Computación y Sistemas Vol. 1 No. 4 pp. 213-227
© 1998, CIC - IPN. *ISSN 1405-5546* Impreso en México

Clustering based on rules and Knowledge Discovery in ill-structured domains

K. Gibert(1) & U. Cortés(2)
(1)Departament d'Estadística i Investigació Operativa (EIO)
(2)Departament de Llenguatges i Sistemes Informàtics (IA)
Universitat Politècnica de Catalunya*
Pau Gargallo 5. Barcelona 08028. Spain.
phone: +34 3 4017323 - Fax: +34 3 4015881
karina@eio.upc.es, ia@lsi.upc.es

Article received on November, 1997; accepted on January 20, 1998.

Abstract

It is clear that nowadays analysis of complex systems is an important handicap for either Statistics, Artificial Intelligence, Information Systems, Data visualization,...Describing the structure or obtaining knowledge from complex systems is known as a difficult task. It is innegable that the combination of Data analysis techniques (clustering among them) , inductive learning (knowledge-based systems), management of data bases and multidimensional graphical representation must produce benefits on this line.

Facing the automated knowledge discovery of *ill-structured domains* raises some problems either from a machine learning or clustering point of view. *Clustering based on rules (CBR)* is a methodology developed in [9] with the aim of finding the structure of *ill-structured domains*. In our proposal, a combination of clustering and inductive learning is focussed to the problem of finding and interpreting special patterns (or concepts) from large data bases, in order to extract useful knowledge to represent real-world domains, giving better performance than traditional clustering algorithms or knowledge based systems approach.

The scope of this paper is to present the methodology itself as well as to show how *CBR* has several connection points with Knowledge Discovery of Data. Some applications are used to illustrate this ideas.

Keywords: Knowledge discovery of data, data mining, clustering, metrics, qualitative and quantitative variables, mixed data, ill-structured domains.

*This research has been partially financed by *TIC'96*

1 Introduction

In apprehending the world, men constantly employ three methods of organization, which pervade all of their thinking: (*i*) the differentiation of experience into particular objects and their attributes; (*ii*) the distinction between whole objects and its parts and (*iii*) the formation and distinction of different classes of objects.

This paper deals only with the third method. Most practical activities, whether on an individual or social level involve classification. Clustering is a mathematical and computational approach for making classes on a set of individuals (also called objects or events). It has been used as a tool for very long from the point of view of Statistics, AI, and now in emerging fields like *Knowledge Discovery of Data (KDD), Data Mining (DM)* ...Objects are described by several numerical variables (see [2], [5]). The classes could simply define a partition over the set of objects or consist of a richer representation such as hierarchical or overlapping categories, and they may be interpreted as diagnoses, predictions...

These kind of methods are interesting from a Machine Learning point of view, because they open a door to the automated generation of classification rules — sets of rules oriented to determining a membership class for a given object —, extremely useful in knowledge-based environments, in particular the diagnostic oriented ones. Indeed, several well known expert systems, as MYCIN [31], MILORD [30] or others, are actually classifiers.

In some sense, classification can be seen as a process of building a knowledge model for a given domain. That is why these methods are also connected with *KDD* and (*DM*) [6].

KDD and *DM* are two very young fields of research. Since the earlier 1990s, great advances in computer technology has taken place and today is possible to generate very large data sets (*i.e.* measured in terabytes). The size of those Data Bases overcomes the actual capabilities of traditional methods of data analysis and machine learning for analyzing, summarizing and extracting knowledge from them. Intelligent tools are needed for the autommatic explotation, as well as for assisting the user in the analysis of those data in order to focuss on the important knowledge; this is the context of *KDD*.

While *KDD* is the overall process of dealing with data to obtain useful knowledge contained in it — including pre and post-processing of the data as well as interpreting-oriented tools —, *DM* refers to those techniques for extracting patterns from raw data [6] which, combined with other tools, leads to a *KDD* process. *DM* is a wide field including, of course, Statistics in general and clustering in particular.

In fact, we agree with the idea that a number of real applications in *KDD* either require a clustering process or can be reduced to it [25]. From this point of view, clustering techniques are an important pile for what is known as *KDD*. *Clustering based on rules* is a methodology developed in [9] with the aim of improving clustering in the specific context of *ill-structured domains.*

Ill-structured domains (ISD) (in [14] a characterization of them may be found) are frequent in real applications. Mental disorders, sea sponges, books classification, fossils, stars... are examples of *ISD*.

Actually, *ISD* refers to complex systems where the consensus among experts is weak — and sometimes non-existent. Nontheless, experts use to have some prior knowledge on the structure of the domain — which is hardly taken into account by clustering methods—. On the other hand, when describing the objects, the use of qualitative variables become very common. Sometimes it is difficult to quantify the concepts, even when they are intrinsically numerical. For example, although the **size of a data base** is clearly quantitative, very often is referred using some categories as *small data base*, *big data base*, *huge data base* and so on. In most cases, quantitative and qualitative information coexists in what we call *non-homogeneous* data bases.

As standard clustering methods were originally conceived to deal with quantitative variables, when qualitative variables appear, previous treatments on the data matrix are needed [11]. Those preprocessing of data either produce some loss of information or the introduction of a certain degree of arbitrariness which will for sure affect the results. In general, traditional clustering methods give bad performances on *ISD*. On the other hand, building a knowledge based system for an *ISD* is also very difficult. Relationships between variables are complex, great quantities of implicit knowledge is used by the experts and formulation of a *complete* knowledge base is almost impossible. So, systems with low predictive capacity use to be obtained as a result.

Clustering based on rules was designed to give better performance than traditional clustering algorithms or knowledge based systems approach. In our proposal, a combination of clustering and inductive learning is focussed to find and interprete special patterns (or concepts) from large data bases, in order to extract useful knowledge to represent real-world domains.

The scope of this paper is to present the methodology of *clustering based on rules* as well as to show how this methodology fits in the context of Knowledge Discovery of Data. First of all, *clustering based on rules* is presented in section §2; then, the connection points with *KDD* are emphasized.

Sections §2.1 and §2.2 focuss on particularities of the methodology. Considering that non-homogeneous matrices are to be analyzed in *ISD*, some modifications on the clustering technique were done. Details about the representation of the classes are introduced in §2.2.1, while §2.2.2, is concerned with a new family of metrics that can measure distances with messy data. The metrics structure of this family is proved and a proposal is made on the values of the parameters of the metrics family. This measure has been succesfully implemented in a clustering based on rules system called **Klass** [9] and [11], and applied to very different domains (sea sponges [12], stars of Milki Way [14], thyroid tests [32]...).

Section §2.3 presents some tools oriented to the autommatic characterization and interpretation of the classes.

Finally, some comparisons with other well-known statistical packages to contrast the performance of our proposal. Although this measure has been also used in other applications, using real data and great ammounts of objects, for the purpose of this paper a simulated study §4 and a data set studied by other authors §3 were chosen. The last section presents conclusions and future work.

2 The methodology: Clustering based on rules

As said before, this methodology was designed with the aim of solving some problems presented by Statistics and Machine Learning in the analysis of *ISD*. We decided

to combine tools from both knowledge areas to built a system able to improve the performance of either clustering and inductive learning in the context of *ISD*.

In this section, *clustering based on rules* (or rule-based clustering) is described. Like most *KDD* processes, it combines prior knowledge from the expert with an automatic clustering method (see §2.1). It is an iterative and interactive process — this are also common characteristics of *KDD* processes [6] —, structured in two major phases which finally organize the set of objects into a set of classes that are presumed to be interpretable: initially, there is a process of acquisition of the available background knowledge even if it is not a complete definition of the domain, followed by the clustering process *strictu sensu*. Figure 1 shows a *schema* of the methodology.

On the other hand, this methodology helps the user to explicit his prior knowledge relevant to the problem and there exists an interactive information exchange between the system and the user, in the following way: using the information automatically produced by the system, the expert fills in the gaps in his knowledge about the domain structure. Then, he can make this knowledge explicit. This implies a new information transfer to the system, which, in turn, will lead to the generation of new results.

