Impact of Mixed Metrics on Clustering*

Karina Gibert and Ramon Nonell

Department of Statistics and Operation Research,
Universitat Politècnica de Catalunya, C. Pau Gargallo 5, Barcelona 08028, SPAIN.
karina@eio.upc.es

Abstract. One of the features involved in clustering is the evaluation of distances between individuals. This paper is related with the use of mixed metrics for clustering messy data. Indeed, when facing complex real domains it becomes natural to deal simultaneously with numerical and symbolic attributes. This can be treated on different approaches. Here, the use of mixed metrics is followed.
In the paper, a family of mixed metrics introduced by Gibert is used with different parameters on an experimental data set, in order to assess the impact on final classes.

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

1 Introduction

Clustering is one of the more used technique to separate data into groups. 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 [18]. Also, 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. That's why, several well known expert systems (MYCIN [23], ...) are actually classifiers.

However, when facing *ill-structured domains* as mental disorders, sea sponges, disabilities... clustering has to be done on heterogeneous data matrices. In this kind of domains (see [5], [6]), the consensus among experts is weak —and sometimes non-existent; when describing objects, 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 who is describing the objects: the more he knows about the domain, the greater is the number of modalities he uses.

In this work, mixed metrics introduced by Gibert in [4],[10] for measuring distances with messy data is used. This measure has been successfully implemented in a clustering system called **Klass** [5], [6] and applied to very different ill-structured domains [7], [9], [11], [12].

* This research has been partially financed by the project *CICYT'2000*.

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 464–471, 2003.

Main goal of this paper is to study the behavior of different metrics of Gibert's family (which also includes Ralambondrainy proposals as particular cases) on a set of experimental data that presents different structures, in order to study which parameters of Gibert's metrics perform better in clustering, according to the data structure. Formal approach to this problem requires a too complex theoretical development. That's why an experimental approach is presented here as a first step of the research. A similar experiment was also performed by Diday in [2], comparing the performance of two metrics [14] and [13] in clustering. Next step of this work is to make a global comparison.

This paper is organized as follows: after the general introduction, an overview of the possibilities of working with messy data is presented. Then, details on the indexed family of distances that combines qualitative and quantitative information introduced by Gibert is presented in 3, together with several proposals on the indexes values. In section 4 the experiment context is provided. Section 4.1 introduces the experimental data sets, while section 5 presents main results. Finally, the last section presents some conclusions and future work.

2 Clustering Heterogeneous Data Matrices

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. Upon [1], data analysis with heterogeneous data bases may follow three main strategies:

Variables partitioning. It consists on partitioning the variables upon their type, then reducing the analysis to the dominant type (determined owing to the group with a greater number of variables, or the group containing the more relevant variables, or the background knowledge on the domain...).

For example, if dominant type is qualitative variables, then correspondence analysis could be used and later a clustering on the factorial components is possible [24], [17]. Since the classification is performed in a fictitious space, additional tools are required to enable the interpretation of the results.

This approach of course misses the information provided by the non dominant groups of variables. A natural extension is to perform independent analysis inside every type of variables. Problem, then, is later integration of results of parallel analysis to produce a consistent, coherent and unique final result. Even in this case, interactions between variables of different types (like cooking temperature and final color of a ceramic) cannot be analyzed under this approach.

Variables converting. It consists on converting all the variables to a unique type, trying to conserve as much original information as possible. First of all, final converting type has to be decided. Conversion is not a trivial process (every variable may be converted to a unique one, or split to a group of variables or several original variables will be grouped to a unique transformed one). In Statistics, traditionally, symbolic variables has been converted to a set of binary variables,

to generate the *complete incidence table*. Then, clustering using χ^2 metrics may be performed [3]. Dimensions of the complete incidence table implies a significant cost increase. In Artificial Intelligence, grouping of quantitative values into a qualitative one [22] is much more popular. This transformation implies a relevant loss of information as well as the introduction of some instability in the results, which depend on the defined grouping.

Many authors, among them [1], [16], discuss different strategies on this line, together with the associated problems of loose of relevant information or even making difficult final interpretation, since the transformed variables could be in a fictitious space. Also, in [5] it is shown how converting all the variables to qualitative ones introduces, almost always, a bias on the results, which can be sometimes even arbitrary.

Compatibility measures. It consists on the use of compatible measures which cover any combination of variable types, making an homogeneous treatment of all the variables. It can, for instance, be defined a non-senseless distance (or similarity) between individuals which uses different expressions for every variable type.

