*Mathware & Soft Computing 4 (1997) 251-266*

# Weighting Quantitative and Qualitative Variables in Clustering Methods

Karina Gibert[1] and Ulises Cortés[2]
[1] Dept. Statistics and Op. Research. *e-mail: karina@eio.upc.es.*
[2] Dept. Software. *e-mail: ia@lsi.upc.es*
Universitat Politècnica de Catalunya
Pau Gargallo, 5. Barcelona. 08028. Spain.

**Abstract**

Description of individuals in *ill-structured* domains produces messy data matrices.

In this context, automated classification requires the management of those kind of matrices.

One of the features involved in clustering is the evaluation of distances between individuals. Then, a special function to calculate distances between individuals partially simultaneously described by qualitative and quantitative variables is required.

In this paper properties and details of the metrics used by **Klass** in this situation is presented — **Klass** is a clustering system oriented to the classification of ill-structured domains which implements an adapted version of the reciprocal neighbors algorithm; it also takes advantage of any additional information that an expert can provide about the target concepts.

**Keywords:** clustering, metrics, qualitative and quantitative variables, messy data, ill-structured domains

## 1 Introduction

Classification is the more used technique to separate data into groups. Classification methods are interesting from an Artificial Intelligence point of view, because they open a door to the automated generation of classification rules, extremely useful in knowledge-based environments, in particular the diagnostic oriented ones. Indeed, several well known expert systems, as **MYCIN** [SHOR76], **MILORD** [SIER89] or others, are actually classifiers.

However, in AI, it is usual to work with *ill-structured domains* (in [GIBE94] a complete characterization of them may be found) as mental disorders, sea sponges, books classification, fossils... In this kind of domains, the consensus among experts is weak — and sometimes non-existent. When describing them, the use of qualitative variables become very common. Experts seem to feel better when using

qualitative terms, even for numerical concepts[1]. That is why in most cases quantitative and qualitative information coexists in what we call *non-homogeneous* data bases. Even more, the number of modalities of qualitative variables depends on the expertise of the person who is describing the objects: the more he knows about the domain, the greater is the number of modalities he uses.

Management of non-homogeneous data matrices requires, indeed, special attention when classifying ill-structured domains. Standard clustering methods were originally conceived to deal with quantitative variables. When qualitative variables appear, previous treatments on the data matrix are needed. Here are different proposals on this line:

- Splitting any qualitative variable to generate the *complete incidence table*. Afterwards, a classification using $\chi^2$ metrics may be performed [DILL84].
- The application of an analysis technique that deals with qualitative information only. Often correspondence analysis is used. Then, a classification on the factorial components is possible [VOLL85], [LEBA85].
- The grouping of quantitative values transforming the corresponding variables into qualitative [ROUX85].

For the first one, there is a significant increase of the cost of the process, due to the dimensions of the complete incidence table. In the second one, since the classification is performed in a ficticious space, additional tools have to be provided to enable the interpretation of the results. For the last one, this transformation implies a loss of information as well as the introduction of some inestability in the results: they depend on the groups made with the quantitative values, what may be rather arbitrary.

In this paper, another alternative is presented: the idea is to allow clustering on a domain simultaneously described by qualitative and quantitative variables without transforming the variables themselves.

In the core of the classification process distances between individuals have to be calculated. Then, a function to do it with non homogeneous data has to be found. In fact, there are some proposals on this line, like [GOWE71] or [GOWD92] presenting similarity coeficients to evaluate *proximitites* between individuals. In this work, a new metrics that can measure distances with messy data is introduced. This measure has been succesfully implemented in a clustering system called **Klass** [GIBE94], [GIBE96], [GIBE94b], and applied to very different domains. In the paper some results are also shown.

This paper is organized in four sections, besides the introduction. First of all, details on the measurement of distances between individuals are given in section 2, presenting a family of functions that combines qualitative and quantitative information. In section 3 the metrics structure of this family is proved while in section 4 a proposal is made on the values of the parameters of the metrics family. Section 5 shows an application, and the last section presents some conclusions and future work.

[1] For example, although the **hair length** is clearly quantitative, no one deals with it in this way, but using some categories as *short hair*, *long hair* and so on.