The main idea is to allow the user to introduce *constraints* on the formation of classes; the expert provides them in a *declarative* way. Therefore, the conditions imposed by the expert induce a sort of *super-structure* on the domain. Clustering will be performed *within* this structure, respecting the user constraints, which may be based on *semantic* arguments. Finally, all the elements are integrated altogether in a global structure. Hierarchical clustering is especially suited for our purposes, mainly considering that the expert can provide heterogeneous knowledge, *i.e.* very specific knowledge of small parts of the domain, together with more general knowledge about other parts.

At the end of this process, the system has *acquired* the knowledge needed to organize the domain, and the expert has succeeded in making explicit his knowledge in a relatively friendly way.

With the set of objects $\mathcal{I} = \{i_1 \ldots i_n\}$, the steps to be followed are described below [9]:

1. **Initialization phase**

- First, an initial hierarchical tree τ^0 is obtained by clustering $\mathcal{I}$ with the algorithm described in §2.2.
- Determine the tree-cut $\mathcal{P}_0$. Tools presented in 2.3 can be used to decide the number of classes to be done as well as to interpret the *meaning* of the classes.
- Analyze $\mathcal{P}_0$ and determine a first set of logic rules $\mathcal{R}^0$ containing part of the expert knowledge on the studied domain.
- Step ξ ($\xi = 0$): Start iteration process

2. **Phase of background knowledge acquisition:**

- Determine the rule-induced partition $\mathcal{P}^{\xi}_{\mathcal{R}}$ on the basis of $\mathcal{R}^{\xi}$. Build a *residual class* C^{ξ}_0 with those objects for which no information is provided.
- **Conflict solving phase**: Analyze the objects selected by contradictory rules: If satisfactory, proceed to the classification step. Otherwise, return to the *background knowledge acquisition phase* and reformulate $\mathcal{R}$.

3. **Clustering phase:**

- **Clustering *within* expert constraints:**

  $\mathcal{P}_{\mathcal{R}^{\xi}}$ is *a priori* satisfying the expert requirements. Perform the clustering for each $\mathcal{C} \in \mathcal{P}^{\xi}_{\mathcal{R}}$ (see §2.2). Notice that every $\mathcal{C}$ is smaller than $\mathcal{I}$. Determine:

  i) The corresponding hierarchical trees $\tau^{\xi}_{\mathcal{C}}$,

  ii) Their prototypes $\vec{i}^{\xi}_{\mathcal{C}}$, by summarizing the class 2.2.1, and

  iii) Their masses $m^{\xi}_{\mathcal{C}} = card\ \mathcal{C}$.

- Add the prototypes $\vec{i}^{\xi}_{\mathcal{C}}$ to the residual class, as if they were ordinary objects, but taking into account their masses. The new data set is:

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

- **Integration phase**: Classify $\tilde{\mathcal{I}}^{\xi}$ to integrate all the trees $\tau^{\xi}_{\mathcal{C}}$, ($\mathcal{C} \in \mathcal{P}_{\mathcal{R}^{\xi}}$) in the sole τ^{ξ}.
- Cut τ^{ξ} into partition $\mathcal{P}^{\xi+1}$, either using heuristic criteria or automatic tools (§2.3) — which provide helpful material for *interpretation*—. Use the tools presented in 2.3 for interpreting the *meaning* of the classes.
- The expert has also to confirm that $\mathcal{R}^{\xi}$ improves the classification obtained with $\mathcal{R}^{\xi-1}$ in the desired way. Tables for comparing different classifications, or terms with major contributions to the distance between them can be used. *i*) If the classification is useful regarding expert goals, proceed to point 4. *ii*) If not, analyze the results to reformulate the rules set. Build $\mathcal{R}^{\xi+1}$, set ($\xi = \xi + 1$) and return to 3.

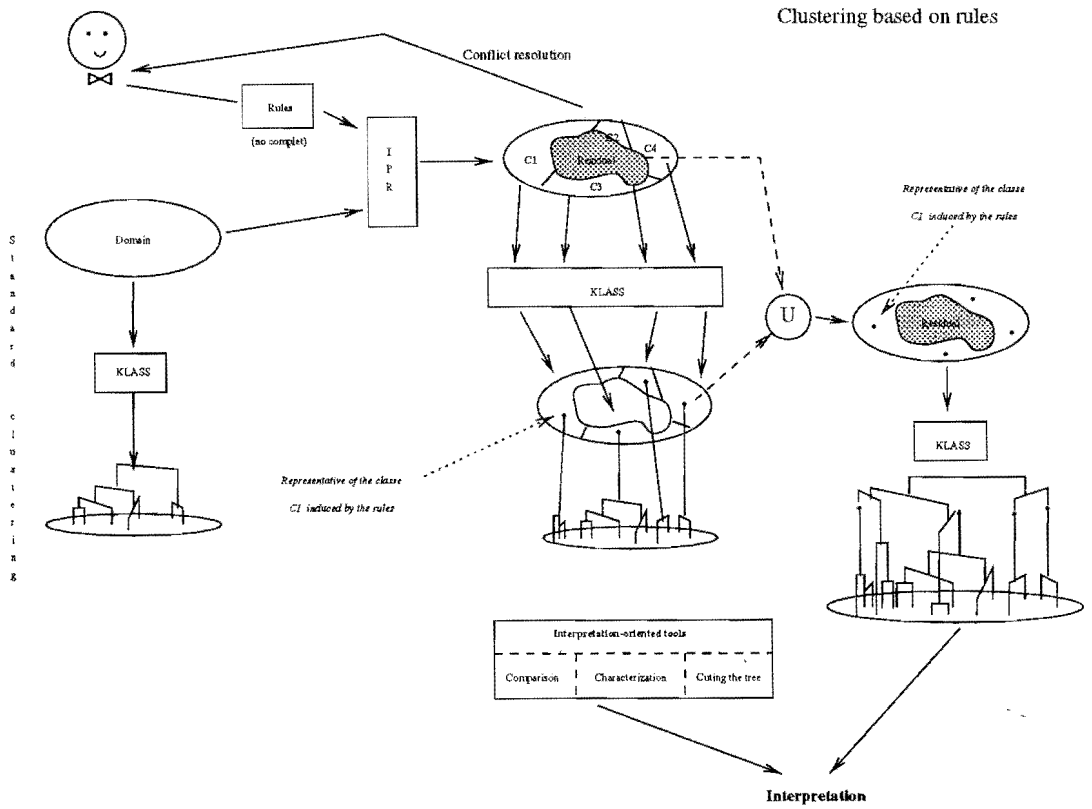


Figure 1: Methodology.

4. **Results production phase**:

   Automatic generation of final reports is available (indeed, LaTeX files can be obtained automatically). A description of the domain is obtained in terms of a prototype that summarizes the contents of each class. Our idea is to provide human interpretable descriptions of the classes.

   There is a last step oriented to the consolidation of the discovered knowledge related to automatic rules generation. This will be useful for later predictive goals.

This methodology is a new possibility to perform classifications of any heterogeneous data matrix as well as for building a set of rules to describe the knowledge contained in the domain. Especially good results are obtained when analyzing data from *ISD*. If we consider that Fayyad defines a *KDD* process as the "*overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms*"[6], it is clear that *clustering based on rules* fits closely this definition, and that the data mining technique is, in our case, the ascendant hierarchical clustering method, which was addapted to deal with non homogeneous data matrices. Following Fayyad, two important key points of *KDD* are: *i*) using domain knowledge and *ii*) domain characterization. It can be seen that those elements take important part in the methodology presented here.

2.1 Prior knowledge

Usually, in *ISD*, experts have some prior knowledge about the structure of the domain. Including this partial information into the clustering process leads to a quality improvement.

Knowledge is represented by means of *If-Then* rules in order to provide maximum flexibility and expressiveness to the expert. Using this approach, collecting prior knowledge becomes easier. There is no restriction on the structure of the antecedent of the rule and it provides a mean of including semantic restrictions on the formation of classes.

Only available knowledge is collected, even if it is a partial description of the domain. The clustering process, acting as a Data Mining tool, will discover, by itself,

the structure of those parts of the domain not described in the rules set.

2.2 The clustering

As said before, in the kernel of the *classification based on rules* there is a clustering process. An ascendant (agglomerative) hierarchical algorithm is chosen for several reasons: first, because hierarchical techniques are widely used in clustering (at present); also, because only a hierarchical technique allows the generation of a unique dendrogramme taking into account the generality level of every rule–induced class — by integration of their prototypical representation into the residual class. This is the key for working with rules of different degrees of generality.