The idea is to allow clustering on a domain simultaneously described by qualitative and quantitative variables without transforming the variables themselves. Since in the core of the classification process distances between individuals have to be calculated, a function to do it with non homogeneous data has to be found. In the literature several proposals on this line can be found. Upon discussions on [4] and [5], this is the approach of this work. Main advantages of this approach are: respecting the original nature of data, there is not loss of information, it is not necessary to take previous arbitrary decisions which can bias results, it is possible to study all the variables together, it is possible to analyze interactions between variables of different types. Proposals on this line could be, chronologically: Gower 71 [14], Gowda & Diday 91 [13], Gibert 91 [4,8], Ichino & Yaguchi 94 [15], Ralambondrainy 95 [21], Ruiz-Schulcloper [19].

3 Gibert's Mixed Metrics

The input of a clustering algorithm is a data matrix with the values of K variables $X_1 \dots X_K$ observed over a set $\mathcal{I} = \{1, \dots n\}$ of individuals. Variables are represented in columns while individuals in the rows of data matrix. The cells contain the value, x_{ik}, taken by individual $i \in \mathcal{I}$ for variable $X_k, (k = 1 : K)$. In our context, heterogeneous data matrices are supposed, so let us name $\zeta \in \{1 \dots K\}$ the indexes of numerical variables and $Q = \{1 \dots K\} - \zeta$ the indexes of categorical variables, being $n_\zeta = card(\zeta)$ and $n_Q = card(Q)$.

Mixed metrics introduced by Gibert in [4], [10] is defined, for clustering purposes as a family of metrics indexed by the pair (α, β):

$$\{d^2_{(\alpha,\beta)}(i, i')\}_{(\alpha,\beta)\ \in [0,1]\times[0,1]} \tag{1}$$

Being, $d^2_{(\alpha,\beta)}(i,i') = \alpha d^2_\zeta(i,i') + \beta d^2_Q(i,i')$; $(\alpha,\ \beta)$ indexes for weighting the influence of variables in ζ versus those in Q; $d^2_\zeta(i,i')$ the normalized euclidian metrics calculated with variables in ζ and $d^2_Q(i,i')$ a rewriting of χ^2 metrics calculated with variables in Q, supporting symbolic representation:

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

where $s^2_k = var(X_k)$. Referring to $d^2_k(i,i')$, I^{k_j} is the number of observations equal to the $j-th$ modality of X_k (namely c^k_j); $I_{k^i} = card(\hat{i} : x_{\hat{i}k} = x_{ik})$. An *extended value* appears for a class representative if X_k is not constant inside the class; it is represented as $(f^{k_1}_i, f^{k_2}_i, ..., f^{k_{n_k}}_i)$ where $f^{k_j}_i, j = 1, 2, ..., n_k$, is the proportion of objects of the class represented by i with $x_{ik} = c^k_j$, then

$$d^2_k(i,i') = \begin{cases} 0 & \text{, if } x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}} & \text{, if } x_{ik} \neq x_{i'k} \\ \frac{(f^{k_s}_i - 1)^2}{I^{k_s}} + \sum_{j=1, j\neq s}^{n_k} \frac{(f^{k_j}_{i'})^2}{I^{k_j}} & \text{, if } x_{ik} = c^k_s \text{ and } i' \text{ extended on } X_k \\ \sum_{j=1}^{n_k} \frac{(f^{k_j}_i - f^{k_j}_{i'})^2}{I^{k_j}} & \text{, for } i \text{ and } i' \text{ extended on } X_k \end{cases}$$

In [10] an heuristic criteria is used to propose proper values for index (α, β):

$$\alpha_0 = \frac{\alpha}{\alpha + \beta} \ \ \& \ \ \beta_0 = \frac{\beta}{\alpha + \beta} \quad ; \quad \alpha = \frac{n_\zeta}{d^2_{\zeta max^*}} \ \ \& \ \ \beta = \frac{n_Q}{d^2_{Q max^*}} \tag{3}$$

with $d^2_{\zeta max^*} = max_{i,i'}\{d^2_\zeta(i,i')\}$ and $d^2_{Q max} = max_{i,i'}\{d^2_Q(i,i')\}$. This values[1] refers the two components of the distance to a common interval, in order to give equal influence in the determination of $d^2_{(\alpha,\beta)}(i,i')$; the numerators give to each component a proportional weight to its presence in the objects description.

