(-25, (\frac{l_y}{4} \cdot i))\ \backslash \mathrm{mbox[c]}\{\mathrm{b_i}\}$

- Marcas de las probabilidades sobre el eje *Y*. Son marcas de longitud 5 sobre el eje *Y*, que representen los valores de la distribución de probabilidad de $X_k$ condicionada a los intervalos

$I^k$ de la clase $C_i$. Considerando la longitud del eje *Y*, $l_y = 100$ puntos, tenemos que el factor de escalamiento *R = 100* y la localización de estas probabilidades es directa. Esto es, $y' = 100 \cdot y$, quedando las marcas para la clase $C_i$ de la siguiente forma:

Para la clase $C_i$ se tiene:

$j \leftarrow j + 1$, $0 \leq j \leq 2\xi - 2$ y su ubicación utilizando la siguiente instrucción:

$$\backslash put(0, p_{jc} \cdot 100)\{\backslash line(-1,0)\{5\}\}$$

**4. A nivel de función de pertenencia.** A este nivel se tiene dos pasos importantes:

- Marcar los límites de los intervalos $I_s$ sobre el eje *X*.
- Dibujar la función escalonada que sobre cada $I_s$ vale $p_{sc}$. Las coordenadas del grafo $Gc_i$ de la función de pertenencia para la clase $C_i$, en el intervalo $I_s$ son

$$((m^k + \sum_{j=1}^{s-1} |I_j|), p_{sc})$$

donde: $I_s$ es el s-ésimo intervalo de $I^k$, $p_{sc}$ es la probabilidad sobre el s-ésimo intervalo y $|I_j|$ es la longitud del j-ésimo intervalo de $I_k$.

Transformando estas coordenadas para ubicarlas en el marco del grafo se tiene:

$$(\sum_{j=1}^{s-1=39} |I_j|, p_{sc} \cdot 100)$$

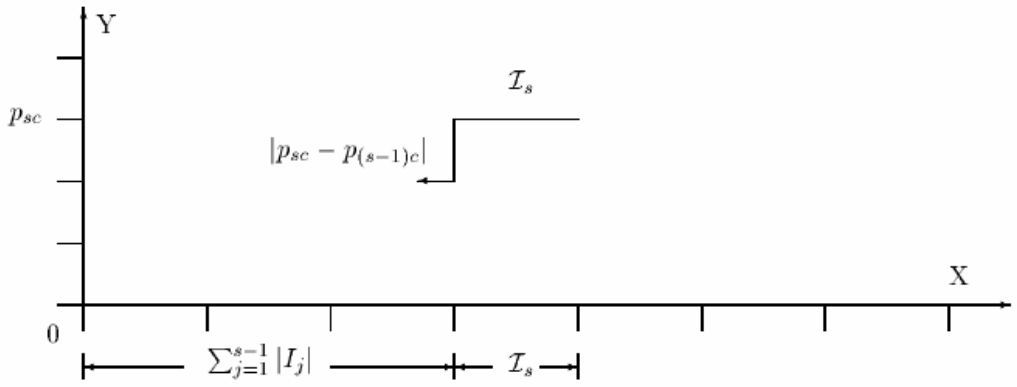



Figura 4. Función de pertenencia para la clase $C_i$

Sobre este intervalo, la función de pertenencia toma valores de línea horizontal de longitud $|\mathrm{I_s}|$, a fin de tener una función continua, uniendo los segmentos horizontales con otros verticales que salvará el salto entre $p_{(s-1)c}$ y $p_{sc}$. Situados en el origen de coordenadas trazamos una línea continua vertical hasta la primera probabilidad y una horizontal sobre el primer intervalo, luego subimos o bajamos hasta el valor de la segunda probabilidad y a continuación una segunda horizontal sobre el segundo intervalo y así, sucesivamente hasta el trazo de una última horizontal sobre el último intervalo $I_{2\xi-1}$. Definimos ahora $I_0 = 0$ y $p_{0c} = 0$, utilizando las siguientes instrucciones, calculamos los segmentos verticales sobre el eje *Y* correspondientes a los saltos de la función de pertenencia y vamos dibujando el valor de dicha función sobre cada $I_s$. Para *s ← s + 1, 0 ≤ s ≤ 2ξ − 2 = 38*