The output of a hierarchical clustering method is a dendrogramme (see figure 11).

Klass uses an adaptation of the *chained reciprocal neighbours algorithm* [3], which is based on the concept of *reciprocal neighbours (RN)* . At every step, a pair of *RN* is aggregated in a new class[1]. The chained version is a quick algorithm of $O(n_{obj}^2)$ worst case complexity.

Significant work has been required on two specific points of the algorithm to allow classification of heterogeneous data matrices: class representation (see §2.2.1) and distance between individuals (§2.2.2).

2.2.1 Class representation. Summarization

Reciprocal neighbours algorithms often work with a representative of any class, treating classes and ordinary objects the same (which improves computational costs). On the other hand, definition of a representative for each class (or subclass) will provide *prototypical* (*conceptual*) descriptions of the classes, which can be understood as a *summary* of each class, very useful for their interpretation.

The calculation of the quantitative components of the centre of gravity of a class is easy. For the qualitative ones, a way to do that is provided here. In [12] the representation of the qualitative components of the centre of gravity of a given class are deduced and justified. For class $\mathcal{C} = \{i_1 \dots i_{n_C}\}$, $\mathcal{C} \subset \mathcal{I}$, where every $i \in \mathcal{C}$ is described by their values in variables $X_k, (k = 1 : K)$ in the form $i = (x_{i1} \dots x_{iK})$, the representative of class $\mathcal{C}$ is defined as $\bar{i}_\mathcal{C} = (\bar{x}_{\mathcal{C}1}, \dots, \bar{x}_{\mathcal{C}K})$, with

$$\bar{x}_{\mathcal{C}k} = \begin{cases} x_{ik} & if \forall i' \in \mathcal{C} : x_{ik} = x_{i'k} \\ \dfrac{\sum_{i \in \mathcal{C}} x_{ik}}{n_\mathcal{C}} & if X_k \text{ quantitative} \\ \left((f_\mathcal{C}^{k_1}, c_1^k), \dots (f_\mathcal{C}^{k_{n_k}}, c_{n_k}^k) \right) & if X_k \text{ qualitative} \end{cases} \quad (1)$$

with $f_\mathcal{C}^{k_j} = \dfrac{I_\mathcal{C}^{k_j}}{\sum_{j=1}^{n_k} I_\mathcal{C}^{k_j}} = \dfrac{I_\mathcal{C}^{k_j}}{n_\mathcal{C}}$,

and $I_\mathcal{C}^{k_j}$ = number of individuals of modality $c_j^k \in \mathcal{D}_k$ contained in[2] subclass $\mathcal{C}$.

and the vector $\left((f_\mathcal{C}^{k_1}, c_1^k), \dots (f_\mathcal{C}^{k_{n_k}}, c_{n_k}^k) \right)$ is the value of the representative of the class $\mathcal{C}$ for qualitative variable X_k.

Actually, the centre of gravity of qualitative variables, defined as in expression (1), can be considered as a generalization of the arithmetic mean for a domain where addition and product are meaningless operations[3].

On the other hand, taking into account that the ascendant hierarchical tree is a binary tree, recurrent expressions were developed for calculating the centre of gravity of a class using the centres of gravity of the two subclasses joined at each step. Thus, the complexity of calculating the centre of gravity is independent of the class size. This property is very interesting in the later iterations of the process, where the classes could contain a large number of objects.

From a formal point of view, it is remarkable that the recurrence found for qualitative variables is exactly the *same* as the existing one for quantitative variables. This is a nice property, that allows homogeneous treatment of qualitative and quantitative variables in the clustering process.

2.2.2 Mixed metrics

The reciprocal neighbours algorithm needs a distance defined on the space of objects, so as to identify reciprocal

[1] Two objects i and $i' \in \mathcal{I}$ are RN *iff* i is the nearest neighbour of i' and vice versa.

[2] $\mathcal{D}_k = \{c_1^k \dots c_{n_k}^k\}$ is the set of values that a qualitative variable X_k can take.

[3] Notice that a mean $\bar{x}_{\mathcal{C}k}$ can also be expressed in terms of the different values taken by the variable and the observed frequencies of these values: $f_\mathcal{C}^1 \, x_k^1 + f_\mathcal{C}^2 \, x_k^2 + \dots + f_\mathcal{C}^T \, x_k^T$, what remembers expression (1).

neighbour pairs. In fact, there are some proposals on this line, like [16] or [15], [19] or [17] presenting similarity coeficients to evaluate *proximitites* between individuals. Although **Klass** is parameterized on the metrics and other measures can easily be incorporated and used, a family of measures was introduced in [10], [13] and detailed in [11] allowing evaluation of distances between objects partially described by quantitative variables, and partially described by qualitative ones. This family of measures depend on two parameters $\alpha, \beta, \in \Re$ and it is defined as:

$$d^2_{(\alpha,\beta)}(i,i') = \alpha \sum_{k\in\mathcal{C}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{\beta}{n_\mathcal{Q}^2} \sum_{k\in\mathcal{Q}} d_k^2(i,i') \qquad (2)$$

what can be written as $d^2_{(\alpha,\beta)}(i,i') = \alpha d^2_\mathcal{C}(i,i) + \beta d^2_\mathcal{Q}(i,i)$.

where $k \in \mathcal{C}$ if variable X_k is quantitative and $k \in \mathcal{Q}$ if variable X_k is qualitative; s_k^2 is the variance of variable X_k; $n_\mathcal{Q}$ is the number of qualitative variables and $d_k^2(i,i')$ is the *contribution of* k^{th} *variable* to $d^2_{(\alpha,\beta)}(i,i)$ (see expression 4).

From a theoretical point of view, it has been demonstrated [11] that the proposed measure (called *mixed distance*) is indeed a metric[4] if:

$$\alpha = 0 \Longrightarrow \mathcal{C} = \emptyset \;\&\; \beta = 0 \Longrightarrow \mathcal{Q} = \emptyset \qquad (3)$$

This condition, which is not very restrictive, means that only when no qualitative variables are recorded in the data matrix the qualitative component can be ignored, and reciprocally for quantitative ones.

In fact, the mixed metrics is a weighting between a canonical normalized Euclidean distance for quantitative components and an enhanced χ^2-distance for qualitative ones, such that the complete incidence table is no longer explicitly built. χ^2 is a metrics commonly used in clustering for qualitative variables. It works on a transformation of the original matrix. For *ISD* it is of significant higher dimension, because usually, there are great number of modalities for each variable. We propose to calculate the distance between two qualitative components in the following way:

$$d_k^2(i,i') = \begin{cases} 0, & \text{if } x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}}, & \text{otherwise,} \\ & \text{for individuals} \\ \frac{(f_i^{k_s}-1)^2}{I^{k_s}} + \sum_{j\neq s}^{n_k} \frac{(f_i^{k_j})^2}{I^{k_j}}, & \text{if } x_{ik} = c_s^k, \text{and} \\ & i' \text{ is a class} \\ \sum_{j=1}^{n_k} \frac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}}, & \text{in general case} \end{cases} \qquad (4)$$

In formula (4), I^{k_j} represents the number of individuals of the sample that are in modality c_j^k; I_{k^i} is the number of individuals in the sample of the same modality as the element i for variable X_k; $f_i^{k_j}$ represents the proportion of individuals from the i^{th} subclass satisfying $X_k = c_j^k$ and n_k is the number of modalities of variable X_k, which is qualitative. In [11] details on the mixed metrics are provided.

In fact, $f_i^{k_j} = \frac{I_i^{k_j}}{\sum_{j=1}^{n_k} I_i^{k_j}}$.

Using this expression, it is possible to calculate $d_{(\alpha,\beta)}$ directly on the original data matrix. In consequence, a lot of calculations can be avoided, as well as physical storage space. This is a relevant advance when processing big data sets.

From the clustering point of view, when the relative distances among objects are preserved, the classes generated are the same (since the same aggregations are done in the same order). For hierarchical methods, the resulting dendrogrammes are also the same, except for an scale factor existing between them.

In consequence, the information provided by some pairs of distances, $d^2_{(\alpha_1,\beta_1)}(i,i')$ and $d^2_{(\alpha_2,\beta_2)}(i,i')$, is equivalent, since both of them produce *equivalent* classification trees. Using this idea, an equivalence relationship over this family of distances $d^2_{(\alpha,\beta)}(i,i')$ may be defined.