# 2 Measuring distances

During data collection, $K$ variables $X_1 \dots X_K$ are observed over a sample of $n$ individuals. The value taken by individual $i, (i = 1 : n)$ for variable $k, (k = 1 : K)$ is denoted by $x_{ik}$. Therefore, an individual $i$ is described by a vector of observations $(x_{i1}, x_{i2}, \dots, x_{iK})$, and an $(n, K)$ matrix is built with the values $x_{ik}$. The rows of the data matrix contain information relative to the individuals to be classified, while each column concerns one of the variables used to describe the sample.

Next subsections provide suitable metrics for classifying different kind of data matrices, and finally a proposal for non-homogeneous ones.

## 2.1 Only quantitative variables

Given a data matrix as the described above, the canonical euclidean metrics has been traditionally used to measure distances between individuals when all the variables are quantitative.

$$d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \tag{1}$$

Using expression (1) to compute distances is not scale invariant. Normalized distances are. Some ways to normalize expression (1): dividing data through the standard deviation or through the range of each variable:

$$d^2(i, i') = \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{s_k} \; ; \quad s_k^2 = \sum_{i=1}^{n} \frac{(x_{ik} - \bar{x})^2}{n-1} \tag{2}$$

$$d^2(i, i') = \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{r_k} \; ; \quad r_k = \max_i\{x_{ik}\} - \min_i\{x_{ik}\}$$

## 2.2 Only qualitative variables

In standard clustering systems, when all the variables are qualitative, the data matrix is usually transformed into a *complete incidence* table [VOLL85] by splitting each qualitative variable $X_k, (k = 1 \dots K)$ which has $n_k$ possible values $\mathcal{D}_k = \{c_1^k, \dots, c_{n_k}^k\}$ into a set of binary variables $\{Z_1^k, \dots, Z_{n_k}^k\}$. Each $Z_j^k$ $(j = 1 : K)$ corresponds to a $c_j^k \in \mathcal{D}_k$, and the complete incidence table is composed of elements

$$z_{ij}^k = \begin{cases} 1, & x_{ik} = c_j^k \\ 0, & \text{otherwise} \end{cases} \; , (i = 1 : n), (k = 1 : K), (j = 1 : n_k)$$

For instance,

$$\begin{array}{c} \\ i_1 \\ \vdots \\ i_n \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \dots & X_K \end{array} \\ \left( \begin{array}{ccc} x_{11} & \dots & x_{1K} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nK} \end{array} \right) \end{array} \quad \text{with} \begin{cases} x_{11} = c_2^1 \\ x_{1K} = c_2^K \\ \vdots \\ x_{n1} = c_1^1 \\ x_{nK} = c_{n_K}^n \end{cases}$$

would be transformed into

$$\begin{array}{c} \\ i_1 \\ \vdots \\ i_n \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1^1 & Z_2^1 & \dots & Z_{n_1}^1 & \dots & Z_1^K & \dots & & Z_{n_K}^K \end{array} \\ \left( \begin{array}{ccccccccc} z_{11}^1 = 0 & 1 & \dots & 0 & & 0 & 1 & \dots & 0 \\ & & & \vdots & & & & & \\ 1 & \dots & \dots & 0 & & 0 & \dots & & z_{nn_k}^k = 1 \end{array} \right) \end{array}$$

When data is presented in such a way, the $\chi^2$ metrics can be used to calculate distances between objects [BENZ80]:

$$d^2(i,i') = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \frac{\left( \frac{z_{ij}^k}{z_{i.}} - \frac{z_{i'j}^k}{z_{i'.}} \right)^2}{z_{.j}^k} \text{ , where } \quad z_{i.} = \sum_{k=1}^{K} \sum_{j=1}^{n_k} z_{ij}^k, \ (i = 1:n) \qquad (3)$$

, and $z_{.j}^k = \sum_{i=1}^{n} z_{ij}^k = card(i : x_{ik} = c_j^k), \quad (k = 1:K), (j = 1:n_k) \quad (z_{.j}^k$ is the number of individuals of the sample that are in the modality $c_j^k$ of $X_k$).

As said before, qualitative variables are very frequent in ill-structured domains and the number of categories of the corresponding qualitative variables increases, according to the expertise of the user.

Splitting qualitative variables with a number of categories leads to *big* and *sparse* binary matrices. This significantly increases the cost of the classification process. That is one of the reasons why it is interesting to process directly the original data matrix, where the values of qualitative variables are represented by means of *symbols*. Hence, we propose a rewriting of expression (3) so that $\chi^2$ distances can be calculated using this raw representation.