Ralambondrainy proposal. In [21], Ralambondrainy also proposes a metrics to work with heterogeneous data matrices; it is defined exactly as expression (1). In [20], two practical ways of standardization for calculating (α, β) are presented:

- by the inertia: $\pi_1 = \frac{1}{n_\zeta}$; $\pi_2 = \frac{1}{\sum\{n_k - 1 \ : \ k \in Q\}}$
- by the norm: $\pi'_1 = \frac{1}{\sqrt{\sum\{\rho^2(X_k, X_{k'}) : k, k' \in \zeta\}}}$; $\pi'_2 = \sqrt{\sum\{n_k - 1 : k \in Q\}}$, $\rho^2(X_k, X_{k'})$ correlation between X_k,$X_{k'}$; n_k number of modalities of X_k.

Those proposals identify two elements of the Gibert family of mixed distances that will also be considered in this paper.

[1] Maximums can also be truncated to the 95% in order to acquire more robustness.

4 The Experiment

As said before the main goal of this paper is to analyze the behavior of different elements of Gibert's family in the clustering of different data sets. An experiment was designed according to that. For this work a single clustering algorithm will be considered: a hierarchical reciprocal neighbors algorithm using Ward criteria (see [24]). In future works other algorithms will also be taken into account. As a first approach, four elements of Gibert's family will be considered in the experiment: $d^2_{(\alpha_0,\beta_0)}(i,i')$ as proposed by Gibert, $d^2_{(\pi_1,\pi_2)}(i,i'), d^2_{(\pi'_1,\pi'_2)}(i,i')$, as proposed by Ralambondrainy, and $d^2_{(0.5,0.5)}(i,i')$ which represents a non-informed option with equal contribution to the distances of both components.

On the other hand, experimental data has to be simulated (see §4.1). Structure of data sets was decided on the basis of factors that can influence into the behavior of the metrics, regarding the clustering process: *distinguishability* of the classes is relevant (that's why some data sets will contain overlapping classes and others separated ones, variance of classes will also be considered); also, the *form* of the classes is important (recognition of convex or filiform classes will be tested); finally different *number of classes* will be tested.

For all the data sets, four clustering processes will be performed, one with every metrics indicated above. On the results of every clustering, relevant information will be codified in a new data matrix. A multivariate analysis will be done with it, to see relationships among different runs. It seems reasonable to determine good behavior on the basis of real data structure recognition, which is easy with simulated data, since real class of every object is a priori known.

4.1 The Simulated Data Set

The basis of experimental data is also following the guidelines presented in [2], where comparison of several hierarchical clustering methods is performed using several kinds of experimental data with different structures. Figure 1 shows the experimental data sets used in this work. It is obvious that it only shows the structure of the numerical part of data sets. Every data set contains also as many categorical variables as numerical, randomly generated with 3 modalities.

Some data sets correspond to the proposal presented in [2], others are specially introduced for our purposes. The basic structures from [2] are: *concentric classes* (fig. 1(d)), *chained classes* (fig. 1(f)), *mixture of convex and concentric classes* (fig. 1 (e)) and *filiform classes in 2D* (figure 1(g)), since it is known that certain clustering algorithms perform confusing recognition in this case. Regarding the discussion previously introduced, and making wider the scope of the analysis, other structures were added in the experiment: *uniform*, representing lack of structure (fig. 1(j)), *convex classes* (fig. 1(a,b,c)), which are supposed to be the easier to recognize; variability of the classes is increasing from (a) to (c) in such a way that distinguishability of classes decreases. Finally, *filiform classes in three dimensions* (see figure 1 (h)).