So any pair of distances, $d^2_{(\alpha_1,\beta_1)}$ and $d^2_{(\alpha_2,\beta_2)}$, such that one of them can be written as a scaling of the other will belong to the same equivalence class. Thus, the equivalence condition is the following:

$$d^2_{(\alpha_1,\beta_1)}(i,i') \equiv d^2_{(\alpha_2,\beta_2)}(i,i') \Longleftrightarrow \alpha_1\beta_2 = \alpha_2\beta_1$$

In [9] it is showed that $\equiv$ satisfies the properties of an equivalence relationship when

$$(\alpha,\beta) \in \Re^+ \times \Re^+ - \{(0,0)\}$$

Using some heuristic criteria, a proposal on the values of the weighting constants α and β is developed (see [12], [28]):

$$\alpha = \frac{n_\mathcal{C}}{d^2_{\mathcal{C}max^*}} \qquad \& \qquad \beta = \frac{n_\mathcal{Q}}{d^2_{\mathcal{Q}max^*}} \qquad (5)$$

where the $d^2_{\mathcal{C}max^*}$ is a truncated[5] maximum of the set $\{d^2_\mathcal{C}(i,i'), \forall i,i' \in \mathcal{I}\}$ and simmetrically for $\mathcal{Q}$.

The values of (α,β) induce an equivalence relationship over the mixed metrics family. It is then possible to work just with the quotient set, taking as the representative of each equivalence class $d^2_{\alpha_0,\beta_0}$:

$$\alpha_0 = \frac{\alpha}{\alpha+\beta} \qquad \& \qquad \beta_0 = \frac{\beta}{\alpha+\beta} \qquad (6)$$

Several applications showed satisfactory results using these values for classifying *ISD*. Section §3 is an example of that. Nonetheless, the system remains open to the use of other values upon the user choice.

[4] This enables the clustering using Ward's aggregation criterion, and all the clustering methods for metric spaces.

[5] The maximum is calculated after eliminating the 5% of extreme values.

2.3 Interpreting tools

Actually, given a partition (classification) of a big set of objects it seems necessary to introduce tools for assisting the user in the interpretation tasks, in order to establish the *meaning* of the resulting classes. Often it is not enough for the user to automatically built the classes, but to help him to understand *why* those classes where detected. This is another important topic of a *KDD* system, and this section gives some ideas about our own approach to this topic.

2.3.1 Class characterization

Some statistical packages, like **SPAD**, include several tools to orient the interpretation of a given classification, as the possibility of calculating the *contribution* of certain variable to the formation of a class. However, at the final stage, the interpretation itself should be done by the user in a non-systematic way. In this paper, a system to find one characterization of a given class in an automatic way is provided. It is based on the representative of each class. This is a summary of what is presented in paper [9].

Let Λ_C^k be the set of *eigen values* of a variable X_k for a given class C. It is defined as the set of values of X_k taken by some element of C that are not taken by any element out of C.

Then, a variable X_k is *characteristic* of a class C if $\Lambda_C^k \neq \emptyset$ and $\forall i \in C, x_{ik} \in \Lambda_C^k$. In *ISD* it is difficult to find characteristic variables for the classes of a given partition $\mathcal{P} = \{C_1, \ldots, C_\xi\}$. For our purposes it is also interesting to consider the variables X_k that are *partially characteristic* of a class C. They are defined as $X_k : \Lambda_C^k \neq \emptyset$.

A partition can be characterized by what we call a *characterization system (CS)*:

$$\mathcal{S} = \{(C, X_k, \Lambda_C^k) : C \in \mathcal{P} \& \Lambda_C^k \neq \emptyset\}$$

If $\mathcal{S}$ contains only a triplet for each class of $\mathcal{P}$, it is called a *minimal and complete characterization system*. Sometimes, the characterization system is not complete:

$$\exists C \in \mathcal{P} : \forall (C', X_k, \Lambda_{C'}^k) \in \mathcal{S} \rightarrow C \neq C'$$

A procedure to complete those kind of *CS* is also developed, based on making some Close-World Assumptions and using negative information. For the scope of this paper, it is no necessary to go into more details. In the application this method is used for automatically characterize the classes.

2.3.2 Comparing several classifications

Sometimes it is interesting to compare two classifications $\mathcal{P}_1, \mathcal{P}_2$ of the same set of objects. In particular, if there exist a reference partition, provided by the expert or so, comparison will turn on an evaluation of the *quality* of the results.

An index $\delta(\mathcal{P}_1, \mathcal{P}_2) \in [0, 1]$ was defined, to evaluate the differences between two classifications. *Grosso modo*, it can be interpreted as the percentage of cases not equally classified by $\mathcal{P}_1$ and $\mathcal{P}_2$. If $\mathcal{P}_1$ is a reference partition of the objects, then $1 - \delta(\mathcal{P}_1, \mathcal{P}_2)$ may also act as a quality coefficient (see the application presented in §micros). Significance test on that index is actually in progress (in [9] details are provided).

2.3.3 Deciding the number of classes

In hierarchical clustering, after a dendrogramme is built, an α-cut on the dendrogramme is required for obtaining the final partition on the objects. The level of that α-cut is usually decided by observing the dendrogram and the level indexes histogram. Big gaps of the cumulated inertia in successive classes determines good levels for the cut of the tree.

An heuristic based upon this idea has been implemented. So, the user may ask for ordering the best partitions of the tree owing to a maximization of the inter-classes inertia. It is recommended to choose, among the k best cuts, the one which provides a better interpretation. For example, in the section 4 this procedure recommends partitions in 2, 3, 5 ...for the classification of *DATASET 1*.

3 An application: Microcomputers

Among other applications [14], [12], where interpretation of the results usually requires a lot of backgfound knowledge on the domain, the one concerned with a set of data presented in [23], [15] has been selected. This is a well studied training set, which makes possible to compare the performance of clustering with mixed metrics against other methods on the basis of a common dataset.

The data matrix is about 12 american microcomputers described by 5 variables, three of which are qualitative: *Display, MP, Keys* (see the data matrix in table 1).

For these data, and using conceptual clustering proposed by Michalski [23], the classification showed in the first column of table (2) was obtained. This training set was also treated in [15] using both reciprocal nearest neighbours algorithm and the single linkage method with a similarity measure proposed in the same paper. The resulting clusters of each method are shown in columns 2 and 3 of table 2. All these classifications contain exactly 4 classes.

A local expert was also consulted. First of all, we want to point out that he intuitively classified the training set on the basis of most relevant variables. He considered that variables *ROM* and *Keys* were much less important for the characterization of microcomputers. Actually, relevance of variables could be used as a biasing rule; nonetheless, this is not taken into account by our system at present.

Objects	Id.	Display	RAM	ROM	MP	Keys
APPLE-II	AP	COLOR-TV	48	10	6502	52
ATARI-800	AT	COLOR-TV	48	10	6502	57-63
COMMODORE-VIC-20-A	CoA	COLOR-TV	32	11	6502A	64-73
COMMODORE-VIC-20-B	CoB	COLOR-TV	32	16	6502A	64-73
EXIDI-SORCERER	ES	B-&-W-TV	48	4	Z80	57-63
ZENITH-H8	ZH8	BUILT-IN	64	1	8080A	64-73
ZENITH-H89	ZH89	BUILT-IN	64	8	Z80	64-73
HP-85	HP	BUILT-IN	32	80	HP	92
HORIZON	Ho	TERMINAL	64	8	Z80	57-63
OHIO-SC.-CHALLENGER	OCh	B-&-W-TV	32	10	6502	53-56
OHIO-SC.-II-SERIES	OS	B-&-W-TV	48	10	6502C	53-56
TRS-80-I	TRI	B-&-W-TV	48	12	Z80	53-56
TRS-80-III	TRIII	BUILT-IN	48	14	Z80	64-73

Table 1: Data matrix for microcomputers.