Since each individual belongs to only one category of a given categorical variable $X_k$, the splitting of $X_k$ leads to a binary subvector containing exactly one element equal to 1.

$$\forall i,k \ (\exists j_0 = \{1 : n_k\} \ : z_{ij_0}^k = 1 \ \& \ \forall j \neq j_0, \ z_{ij}^k = 0)$$

and therefore $z_{i.} = K, \quad \forall i = (1:n)$.

Using this property, expression (3) may then be rewritten as

$$d^2(i,i') = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \frac{(z_{ij}^k - z_{i'j}^k)^2}{z_{.j}^k} = \frac{1}{K^2} \sum_{k=1}^{K} d_k^2(i,i') \qquad (4)$$

where $d_k^2(i,i')$ considers together the set of columns $Z_1^k \dots Z_{n_k}^k$ which come from splitting the qualitative variable $X_k$ into binary variables. Thus, $d_k^2(i,i')$ is called **contribution of the $k^{th}$ variable** to the total distance. Defining $I_{k^i} = card(\hat{\imath} : x_{ik} = x_{ik})$ as the number of individuals of the sample that chose for variable $k$ the same modality as individual $i$, it holds that

$$d_k^2(i,i') = \begin{cases} 0, & \text{if } x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}}, & \text{otherwise} \end{cases} \tag{5}$$

See [GIBE94] for a detailed developement of these expressions.

Anyway, during the clustering process, the subclasses generated by the classifier are prototypically represented by their gravity center. Representation of qualitative components of the class gravity center also requires some work.

For quantitative variables, the gravity center of a class $\mathcal{C} = \{i_1, \dots, i_{n_\mathcal{C}}\}$ is calculated as the arithmetic mean. If variable $X_k$, in class $\mathcal{C}$, takes values $x_t \in \Re$:

$$\bar{x}_{\mathcal{C}k} = \sum_{i \in \mathcal{C}} = \frac{x_{ik}}{n_\mathcal{C}} \text{ which may also be written as } \bar{x}_{\mathcal{C}k} = \sum_{t=1}^{T} f_t x_t$$

where $f_t$ is the proportion of objects $i \in \mathcal{C}$ such that $x_{ik} = x_t$.

Considering that the sum is non-sense for qualitative variables, a vectorial representation of the gravity center may be adopted in this case.

$$\bar{x}_{\mathcal{C}k} = \left( (f_\mathcal{C}^{k_1}\ c_1^k) \dots (f_\mathcal{C}^{k_{n_k}}\ c_{n_k}^k) \right), \quad f_\mathcal{C}^{k_j} = \frac{card\{i \in \mathcal{C} :\ x_{ik} = c_j^k\}}{n_\mathcal{C}} \tag{6}$$

For vectorial components, ordinary $\chi^2$ distance is used and, finally, the contribution of variable $k$ to the total distance is calculated as follows:

$$d_k^2(i,i') = \begin{cases} 0, & \text{if } x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}}, & \text{otherwise, for individuals} \\ \frac{(f_i^{k_s} - 1)^2}{I^{k_s}} + \sum_{j \neq s}^{n_k} \frac{(f_i^{k_j})^2}{I^{k_j}}, & \text{if } x_{ik} = c_s^k, \text{ and } i' \text{ is a subclass} \\ \sum_{j=1}^{n_k} \frac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}}, & \text{in the general case} \end{cases} \tag{7}$$

In formula (7), $I^{k_j}$ is the number of individuals of the sample that are in modality $c_j^k$; $I_i^{k_j}$ the same, concerning the $i^{th}$ class; $I_{k^i}$ defined previously (equation 5); and $f_i^{k_j}$ represents the proportion of individuals from the $i^{th}$ subclass satisfying

that $X_k = c_j^k$. In fact,

$$f_i^{k_j} = \frac{I_i^{k_j}}{\sum_{j=1}^{n_k} I_i^{k_j}}. \tag{8}$$

It can be observed that expression (8) may be directly evaluated on the symbolic representation of the data matrix. It is no more necessary to explicitly built the complete incidence table corresponding to the original data matrix.

## 2.3 Combining qualitative and quantitative variables

When both quantitative and qualitative variables appear in the data matrix, a new distance has to be defined.