Si $p_{sc} - p_{(s-1)c} \geq 0$ entonces:

$$\backslash put((\sum_1^s |\mathrm{I_{s\text{-}1}}|\cdot \mathrm{T_x}, p_{sc}\cdot T_y)\{\backslash line(0,1)\{p_{sc} - p_{(s-1)c}\}\}$$

sino:

$$\backslash put((\sum_{j=1}^{s-1} |\mathrm{I_j}|\cdot \mathrm{T_x}, p_{sc}\cdot T_y)\{\backslash line(0,-1)\{|\, p_{sc} - p_{(s-1)c}\,|\}\}$$

después:

$$\backslash put((\sum_{j=1}^{s} |\mathrm{I_j}|\cdot \mathrm{T_x}, p_{sc}\cdot T_y)\{\backslash line(1,0)\{|\mathrm{I_s}|\cdot \mathrm{T_x}\}\}$$

Utilizando la misma forma imbricada de figuras, el otro tipo de gráfico en este módulo, el de gráficos que representan la combinación de atributos se hizo de la siguiente forma: Considerando como entrada el conjunto de prueba $P_0$, se lee para el primer individuo $i = 1$ el valor de la primer atributo $X_1$, se localiza el intervalo que le corresponde en la matriz de distribuciones condicionadas a intervalos de ese atributo y se toma ese renglón con sus clases y probabilidades correspondientes y se construye el primer gráfico; luego se lee el segundo atributo $X_2$ se localiza el intervalo a que corresponde en la matriz de distribuciones de ese atributo y se construye este gráfico, que representará todas las clases asociadas con sus correspondientes probabilidades y así, sucesivamente hasta el atributo $X_{17}$, de todo esto obtenemos 17 gráficos para el primer individuo. En seguida se considera el segundo individuo $i = 2$ y se repite el proceso anterior y así, hasta agotar todos los individuos del conjunto de prueba (30 elementos), de tal forma que obtengamos 17 gráficos que representan las clases y sus correspondientes probabilidades de los 17 atributos seleccionadas por individuos en el conjunto de prueba $P_0$.

A nivel algoritmo se definen dos constructores: Class GeneradorGrafico y Class GeneradorLaTeX.

## 1.3 ESTRUCTURA DE DATOS

En este apartado se explica la representación de datos y la estructura de archivos que el sistema CIADEC necesita para que funcione.

### 1.3.1 Representación de datos

Los individuos que forman el conjunto $T_0$ están descritos por una serie de atributos o características y pueden ser de dos tipos:

- Variables cualitativas ó categóricas: Corresponden a un tipo de característica de los individuos que se expresan mediante adjetivos. Estas variables cualitativas se dividen en ordinales dos y nominales.
- variables cuantitativas: Son características medibles y se expresan en forma numérica.

Si se dispone de *n* individuos y de *k* atributos que los describen, los valores de todas estas variables para el conjunto de individuos se representan mediante una matriz rectangular $X$ de dimensiones $(n,k)$. Las filas de la matriz contendrán la información de los individuos, mientras que las columnas hacen referencia a las variables. Si los individuos son caracterizados simultáneamente con variables cuantitativas y cualitativas, la matriz de datos se considera heterogénea.

A las observaciones no presentes en la matriz de datos $X$ se les denomina valores faltantes. En caso de valores faltantes les asignamos un "*" con valor NaN (Not a Number) para que sea tratable desde el punto de vista del algoritmo.

### 1.3.2 Estructura de archivos

En este apartado se la estructura de archivos que el sistema CIADEC necesita para que funcione. Las estructuras de los archivos que describen el flujo de datos en el sistema CIADEC son de dos tipos de entrada y salida y extensiones: dat, par, iks, dci, tex, srg, srr, tcp, vsg, vsm y coi.