Id.	MIC83 Conc. clust.	GOW92 Rec. Neigh.	GOW92 Sing. Link.	Klass	Exp.
AP	1	1	1	1	1
AT	1	1	1	1	1
CoA	1	3	3	1	1
CoB	1	3	3	1	1
ES	4	4	1	2	2
ZH8	3	4	4	3	3
ZH89	3	4	4	3	3
HP	2	2	2	4	3
Ho	4	4	4	3	4
OCh	1	3	1	2	2
OS	1	1	1	2	2
TRI	4	1	1	2	2
TRIII	3	4	4	3	3

Table 2: Different classifications of microcomputers provided by different algorithms and distances.

Next, he proceed to determine how the values of each categorical variable could be grouped. In fact, he was looking for the structure of qualitative variables (see table 3), based on his background knowledge and experience.

After that, the expert proposed three general classifications according to the values taken by variables *Display, ROM* and *RAM* respectively, and he accepted as meaningful any

Display	*TV*	*Black & White*
		Color
	Built in	
	Terminal	
Microprocessor	*Motorola*	*6502*
		6502A
		6502C
	Intel (and similar)	*Z80*
		8080A
	Hewlett Packard	

Table 3: Structure of categorical variables (by expert).

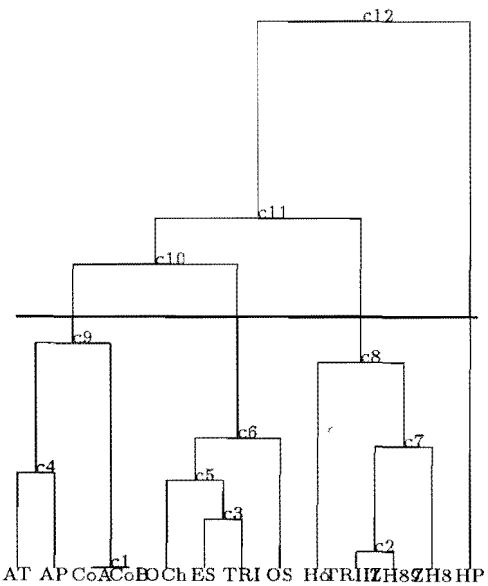


Figure 2: Dendrogramme for microcomputers. Mixed metrics and Ward's criterion with ($\alpha = 0.014$ and $\beta = 0.986$).

combination of this initial classifications. None of them had four classes, except the one regarding *Display*, which is shown in the last column of table (2).

On the basis of the distance defined in this paper, it is possible to perform a classification of the data using the Ward's criterion. The dendrogramme obtained using the mixed distance $d_{(\alpha,\beta)}$, with $\alpha = 0.014$ and $\beta = 0.986$ as suggested by formula (6) is presented in figure (2). A classification with four clusters has been chosen in order to make easier comparison against the other methods considered here. Extensional and prototypical representations of the produced classes are described in table (4).

As a first approach, the expert was asked to interpret the

Class			1	2	3	4
Proto-type de la classe	Display		COLOR-TV	B-&-W-TV	Built-in 3/4 Terminal 1/4	TERMINAL
	RAM		40	44	60	64
	ROM		47/4	9	31/4	8
	Micro-processor	6502	1/2	1/4		HP
		6502A	1/2	1/2		
		Z80			3/4	
		8080A			1/4	
		6502C		1/4		
	Keyword	52	1/4			92
		57–63	1/4	1/4	1/4	
		64–73	1/2		3/4	
		53–56		3/4		
Extensional description			APPLE-II ATARI-800 COMMODO-RE-VIC-20-A COMMODO-RE-VIC-20-B	EXIDI-SORCERER TRS-80-I OHIO-SC.-CHALLEN. OHIO-SC.-II-SERIES	ZENITH-H8 ZENITH-H89 HORIZON TRS-80-III	HP-85

Table 4: Intensional and extensional description of the classes proposed by Klass.

results from the different methods, in order to evaluate the performance of our metrics. From his opinion, the results of **Klass** were based on clear classification criteria: the *Display* followed, with less influence, by *Microprocessor*. This is clear from the prototypical descriptions provided by **Klass**. Michalski's proposal is also meaningful from the expert's point of view, whereas Gowda's results are less understandable in terms of finding clear clustering criteria. Other values of α, β give not so clear results too.

After that, we proceed to perform the automatic characterization of those partitions, following the definitions provided in §2.3. For the results provided by our approach, the interpretation given by the system is extremely similar to that provided by the expert:

$$\begin{aligned} \mathcal{S}_{\mathcal{P}_4} = \{ \quad & (\mathcal{C}_1,\ Display,\ Color-TV), \\ & (\mathcal{C}_2,\ Display,\ B\&W-TV), \\ & (\mathcal{C}_3,\ MP, \{Z-80, 8080A\}), \\ & (\mathcal{C}_4,\ MP,\ HP)\} \end{aligned}$$

This can be read in the following way: $\mathcal{C}_1$ is the class of the computers with color-TV display; $\mathcal{C}_2$ gathers those with black and white TV displays; $\mathcal{C}_3$ contains those computers with Intel micropocessor and $\mathcal{C}_4$ those with Hewlet-Packard microprocessor.

It is no possible to obtain a complete characterization for the classifications proposed in [23] and [15]. For example, for Michalski's results, the following system is obtained:

$$\begin{aligned} \mathcal{S}_{\mathcal{P}_1} = \{ \quad & (\mathcal{C}_1,\ MP, \{6502, 6502A, 6502C\}), \\ & (\mathcal{C}_2,\ MP,\ HP), \\ & (\mathcal{C}_4,\ MP,\ Z-80)\} \end{aligned}$$

and elements of $\mathcal{C}_3$ do not have any characteristic value for any variable. However, according to the expert intuition that

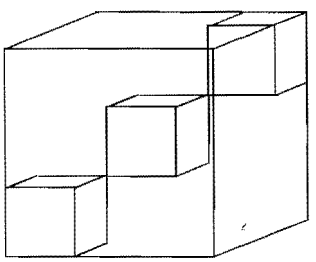


Figure 3: Target population.

Michalski's results are also interpretable, the CS may be completed by using negative information: conditioning to the element $(\mathcal{C}_3,\ MP,\ \overline{HP})$ the element $(\mathcal{C}_3,\ Display,\ Built-in)$ characterizes $\mathcal{C}_3$.

Trying to evaluate in an *objective* way the *proximity* between pairs of classifications, the distances between them are calculated (see 2.3). Again, the results produced by **Klass** are the more similar to the expert proposal — with $\delta(Exp, Klass) = 0.15$ what represents two objects (15%) classified in different classes —, followed by those presented in [23] — with $\delta(Mich, Exp) = 0.46$ of differences.

4 Comparing with other systems

In order to test the performance of the clustering based on rules, a sample of 150 points from three equal cubes located in the main diagonal of the unitary cube $[0,1]^3$ has been sim-

DATASET	X'_1	X'_2	X'_3	Data matrix type
1	X_1	X_2	X_3	Quantitative
2	Code / Interval: a [0.0,0.2), b [0.2,0.4), c [0.4,0.6), d [0.6,0.8), e [0.8,1.0]y	Code / Interval: a [0, 0.5), b [0.5, 1]	Code / Interval: a [0.0,0.25), b [0.25,0.50), c [0.50,0.75), d 0.75,1.00]	Qualitative
3	Code / Interval: a [0.0,0.2), b [0.2,0.4), c [0.4,0.6), d [0.6,0.8), e [0.8,1.0]	X_2	X_3	Mixed

Table 5: The three simulated data sets.

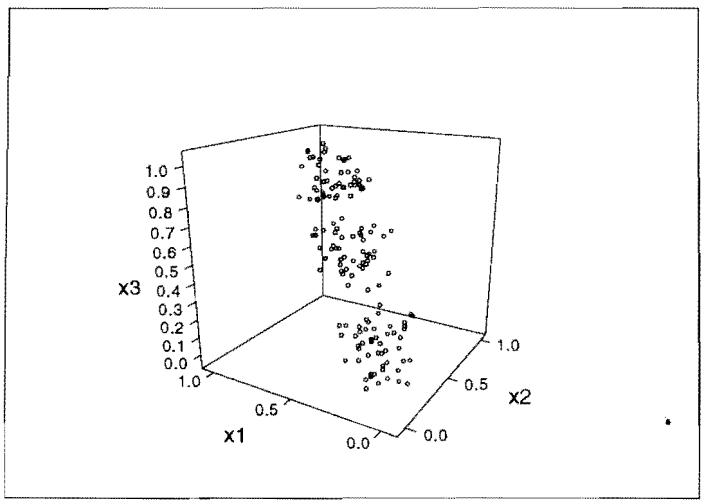


Figure 4: Sample to analyze

ulated. In figure 3 there is represented the target population. Points are described by their cartesian coordinates, namely X_1, X_2, X_3 (*DATASET 1*). Restriction to three variables is interesting since it makes possible understandable graphical representations.