Let $\mathcal{V} = (1 : K)$ be the set of indexes of the variables $X_k$. Expressions (1) and (7) are particular cases of

$$d^2(i, i') = \sum_{\forall k \in \mathcal{V}} \delta_k^2(i, i') \tag{9}$$

where $\delta_k^2(i, i')$ measures the distance between $i$ and $i'$ related to variable $X_k$.

In the general case, $\mathcal{V}$ may be partitioned as

$$\mathcal{Q} = \{k \in \mathcal{V} : X_k \textit{ is qualitative}\} \quad \text{and} \quad \mathcal{C} = \{k \in \mathcal{V} : X_k \textit{ is quantitative}\}$$

such that $\mathcal{V} = \mathcal{Q} \cup \mathcal{C}$, and expression (9) may be generalized to the case where $X_k$ can be either quantitative or qualitative. Then, $\delta_k^2$ will be calculated according to the type of the variable $X_k$:

$$\delta_k^2(i, i') = \begin{cases} \frac{(x_{ik} - x_{i'k})^2}{s_k^2}, & k \in \mathcal{C} \\ \\ \frac{d_k^2(i,i')}{n_{\mathcal{Q}}^2}, & k \in \mathcal{Q}^{(2)} \end{cases}$$

In this case, expression (9) may be rewritten as

$$d^2(i, i') = \sum_{\forall k \in \mathcal{C}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{1}{n_{\mathcal{Q}}^2} \sum_{\forall k \in \mathcal{Q}} d_k^2(i, i') \tag{10}$$

In order to balance the influence given by each group of variables, it seems reasonable to give different weights, namely $\alpha$ and $\beta$ (with $\alpha \geq 0, \beta \geq 0$), to these two components of the distance. So expression (10) is modified to

$$d_{(\alpha,\beta)}^2(i, i') = \alpha \sum_{\forall k \in \mathcal{C}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \beta \frac{1}{n_{\mathcal{Q}}^2} \sum_{\forall k \in \mathcal{Q}} d_k^2(i, i') \tag{11}$$

and it is called *mixed function*.

[2] From now on $n_{\mathcal{Q}} = card\mathcal{Q}$.

The mixed function may be seen as the combination of two major components:

$$d^2_{\mathcal{C}}(i,i') = \sum_{\forall k \in \mathcal{C}} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} \quad ; \quad d^2_{\mathcal{Q}}(i,i') = \frac{1}{n^2_{\mathcal{Q}}} \sum_{\forall k \in \mathcal{Q}} d^2_k(i,i')$$

$d^2_{\mathcal{C}}(i,i')$ will be called *quantitative subdistance*, whereas $d^2_{\mathcal{Q}}(i,i')$ can be called *qualitative subdistance*.

Expression (11) may then be written as:

$$d^2_{(\alpha,\beta)}(i,i') = \alpha d^2_{\mathcal{C}}(i,i') + \beta d^2_{\mathcal{Q}}(i,i')$$

In fact, the *mixed function* is a family of functions indexed by the pair $(\alpha, \beta)$:

$$\{d^2_{(\alpha,\beta)}(i,i')\}_{(\alpha,\beta)\ \in(\Re^* \times \Re^* - \{(0,0)\})} \ , \quad \text{, with}^3 \ \ \Re^* = \Re^+ \cup 0.$$

With this formulation, distance expressed in (4) corresponds to the element $d^2_{(0,1)}(i,i')$ of *mixed family*, while in (2) is $d^2_{(1,0)}(i,i')$.

# 3 The mixed metrics

## 3.1 Structure of mixed function

It is shown in [GIBE94] that *mixed function* satisfies the properties of a distance under the following condition:

$$\alpha = 0 \Longrightarrow \mathcal{C} = \emptyset \quad \& \quad \beta = 0 \Longrightarrow \mathcal{Q} = \emptyset \tag{12}$$

Considering that $d_{\mathcal{C}}(i,i')$ and $d_{\mathcal{Q}}(i,i')$ are both metrics over the spaces $\{X_k : k \in \mathcal{C}\}$ and $\{X_k : k \in \mathcal{Q}\}$ respectively, the following properties hold:

1. Simetry: $d_{(\alpha,\beta)}(i,i') = d_{(\beta,\alpha)}(i,i'),\ \forall i,i'$
2. Triangular inequality: $d_{(\alpha,\beta)}(i',i'') \leq d_{(\alpha,\beta)}(i',i) + d_{(\alpha,\beta)}(i,i''), \forall i$
3. Identity: $d_{(\alpha,\beta)}(i,i') = 0 \Longleftrightarrow i = i', \forall i,i'$

Demonstrations related to properties 1 and 2 are simple on the basis of the metrics properties of each $d^2_k(i,i')$. However, condition (12) needs to be imposed to guarantee the property 3. Otherwise, different objects may work out at null distances. In next paragraph, the violation of property 3 when contition (12) is not satisfied is shown.