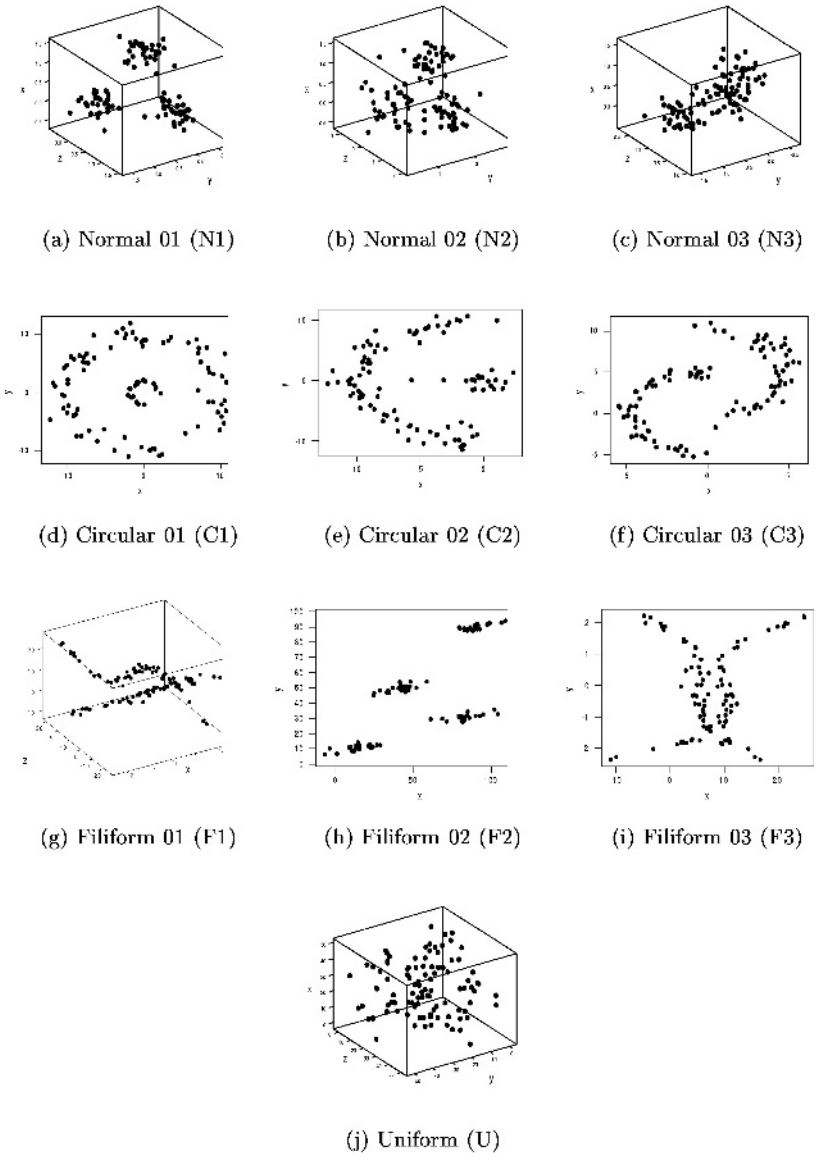
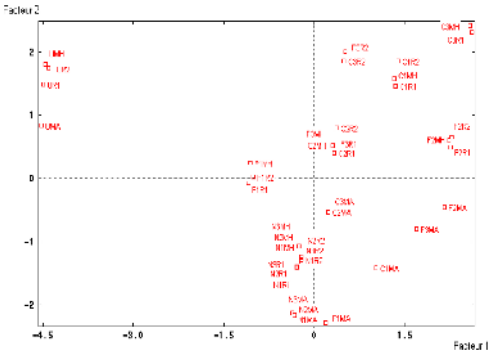


Fig. 1. Experimental data sets

5 Results

Every data set is clustered using the four metrics given in §4. Then some relevant information on the results (like number of resulting classes —which is an output in hierarchical clustering, size of every class, number of real classes, real classes form, etc) is used for a later Principal Components Analysis, in order to study

relationships among runs; tax of misclassification is used as a quality measure of runs (fig. 2 shows the projection of the runs on the first factorial plane).



Code	Metrics
MA	$d^2_{(\alpha_0,\beta_0)}(i,i')$
MH	$d^2_{(0.5,0.5)}(i,i')$
R1	$d^2_{(\pi_1,\pi_2)}(i,i')$
R2	$d^2_{(\pi'_1,\pi'_2)}(i,i')$

For data set codes, see fig. 1

Fig. 2. First factorial plane with experimental run represented.

It seems that first axis is opposing less structured data sets (on the left hand side, like *uniform*) against more structured ones (on the right, *filiform 02* or *circular 03*). The more remarkable thing, regarding the second axis, is that given a data set, its four runs use to be vertically displayed in two subsets: on the lower side the clustering performed with Gibert proposal $d^2_{(\alpha_0,\beta_0)}(i,i')$ (*–MA* in the figure), much more down than a second group, where the rest of runs appear in close neighborhood (except for *circular 02* and *filiform 03* for which runs with $d^2_{(\pi'_1,\pi'_2)}(i,i')$ (*–R2* in the figure) are projected in intermediate positions). So, in general it can be said that the second axis is opposing Gibert proposal for (α,β) to the other ones, which are difficult to distinguish.