Assuming a uniform distribution, 50 points of each subcube are simulated (see figure 4). In this example, the existence of three well-defined classes, each of 50 points, is previously known.

From the simulated sample, two more data sets were were generated by means of transforming one or more variables to a qualitative form. Table 5 shows those data sets. The advantage of working with simulated data is that real class of every object is known. Therefore the degree of missclassification can be exactly calculated.

Each data set has been analyzed using 3 statistical packages: **Klass** — which can perform clustering based on rules using mixed metrics—, **SPSS** [21] — a general purpose statistical package—, and **SPAD** [22] — which is oriented to multivariate analysis. In all the cases, an ascendant hierar-

Figure 5: Dendrogramme for first prove.

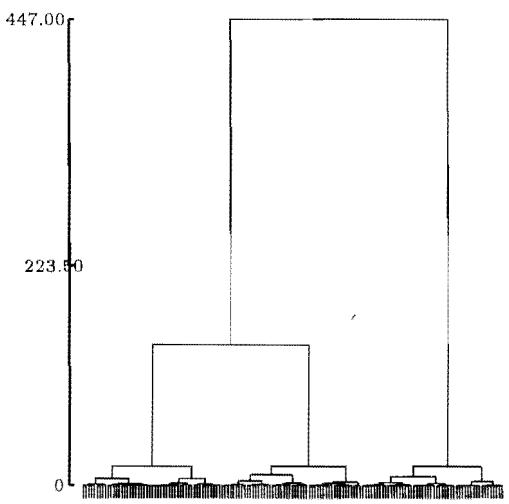
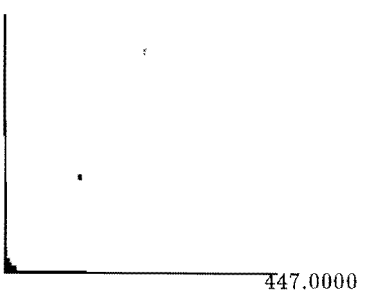


Figure 6: Inertias histogram.



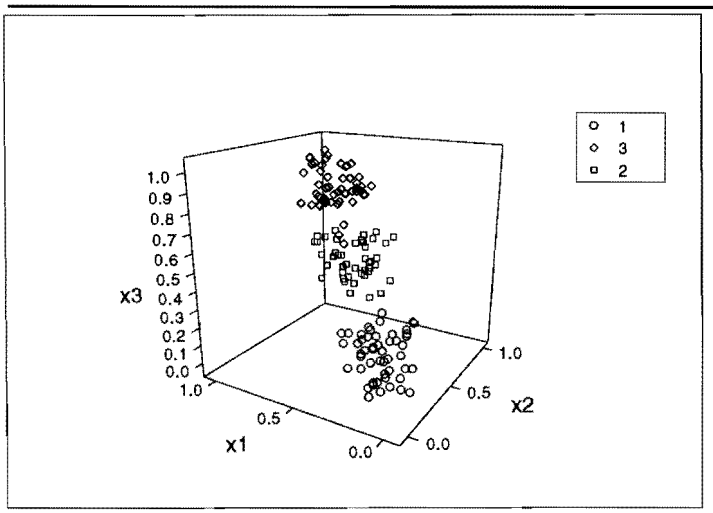


Figure 7: Classification in three groups.

chical clustering using the reciprocal neighbors algorithm and Ward's criterion is used. In **Klass** mixed distance is selected to perform the clustering. Results are summarized bellow.

For the *DATASET 1* The results provided by the three packages coincide, as expected, since the mixed metrics is equal to the normalized euclidean distance in this case. The dendrogramme is depicted in figure 5. The 3–classes structure is clear, especially when consulting the histogram of the level indexes (see figure 6). The three-classes cut is:

$$\begin{aligned} \mathcal{P} = \quad & \{\mathcal{C}_1 = \{i_1 \dots i_{50}\}, \\ & \mathcal{C}_2 = \{i_{51} \dots i_{100}\}, \\ & \mathcal{C}_3 = \{i_{101} \dots i_{150}\}\} \end{aligned} \tag{7}$$

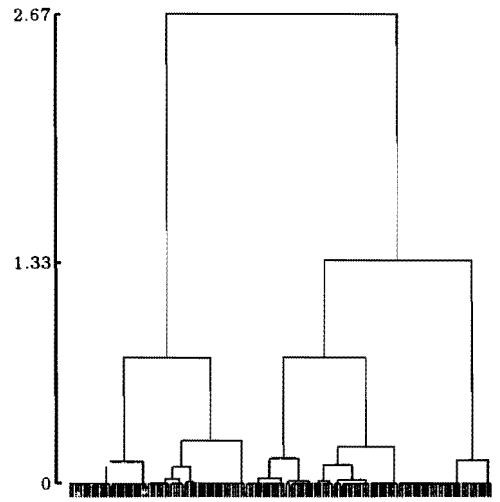
exactly corresponding to the three simulated cubes. Recognition of the classes was correct, as expected. Table 6 describe the existing classes. Figure 7 represents in a graphical way the partition identified by the three classifiers.

SPAD present the disadvantage that clustering is only allowed on the principal components. So, perform of PCA is first needed. For this particular application, the first axis can be interpreted as the main diagonal of the cube, which, by itself, gathers information enough about the location of the points.

For *DATASET 2*, the data matrix is qualitative (see codification in table 5): χ^2 metrics can be used. All the packages produce a five classes cut, directly induced by the values of X_1. However, neither **SPSS** nor **SPAD** can work directly with the qualitative data matrix. Data preprocessing is needed. First of all, transformation of modalities in numerical codes; then, a multiple correspondence analysis; finally, a clustering with euclidean metrics on the first five principal components can be done. Only **Klass** can directly deal on the data matrix.

DATASET 3 is an heterogeneous data matrix. Using the mixed metrics with the weighing values proposed in expression 6, partitions in 2, 4, 6 i 5 classes are suggested by **Klass**.

Figure 8: CAJ. Dendrogramme for third prove.



The second one determined, again, by the categorization of X_1 as in *DATASET 2*. In the 3 classes cut, X_1 is still a *characteristic variable* of the classes, in the sense of §2.3.1. The corresponding dendrogramme is shown in figure 8.

Again, neither **SPAD** nor **SPSS** can deal directly with the data matrix. Previous transformation is required. In this case, codification of all the continnuous variables[6]; next, split of all the variables in blocks of binary ones[7]; finally, classify the complete incidence matrix using χ^2 metrics[8].

Three and 5 classes cuts are identical in all classifications. After that, no more coincidences are detected. As an overall idea, performing a 6-classes cut produces a 3.3% of differences between **Klass** and the other packages; with a 7-classes cut, an 8% of differences; or even a 23% in a 7-classes cut.

That is, for this concrete case, the α–cuts of higher levels will be the same — this is because of the strong structure of data, but not for more classes. The use of the mixed distance offers a *different* possibility for classification over mixed matrices from those available in **SPSS** or **SPAD**, with the advantage of processing directly the mixed matrix, without previous transformations on the data. Often, results provided by mixed metrics allows a successfull interpretation from an expert point of view or even by means of autommatic tools.

[6]This implies the introduction of some arbitraryness in the process, since the result is then highly depending on the codes definition. And there are no tools to know *a priori* how to define them. From our opinion, this introduces unstability into the system, what is not desireable.

[7]Expanding the data matrix to the binary form transforms a data matrix of 150×3 cells to a bigger one (150×7). The increase is no critical for this particular application, but it use to be in real *ISD*.

[8]For technical reasons, this step is done as a multiple correspondence analysis of the complete incidence table, followed by clustering of the resulting principal components.

Class	X_1		X_2		X_3		Elements
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	
classe147	0:82903	0.10527	0.83743	0.10215	0.80769	0.09213	50
classe143	0.50071	0.10763	0.48819	0.09464	0.51179	0.10053	50
classe82	0.15738	0.08975	0.18455	0.09821	0.17996	0.10913	50

Table 6: Description of the classes.