Let us suppose a set of variables $X_1 \ldots X_k$ where $\mathcal{C} \neq \emptyset$ and $\mathcal{Q} \neq \emptyset$, and consider the element $d^2_{(0,\beta)}$, $\beta > 0$. It holds that $d^2_{(0,\beta)}(i,i') = 0$ for any pair of objects $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $i' = (x_{i'1}, x_{i'2}, \ldots, x_{i'n})$, such that $x_{ik} = x_{i'k}, \forall k \in \mathcal{Q}$, independently of the quantitative components, even differents. Expression (12) is necessary and suficient for the metrics structure of the mixed function.

Therefore, the **family of mixed metrics** is defined as the set of elements of the mixed function satisfying condition (12).

[3] The domain of $(\alpha, \beta)$ excludes the element (0,0) because $d^2_{(0,0)}(i,i')$ is the constant function 0 that is nonsense in this context.

### 3.2 Equivalence on mixed distances

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, when 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:

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

It can be shown that $\equiv$ satisfies the properties of an equivalence relationship when

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

Given this equivalence relationship it is possible to work with its quotient set, choosing the distance

$$d^2_{(\alpha_0,\beta_0)}(i,i'), \text{ such that } \alpha_0,\beta_0 \in [0,1] \text{ and } \alpha_0+\beta_0 = 1,$$

as a representative element for each equivalence class and identifying the class by its representative.

The representative of the equivalence class of a given distance $d^2_{(\alpha,\beta)}(i,i')$ is $d^2_{(\alpha_0,\beta_0)}(i,i')$:

$$\alpha_0 = \frac{\alpha}{\alpha+\beta} \qquad \& \qquad \beta_0 = \frac{\beta}{\alpha+\beta} = 1-\alpha_0 \tag{13}$$

## 4 On the values for $\alpha$ and $\beta$

In this section some heuristic criteria are introduced to find acceptable values for the weighting constants $\alpha$ and $\beta$. As said before, the mixed function may be seen as the combination of $d^2_{\mathcal{C}}$ and $d^2_{\mathcal{Q}}$. The first point is to adjust these components. The range of quantitative subdistances depends on the measuring magnitude of the corresponding variable. The range of qualitative ones also depends on the number of categories of the qualitative variables.

The equilibrium is found when the two subdistances are referred to a common interval, for example [0,1]. In order to do that, a first approach is to divide each subdistance by their maximum, previously eliminating a 5% of extreme values, in order to acquire more robustness. These truncated maximums are respectively denoted by $d^2_{\mathcal{C}max^*}$ and $d^2_{\mathcal{Q}max^*}$. Hence,

$$\alpha \propto \frac{1}{d^2_{\mathcal{C}max^*}} \qquad \& \qquad \beta \propto \frac{1}{d^2_{\mathcal{Q}max^*}} \tag{14}$$

This operation guarantees that the two components will have equal influence in the determination of $d^2(i,i')$.

Doing so, if there is an outlier with big distances respect to the other objects it will not be taken as reference point, otherwise the other distances would be very little and concentrated in a subinterval [0,$c_0$], $c_0 \ll 1$.

Moreover, if there is not an outlier, the eliminated distances will be almost of the same range as $d^2_{\mathcal{C}max^*}$, and $d^2_{\mathcal{Q}max^*}$ respectively, and the real working interval will be [0,$c_0$], $c_0 \approx 1$, what does not imply a major change.