- < nombre archivo.dat > Contiene la matriz de datos $X$ por renglones. Para cada individuo *u* objeto i hay una lista con las coordenadas que le definen en cada variable y su formato es el estándar de este tipo de archivo: Los elementos de una línea estarán separados por al menos un espacio. En la primera línea van los nombres de las variables, en caso de que no estén los nombres se asignarán a las variables los nombres por defecto NONAMEk, donde k es el número de la variable en

consideración. En la Tabla.1 muestra el formato de un fichero con extensión .dat.

| $x_1$ | $x_2$ | ... | ... | $x_n$ | XX |
|---|---|---|---|---|---|
| $v_{11}$ | $v_{12}$ | ... | ... | $v_{1n}$ | $id_1$ |
| $v_{21}$ | $v_{22}$ | ... | ... | $v_{2n}$ | $id_2$ |
| ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... |
| $v_{m1}$ | $v_{m2}$ | ... | ... | $v_{mn}$ | $id_m$ |

Tabla 1. Estructura de un archivo extensión .dat

- <nombre archivo.par> Cuando la partición de referencia no está incluida en la matriz de datos $X$, este archivo contiene información referida a dicha partición, su estructura es una columna que conserva el orden de asignación de la clase de referencia con respecto al orden de los individuos en el conjunto de datos de la matriz $X$. La Tabla 2 muestra el formato de un archivo con extensión par.

| nom_Obj | Clase |
|---|---|
| nom_1 | $id_1$ |
| nom_2 | $id_2$ |
| ... | ... |
| ... | ... |
| ... | ... |
| nom_n | $id_n$ |

Tabla 1. Estructura de un archivo extensión par

- <nombre archivo.iks> Un archivo con extensión iks contiene información sobre el sistema de intervalos de longitud variable correspondiente a la variable $X_k$, su estructura consiste de un renglón donde se encuentran los 2ξ valores límites del sistema de intervalos separados por al menos un espacio. La Tabla 3 muestra el formato de un archivo con extensión iks.

| $Z_1$ | $Z_2$ | $Z_3$ | ... | ... | $Z_{2\xi-1}$ | $Z_{2\xi}$ |
|---|---|---|---|---|---|---|

Tabla 3. Estructura de un archivo extensión iks

- <nombre archivo.dci> Un archivo con extensión .dci contiene información sobre la tabla de distribuciones condicionadas a intervalos para un cierto atributo $X_k$ cuya estructura es la siguiente: en la primera línea van los límites de los intervalos y en el resto de las casillas los valores de la función de pertenencia $p_{sc}$ por clase $C$, todos sus

elementos están separados por al menos un espacio. La Tabla 4 muestra el formato de un fichero con extensión dci.

| | | | | | |
|---|---|---|---|---|---|
| $p_{11}$ | $p_{21}$ | $p_{31}$ | $\cdots$ | $\cdots$ | $p_{(2\xi-1)1}$ |
| $p_{12}$ | $p_{22}$ | $p_{32}$ | $\cdots$ | $\cdots$ | $p_{(2\xi-1)2}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $p_{1\xi}$ | $p_{2\xi}$ | $p_{3\xi}$ | $\cdots$ | $\cdots$ | $p_{(2\xi-1)\xi}$ |

Tabla 4. Estructura de un archivo extensión .dci

- <nombre archivo.tex> Archivo que contiene la estructura de las instrucciones en código LaTeX de los gráficos de las funciones de pertenencia condicionadas a intervalos por atributo $X_k$ y por clase $C$. La Tabla 5 muestra el formato de un fichero con extensión .tex.

% Contenido de la *figura grande* para el atributo $X_k$
$\backslash begin\{figure\}$
$\{\backslash setlength\ \{\backslash unitlength\{1pt\}\}$
$\backslash begin\{picture\}(540, 700)(50, -175)$
% Para cada clase $Ci$ el origen de cada grafo se coloca en la posición
% $(0, 140 \cdot c)$, donde $c$ varía de 0 a 3 en cada hoja tamaño
a4, utilizando la instrucción:
$\backslash put(0, 140 \cdot c)\{G_C\}$
donde $G_C$ es el código LaTeX del grafo para la clase $C$, con $1 \leq C \leq \xi$.
$\backslash end\{picture\}$
$\backslash end\{figure\}$

Tabla 5. Estructura de un archivo extensión tex

- <nombre archivo.srg> Archivo que contiene la estructura del sistema de reglas globales $\Re(X_k, P)$ para la variable seleccionada $X_k$, el contenido de este archivo es recomendable que sea tipo latex. Es un archivo en forma de columna de (2ξ −1)(ξ) elementos, el sub-índice de la regla nos marca la posición en que se dispara la regla. La Tabla 6 muestra el formato de un archivo con extensión .srg.