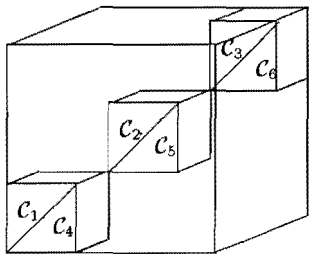


Figure 9: Graphical representation of the rules-induced partition.

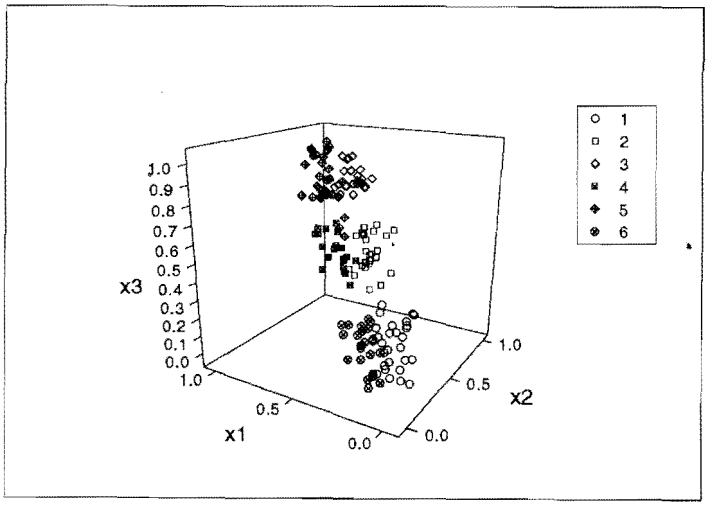


Figure 10: A 6-classes cut, using rules.

4.1 Introducing rules

Let us introduce semmantics into the system. For this last experiment *DATASET 1* is used. Suppose that X_1 is the *width* of a box, X_2 is its *height* and X_3 represents *length*. User could be interested (for storing purposes, for example) in separating *tall* boxes from *short* ones. The set of rules $\mathcal{R} = \{r_1, r_2\}$, where $r_1 : X_1 \leq X_2 \rightarrow Tall$ and $r_2 : X_1 > X_2 \rightarrow Short$, is introduced into the system. In fact, this implies to restrict clustering according to the plane $X_1 = X_2$ (see figure §10). In spite of the hard structure of the data set, the introduction of the rules is strongly biassing the clustering process and the resulting dendrogramme (figure 11) shows a different organization of the domain.

No three classes are detected anymore, but the two blocks

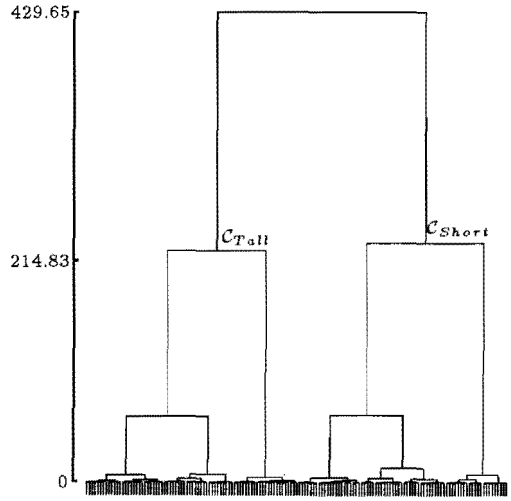


Figure 11: Classification of *DATASET 1* using rules.

$\mathcal{P}_4 \mid \mathcal{P}_1$	down-left	central	up-right
Short	23	25	28
Tall	27	25	22

Table 7: Comparing $\mathcal{P}_1$ with $\mathcal{P}_4$.

of *tall* boxes and *short* boxes, as well as the three *semi-cubes* are perfectly identified inside them. Other sets of rules could produce different results. Rules can be used to specify the clustering objectives, semmantical restrictions or prior kowledge relative to the structure of the domain. But strong structures are still recognized by the clustering process.

In the final step, the user can choose with the assistance of the system which is the better level to cut the tree. For this example, the first five suggested cuts are those of 2, 4, 6, 3, and 8 classes. For a 6-classes partition, comparison to initial partition of *DATASET* 1 can be performed. The distribution of the objects is shown in table 7 and about half of the objects are classified in different ways from both performances ($\delta = 0.63$). In fact, *tall* and *short* boxes are separated inside every subcube.

5 Conclusions and future work

In this paper, the methodology of *clustering based on rules* is presented. It successfully combines Artificial Intelligence

techniques with Statistical methods, for finding the structure of *ill-structured* domains (see §2).

Clustering based on rules uses the expert's knowledge to guide the clustering process. The use of this knowledge produces a great reduction in the ammount of computation required to classify the domain [8]; it also increases the quality of the results. In most of the cases, this process helps the expert to make his knowledge relative to some parts of the domain explicit.

Clear connections between *clustering based on rules* and *KDD* are shown along the paper: taking into account prior knowledge, applying a repeated Data Mining technique (in this case, clustering), including some tools interpretation-oriented to help the user to find the *meaning* of the classes... are some of the features that remain common between a *KDD* process and *clustering based on rules.*

Clustering based on rules is a methodology for the automatic clustering of objects described using both qualitative and quantitative variables. A family of metrics to measure distances between individuals that combine qualitative and quantitative variables, avoiding as much as possible the loose of information and making easier the implementation process, was first introduced in [13] and further developed in [11], although there are similar proposals as that of [28]. Among other characteristics of this metrics, we can mention:

- To simultaneously take advantage of the qualitative and quantitative information as well as the possibility to deal with the variables in their original form §3, avoiding intermediate transformation of the data matrix.
- It is no necessary to encode the categorical variables to obtain their numerical representation. The grouping of quantitative values — with the corresponding loss of information — to get an homogeneous data matrix of categorical variables may be suppressed.
- Considering that the quality of the results may depend on the way in which these groups are formed, elimination of this process is likely to produce more *objective* results on those classifications.
- It makes possible to use all those clustering methods that require a metric space, like Ward's method, with non homogeneous data matrices.

Different ranges of different kind of variables give a solid reason for proposing the mixed distance as a weighted distance. Different values of α and β may be used upon the user requirements. If the pair $\alpha = 1, \beta = 0$ is used, only numerical variables are considered to measure the distances. On the contrary, $\alpha = 0, \beta = 1$ represents the exclusive use of qualitative variables. Any pair α, β between these two cases represents an intermediate weighing of quantitative and qualitative information. The more α increases, the more influence quantitative variables in the final distance, and similarly occurs with β and qualitative variables.

The values proposed in formula (6) for the constants α, β are determined on the basis of some heuristic criteria shown in [11]. Apart from preserving the metrics structure,

- they represent a neutral situation where every variable is equally considered and
- they provide, in a number of cases, clear interpretable results.

Presenting a family of distances is a general situation that may include, as particular cases, the results provided by other methods. Indeed, the clusters obtained with other methods for a given data set may be obtained using mixed metrics, with appropriate values for α and β.

For example, for the application presented in §3, using $\alpha = 0.05$, $\beta = 0.95$ and an α-cut at level 2.5, the clusters provided by the single linkage method, and presented in table (2), are obtained. In this case, the values $\alpha_0 = 0.014$ and $\beta_0 = 0.986$, suggested by expression 6, give even more importance to qualitative variables according to the fact that they represent the 40% of the available information.

Regarding the §3 it can be seen that the characterization system proposed in §2.3.1 can help to find the more relevant features of every class, contributing to an easy interpretation of them.

In §2.2.2, mixed distance with the α and β values proposed in 6 generated higher quality results that other methods. In fact, for this particular application, this metrics produce an automatic characterization close to the one made by the expert.

From §4 the first observation is that **SPAD** deals only with principal components and the coincidence degree with the other packages depends on how well are those components representing the whole set of variables.

For categorical or mixed data matrices, multiple correspondence analysis and other preprocessing methods are required either in **SPAD** or **SPSS** , while **Klass** can process the original data matrix.

In the analyses of mixed data matrices, mixed distance offers a new possibility from those available in **SPSS** or **SPAD**, with the advantage of processing directly the mixed matrix, without previous transformations on the data. Often, results provided by mixed metrics allows a successfull interpretation from an expert point of view or even by means of autommatic tools.