After finding this common reference, it seems reasonable to give more importance to the qualitative component if the objects are mainly described by qualitative variables, and *vice versa*. Therefore, the weighting constants are defined proportionally. So, being $n_{\mathcal{C}} = card\{\mathcal{C}\}$ and $n_{\mathcal{Q}} = card\{\mathcal{Q}\}$,

$$\alpha \propto n_{\mathcal{C}} \qquad \& \qquad \beta \propto n_{\mathcal{Q}} \tag{15}$$

Condition (15) also guarantees that the resulting distance will be an element of the *mixed metrics family*. Combining this condition with (14), the following values may be proposed for $\alpha$ and $\beta$:

$$\alpha = \frac{n_{\mathcal{C}}}{d^2_{\mathcal{C}max^*}} \qquad \& \qquad \beta = \frac{n_{\mathcal{Q}}}{d^2_{\mathcal{Q}max^*}} \tag{16}$$

and considering the equivalence relationship defined over the mixed metrics family, the representative of the equivalence class of $(\alpha, \beta)$ is taken. So, the final proposal is:

$$\alpha_0 = \frac{\alpha}{\alpha + \beta} \qquad \& \qquad \beta_0 = \frac{\beta}{\alpha + \beta} \tag{17}$$

# 5 An application

Among other applications [GIBE92], [GIBE94], where data usually requires a lot of background knowledge on the domain to interpret the results (sea sponges, stellar populations,...) the one concerned with a set of data presented in [MICH83], [GOWD92] has been selected to compare the performance of clustering with mixed metrics against other methods on the basis of a common and well studied training set.

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, [MICH83] gave the classification showed in table (2). This training set was also treated in [GOWD92] using 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 table 2. All these classifications contain exactly 4 classes. Our local expert has also been 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 shortly important to characterize microcomputers. However, relevance of the variables is a biasing rule which, at present, is not taken into account by our system. Next, he proceed to determine how the values of each categorical variable could be grouped. In fact, he was looking for the

| 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.

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 *Display, ROM* and *RAM* respectively, and he accepted as meaningful any combination of this initial classifications. None of them had four classes, except the one regarding *Display*, which is shown in 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 provided by **Klass** in this case using the mixed distance $d_{(\alpha,\beta)}$, with $\alpha = 0.014$ and $\beta = 0.986$ as suggested by formula (17) is presented in figure (1). A classification with four clusters has been chosen in order to make easier comparison against the other methods considered. Extensional and prototypical representations of the classes are described in table (4).

In order to evaluate the performance of this metrics, the expert was asked to interpret the results from the different methods. From his opinion **Klass** results were based on clear classification criteria: the *Display* combined, with less influence, by *Microprocessor*. This is obvious 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 meaningful clustering criteria.

Trying to evaluate in an *objective* way the *proximity* between pairs of classifications, the number of individuals equally classified in a given pair of partitions was used. Again, the results produced by **Klass** are the more similar to the expert proposal — with a 15 %, what represents two objects classified in different classes —, followed by those presented in [MICH83] — with a 46% of differences. Table (2–right) shows how **Klass**, followed by Michalski's results, are the more similar to the expert classification.

| Id. | MIC83 Conc. clust. | GOW92 Rec. Neigh. | GOW92 Sing. Link. | **Klass** | Expert |
|---|---|---|---|---|---|
| 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 |

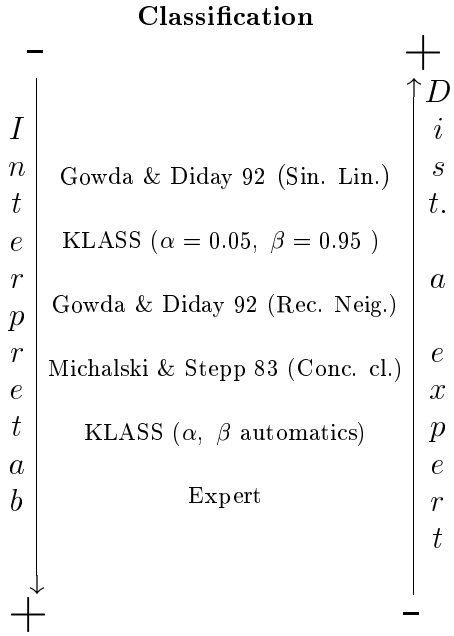


Table 2: Different classifications of microcomputers provided by different algorithms and distances.

# 6 Conclusions and future work

In this paper, a family of metrics to measure distances between individuals that combine qualitative and quantitative variables is presented (see §3). The use of mixed distance implies

- To take advantage of the qualitative and quantitative information simultaneously and the possibility to deal with the variables directly in the way they arrive to the system, 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 also be suppressed.
- Considering that the quality of the results may depend on the way in which these groups are performed, elimination of this process is likely to produce more *objective* results on the classifications.
- It is remarcable that the necessary and suficient condition for preserving the metrics structure of the mixed metrics has been found. It is then possible to use all those clustering methods that require a metric space, like Ward's method, with non homogeneous data matrices.