$$
\begin{aligned}
r_{11}: \; & x_{ik} \in I_1^k \xrightarrow{p_{sc}} C1 \\
r_{12}: \; & x_{ik} \in I_1^k \xrightarrow{p_{sc}} C1 \\
& \dots \\
& \dots \\
r_{2\xi-1,\xi}: \; & x_{ik} \in I_{2\xi-1}^k \xrightarrow{p_{sc}} C\xi
\end{aligned}
$$

Tabla 6. Estructura de un archivo extensión srg

- <nombre archivo.srr> Archivo que contiene la estructura de un sistema reducido de reglas $\Re^*(X_k, P)$ para la variable seleccionada $X_k$ cuando se ha optado por escoger algún criterio de agregación. Como una primera aproximación en este trabajo hemos elegido el criterio de máxima probabilidad; así, que tenemos una regla por cada intervalo $I_s^k$ del sistema $I^k$. La Tabla 7 muestra el formato de un fichero con extensión srr.

$$
\begin{aligned}
r_1: \; & x_{ik} \in I_1^k \xrightarrow{p_{max}} C \\
r_2: \; & x_{ik} \in I_2^k \xrightarrow{p_{max}} C \\
& \dots \\
& \dots \\
r_{2\xi-1}: \; & x_{ik} \in I_{2\xi-1}^k \xrightarrow{p_{max}} C
\end{aligned}
$$

Tabla 7. Estructura de un archivo extensión srr

<nombre archivo.tcp> Archivo que contiene información sobre la comparación entre la clasificación de referencia ($C$) y la obtenida por el sistema de reglas ($\hat{C}$) en cada atributo, donde $c_{ij}$ es el número de coincidencias entre ambas clasificaciones para el atributo $X_k$. La Tabla 8 muestra el formato de un archivo con extensión .tcp.

| | $C_1$ | ... | ... | $C_\xi$ |
|---|---|---|---|---|
| $\hat{C}1$ | $c_{11}$ | ... | ... | $c_{1\xi}$ |
| $\hat{C}2$ | $c_{21}$ | ... | ... | $c_{2\xi}$ |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |
| $\hat{C}\xi$ | $c_{\xi 1}$ | ... | ... | $c_{\xi\xi}$ |

Tabla 8. Estructura de un archivo extensión tcp

<nombre archivo.vsg> Archivo que contiene la información de las probabilidades y consecuentes de las reglas que se disparan para los individuos

del conjunto de prueba $P_0$. La Tabla 9 muestra el formato de este tipo de archivo.

| No. | $C$ | $P$ | ... | $C$ | $P$ |
|---|---|---|---|---|---|
| 1 | $C_{11}$ | $p_{11}$ | ... | $C_{1\xi}$ | $p_{1\xi}$ |
| 2 | $C_{21}$ | $p_{21}$ | ... | $C_{2\xi}$ | $p_{2\xi}$ |
| ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... |
| n | $C_{n1}$ | $p_{n1}$ | ... | $c_{n\xi}$ | $p_{n\xi}$ |

Tabla 9. Estructura de un archivo extensión .vsg

<nombre archivo.vsm> Archivo que contiene la información de las probabilidades máximas y consecuentes de las reglas que se disparan para los individuos del conjunto de prueba $P_0$. La Tabla 10 muestra el formato de este tipo de archivo.

| No. | $C$ | $P_{max}$ |
|---|---|---|
| 1 | $C_1$ | $p_{1max}$ |
| 2 | $C_2$ | $p_{2max}$ |
| ... | ... | ... |
| ... | ... | ... |
| n | $C_n$ | $p_{nmax}$ |

Tabla 10. Estructura de un archivo extensión .vsm

<nombre archivo.coi> Archivo que contiene la información sobre la coincidencias entre las clases de predicción ($\hat{C}$) y la referencia ($C$) para cada uno de los individuos del conjunto de prueba $P_0$. La Tabla 11 muestra la estructura de este tipo de archivo con extensión .coi.

| No. | $\hat{C}$ | $C$ |
|---|---|---|
| 1 | $\hat{C}_1$ | $C_1$ |
| 2 | $\hat{C}_2$ | $C_2$ |
| ... | ... | ... |
| ... | ... | ... |
| n | $\hat{C}_n$ | $C_n$ |
| Número | de | coincidencias: |

Tabla 11. Estructura de un archivo extensión.coi