On the other hand, it is clear that the introduction of rules into the clustering process allows the management of semmatics, what is impossible from other clustering systems. In general, this is also improving the quality of the results, especially in terms of interpretability (see §4). Anyway, strong structures on the data, overcome any expert constraint expressed in the rules base, and remain evident in the final classification. This is the case of the cubes application, where the rules imposed by the expert cannot hide the general structure of the data in three subcubes. This ows to the fact that a case bas-

ed technique (clustering) is combined with a knowledge based one and the results are the combination of the information provided by the data (contained in the data matrix) and that provided by the expert (contained in the rules).

From a validation point of view, apart from some tools provided by **Klass**, like the similarity between a classification and the expert proposal, the *interpretability* of the results has been used as a criterion on the classifications quality, since, at present, assessing the clustering results is not a very well solved question [4]. For the specific application presented in §3, clusters provided by **Klass** using the α_0 and β_0 values suggested in formula (6) with mixed metrics fit rather well the classification proposed by the expert.

We can conclude that experts use to be able to interpret the results obtained with the heuristic presented here, as it has been observed from other applications in different fields (as sea sponges [12]).

Anyway, comparisons among classifications are still informal, and it will be interesting to have more objective criteria to validate them. Distances between "*expert classifications*" and "*automatic classifications*" would be a numerical way to do that. This research is actually in progress [9], and it hopefully will provide a tool to accept or reject a classification according to expert's criteria.

On the other hand, it will be also interesting to introduce the concept of *relevance* of a variable into the system [1]. As a first approach, giving weights to the variables may be considered, although there may be some other possibilities.

References

[1] Belanche, L., Cortés, U., The nought attributes in KBS. In *EUROVAV-91*, 77–103, 1991.

[2] Benzecri, J.P., *L'analyse des données.* Dunod, Paris, 1980.

[3] de Rham, C., La classif. hierarch. selon la méthode des voisins réciproques. *Cahiers d'Analyse des Données* **V**, n.2: 135–144.

[4] Dillon W.R., Goldstein M., *Multivariate analysis. Methods & applications.* Wiley, USA, 1984.

[5] Everitt, B., *Cluster analysis.* Heinemann Educational Books Ltd, London, 1974.

[6] Fayyad, U., Piaetsky-Shapiro, G. Smyth, P. From Data Mining to Knowledge Discovery: An overview. In Fayyad, U., Piaetsky-Shapiro, G. Smyth, P, Uthurusamy, R. (Eds.) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT, 1996.

[7] Frawley, W., Piaetsky-Shapiro, G., Mathews, C., KDD: An overview. *AI Magazine* **14**(3):57–70, 1992.

[8] Gibert, K. On the Uses and Costs of Rule-Based Classification In proc. *COMPSTAT'96*, pp 265–270, Barcelona, Springer-Verlag, 1996.

[9] — L'ús de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats. Departamento EIO, UPC, Barcelona, 1994 (Ph. D. thesis).

[10] Gibert, K. KLASS: Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Universidad Politécnica de Catalunya, 1991 (master thesis).

[11] Gibert, K., Cortés, U. Weighing quantitative and qualitative variables in clustering methods, *MATHWARE* 4(3): 251–266, May 1997.

[12] — Combining a knowledge based system with a clustering method for an inductive construction of models. In P. Cheeseman *et al.* (Eds.), *Selecting Models from Data: AI and Statistics IV*, LNS n° 89, pp 351–360, New York, Springer-Verlag, 1994.

[13] — **KLASS:** Una herramienta estadística para la creación de prototipos en dominios poco estructurados. In proc. *IBERAMIA-92*, pp 483–497, México, Noriega Eds., 1992.

[14] Gibert, K., Hernández-Pajares, M., Cortés, U. Classification based on rules: an application to Astronomy. In *IFCS'96*, pp 69–72, Kobe, Japan, 1996.

[15] Gowda, K. C., Diday, E. Symbolic clustering using a new similarity measure, *IEEE Trans. on systems, man, and cibernetics*, **22**(2): 368–378, 1992.

[16] Gower,J. C., A general coefficient for similarity, *Biometrics*, (27): 857–872.

[17] Krzanowski, W. J. Distance between populations using mixed continuous and categorical variables. *Biometrika*, **70**: 235–43, 1983.

[18] Lebart, L., *et al.*, *Traitement des données statistiques.* DUNOD, Paris, 1982. (2^d ed.)

[19] Lerman, I. C., Peter, Ph., " Organization et consultation d'une banque de "petites annonces" a partir d'une méthode de classification hierarchique en parallele". In *Data analysis and informatics, IV*, pp 121–135, North Holland, Elsevier Science Publishers BV, 1986.

[20] Martín, M. *et al.*, Knowledge acquisition combining analytical and empirical techniques, *Machine Learning*: 657–661, 1991.

[21] Norusis, M.j., SPSS base system user's guide. Chicago, *SPSS* Inc., 1990.

[22] Lebart, L., Morineau, A., Lambert, *Système portable pour l'analyse des donées.* Sèvres: *CISIA*, 1987.

[23] Michalski, R. S., Stepp, R. E., Automated construction of classifications: Conceptual clustering versus numerical taxonomy, *IEEE Trans. on PAMI* (5): 396–410, 1983.

[24] Michalski, R. S. *et al.*, PLANT/ds: An expert consulting system for the diagnosis of soybean diseases. In *Eur. Conf. of AI.* Orsay, France, 1982.

[25] Nakhaeizadeh, G. Classification as a subtask of of Data Mining experiences form some industrial projects. In *IFCS'96*, pp 17–20. Kobe, Japan.

[26] Quinlan, J. R., Learning efficient classification procedures and their application to chess and games. In Michalski, R.S. *et al.* (Eds.), *ML: An AI Approach.*, pp 463–482. Tioga, PA, 1984.

[27] — , Learning efficient classification procedures, *ML. an A.I. perspective*, Tioga, PA, 1983.

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

[29] Roux, M., *Algorithmes de classification*, Masson, Paris, 1985.

[30] L. Godo, R. López de M´ntaras, C. Sierra and A. Verdaguer, **MILORD II**: The Architecture and the Management of Linguistically Expressed Uncertainty Chapter 23. In D. Dubois, H. Prade, and R. R. Yager (Eds.), *Fuzzy information engineering: a guided tour of applications* , John Wiley & Sons, New York 1997.

[31] Shortlife, E. H., **MYCIN**: A rule-based computer program for advising physicians regarding antimicrobial therapy selection, Standford, Univ. USA, 1976(Ph. D. thesis).

[32] Sonicki, Z. *et al.* The use of induction in routine laboratory diagnostics of thyroid, *LIJECNICKI VJESNIK* **115**: 306–309, 1993 (in croatian).

[33] Volle, M., *Analyse des données*, Economica, Paris, 1985.

[34] Wolfe, J. H., Pattern clustering by multivariate mixture analysis, *MBR* (5): 329-350, 1971.

Karina Gibert Oliveras *received a Ph.D. degree in Informatics at the Universitat Politècnica de Catalunya (UPC), Spain. Her main areas of interest are: data mining, knowledge discovery, automatic classification, knowledge bases, and mixed data metrics. Currently she is a professor at the Operations Research and Statistics Department at the Informatics, and Mathematics and Statistics Faculties,UPC. She is also an advisor of the technical Engineering Informatic studies at the Universitat Oberta de Catalunya.*

Ulises Cortés *recived a B.Sc. in Industrial and Systems Engineering from the Instituto Tecnológico de Estudios Superiores de Monterrey (ITESM) in 1982, and a Ph. D. from the Universitat Politécnica de Catalunya (UPC) in 1984. He is an Associate Professor in the Software Department of the UPC since 1988. Head of the Artificial Intelligence Ph. D. program and Vice-dean of the International Affairs of the Faculty of Computer Science of Barcelona (FIB). He is member of the AEPIA (Spanish Association for Artificial Intelligence). He is pioneer member of ACIA (Catalan Association of Artificial Intelligence). He is member of SMIA (Mexican Society of Artificial Intelligence). He has an extensive list of publications, conference papers, and tutorial and supervisory contributions. His main research topics are machine learning, knowledge acquisition an LISP-like languages.*