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

Table 3: Structure of categorical variables found by an expert.

Different ranges of different kind of variables give a solid reason for proposing the mixed distance as a weighted distance. Different values of $\alpha$ and $\beta$ may be used upon the user requirements. If the pair $\alpha = 1, \beta = 0$ is used, only numerical variables are considered. On the contrary, $\alpha = 0, \beta = 1$ represents the exclusive use of qualitative variables. Any pair $\alpha$, $\beta$ between these two cases represents an intermediate weighing of quantitative and qualitative information. The more $\alpha$ increases, the more influence quantitative variables in the final distance, and *vice versa*. The values proposed in formula (17) are determined on the basis of some heuristic criteria shown in §4. They represent a neutral situation where every variable is equally considered and preserve the metrics structure.

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 may be obtained by **Klass** using appropriate values for $\alpha$ and $\beta$.

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

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 validation criterion on the classifications quality, since, at present, assessing the clustering results is not a very well solved question [DILL84]. For the specific application presented here, clusters provided by **KLASS** using the $\alpha_0$ and $\beta_0$ values suggested in formula (17) 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 by means of the heuristic presented here, as it has been observed from other applications in different fields (as sea sponges [GIBE92]).

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

| **Classe** | | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| **Proto-** | **Display** | | COLOR-TV | B-&-W-TV | Built-in 3/4<br>Terminal 1/4 | TERMINAL |
| | **RAM** | | 40 | 44 | 60 | 64 |
| | **ROM** | | 47/4 | 9 | 31/4 | 8 |
| **tipus** | **Micro-** | 6502 | 1/2 | 1/4 | | HP |
| **de** | **proces-** | 6502A | 1/2 | 1/2 | | |
| **la** | **sador** | Z80 | | | 3/4 | |
| | | 8080A | | | 1/4 | |
| | | 6502C | | 1/4 | | |
| **classe** | **Teclat** | 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<br>ATARI-800<br>COMMODO-<br>RE-VIC-20-A<br>COMMODO-<br>RE-VIC-20-B | EXIDI-<br>SORCERER<br>TRS-80-I<br>OHIO-SC.-<br>CHALLEN.<br>OHIO-SC.-<br>II-SERIES | ZENITH-<br>H8<br>ZENITH-<br>H89<br>HORIZON<br>TRS-80-III | HP-85 |

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

to do that. This research is actually in progress [GIBE94], 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 [BELA91]. As a first approach, giving weights to the variables may be considered, although there may be some other possibilities.

# 7 Aknowledgements

Special thanks to Dr. Ramon Nonell, Dr. Miquel Barceló and Dr. Tomàs Aluja for their invaluable comments, and their support.

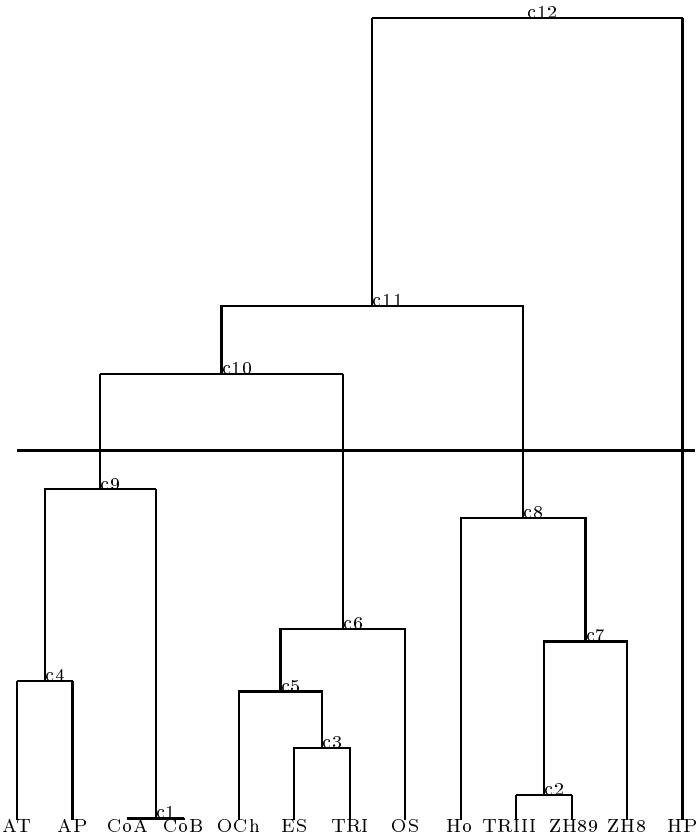


Figure 1: Dendrogramme for microcomputers. Ward's criterion and mixed metrics ($\alpha = 0.014$ and $\beta = 0.986$).

## References

[BÉJA94] Béjar, J., Adquisición de conocimientos en dominios poco estructurados (Dep. LSI, UPC, Barcelona, 1994) Ph. D. thesis.

[BELA91] Belanche, L., Cortés, U., The nought attributes in KBS (*EUROVAV-91*) 77–103.

[BENZ80] Benzecri, J.P., *L'analyse des données* (Dunod, Paris, 1980).

[DILL84] Dillon W.R., Goldstein M., *Multivariate analysis. Methods & applications* (Wiley, USA, 1984).

[EVER74] Everitt, B., *Cluster analysis* (Heinemann Educational Books Ltd, London, 1974).

[GIBE95] Gibert, K. Classification based on rules. Boletín de la *ACIA,n^o* **4**. IIIA, Barcelona 1995. 115–119.

[GIBE94] Gibert, K. L'ús de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats (Dep. EIO, UPC, 1994) Ph. D. thesis.

[GIBE91] — , **KLASS**: Estudi d'un sistema d'ajuda al tractament estad. de grans bases de dades (Dep. LSI, UPC, 1991) Master thesis.

[GIBE94b] Gibert, K., Cortés, U. 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$^o$ 89 (Springer-Verlag, New York, 1994) 351 – 360.

[GIBE93] Gibert, K., Cortés, U. On the uses of the expert knowledge for automatic biaising of a clustering method. Proceedings of *ITI'93*. University Computing Center of Croatia. 219–224.

[GIBE92] — , **KLASS:** Una herramienta estadística para la creación de prototipos en dominios poco estructurados (*IBERAMIA-92*, Noriega, México, 1992) 483–497.

[GIBE96] Gibert, K., Hernández-Pajares, M., Cortés, U. Classification based on rules: an application to Astronomy. In proc *IFCS'96*. Kobe, Japan, pp. .

[GOWD92] Gowda, K. C., Diday, E. Symbolic clustering using a new similarity measure, *IEEE Trans. on systems, man, and cib.*, **22**(2) (1992) 368–378.

[GOWE71] Gower,J. C., A general coefficient for similarity, *Biometrics*, ( 27) 857–872.

[LEBA85] Lebart, L., *et al.*, Traitement des données statistiques (DUNOD, Paris, (1982) (2$^d$ ed.)

[MART91] Martín, M. *et al.*, Knowledge acquisition combining analytical and empirical techniques, *ML* (91) 657–661.

[MICH83] Michalski, R. S., Stepp, R. E., Automated construction of classifications: Conceptual clustering versus numerical taxonomy, *IEEE Trans. on PAMI* (5) (1983) 396–410.

[MICH82] Michalski, R. S. *et al.*, PLANT/ds: An expert consulting system for the diagnosis of soybean diseases, (*Eur. Conf. of AI.* Orsay, France, 1982).

[QUIN84] 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.* (Tioga, PA, 1984) 463–482.

[QUIN83] — , Learning efficient classification procedures, *ML an A.I. perspective* (Tioga, PA, 1983).

[ROUX85] Roux, M., *Algorithmes de classification* (Masson, Paris, 1985).

[SIER89] Sierra, C. **MILORD II**: Arquitectura multi-nivell per a sistemes experts en classificació (Dep. LSI, UPC, Barcelona, 1989) Ph. D. thesis.

[SHOR76] Shortlife, E. H., **MYCIN**: A rule-based computer program for advising physicians regarding antimicrobial therapy selection (Standford, Univ. USA, 1976) Ph. D. thesis.

[VOLL85] Volle, M., *Analyse des données* (Economica, Paris, 1985).

[WOLF70] Wolfe, J. H., Pattern clustering by multivariate mixture analysis, *MBR* (5) (1971) 329-350.