6 Conclusions and Future Work

It has been seen that changing metrics produces real effects on the clustering results. It is then important to know when different metrics have better behaviour for recognizing real classes.

From the four studied elements of Giber't family, $d^2_{(\alpha_0,\beta_0)}(i,i')$ is the one which produces more different results on the used experimental data sets. It seems, from this work, that the other three possibilities do not produce great differences on the used clustering method. In addition, for case *filiform 03*, $d^2_{(\alpha_0,\beta_0)}(i,i')$ is the only one that allow recognition of real classes.

Next step is to complete the experiment in order to check if this separate behavior of $d^2_{(\alpha_0,\beta_0)}(i,i')$ is maintained, and if it is possible to obtain more knowledge on the other metrics; it will be interesting to work with different structures on the categorical part of data matrix, which was blocked for this work to uniform distribution. After that, comparison with results reported in [2] will also be done, as well as with other proposals from the literature, like Gower coefficient.

In the last step, including other clustering algorithms will enable study of more general properties of those metrics.

Acknowledgements. To Àngela Twose for his help on running experiments.

References

1. Michel R. Anderberg. *Cluster Analysis for aplications.* Academic Press, 1973.
2. E. Diday and J.V. Moreau. Learning hierarchical clustering from examples. In Rapport N 289 Centre de Rocquencourt, editor, *INRIA*, 1984.
3. W.R. et al. Dillon. *Multivariate analysis. Methods & applications.* Wiley, 1984.
4. K. Gibert. Klass. Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Master's thesis, UPC, 1991.
5. K. Gibert. *L'us de la Informació Simbòlica en l'Automatització del Tractament Estadístic de Dominis Poc Estructurats.* phd. thesis., UPC, Barcelona, Spain, 1994.
6. K. Gibert. The use of symbolic information in automation of statistical treatment for ill-structured domains. *AI Communications.*, 9(1):36–37, march 1996.
7. K. Gibert and R. et al. Annicchiarico. Kdd on functional disabilities using clustering based on rules on who-das ii. In *ITI 03.*, pages 181–186, Croatia, 2003.
8. K. Gibert and U. Cortés. KLASS: Una herramienta estadística para ... poco estructurados. *proc. IBERAMIA-92.*, pages 483–497, 1992. Noriega Eds. México.
9. K. Gibert and U. Cortés. Combining a knowledge-based system and a clustering method ... volume 89 of *LNS*, pages 351–360. Springer-Verlag, 1994.
10. K. Gibert and U. Cortés. Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing*, 4(3):251–266, 1997.
11. K. Gibert and U. Cortés. Clustering based on rules and knowledge discovery in ill-structured domains. *Computación y Sistemas.* 1(4):213–227, 1998.
12. K. Gibert and Z. Sonicki. Classification Based on Rules and Thyroids Dysfunctions. *Applied Stochastic Models in Business and Industry*, 15(4):319–324, octuber 1999.
13. K. Chidananda Gowda and Diday E. *Symbolic clustering using a new similarity measure.* IEEE Tr SMC, Vol 22, No 2, March/April, 1991.
14. J.C. Gower. A General coefficient if similarity ... *Biometrics*, 27:857–874, 1971.
15. M. Ichino and H. Yaguchi. Generalized Minkowski Metrics for Mixed feature-type data analysis. *IEEE Tr SMC*, 22(2):146–153, 1994.
16. L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data, an Introduction to Cluster Analysis.* John Wiley & Sons, London (England), 1990.
17. L. Lebart. *Traitement statistique des données.* Dunod, Paris., 1990.
18. G. Nakhaeizadeh. Classification as a subtask of of Data Mining experiences form some industrial projects. In *IFCS*, volume 1, pages 17–20, Kobe, JAPAN, 1996.
19. J. R-Schulcoper. Data analysis between sets of objects. In *ICSRIC'96.* 85-81, vIII.
20. H. Ralambondrainy. *A clustering method for nominal data and mixture ...* H.H.Bock, Elsevier Science Publishers, B.V. (North-Holland), 1988.
21. H. Ralambondrainy. *A conceptual version of the K-means algorithm.* Lifetime Learning Publications, Belmont, California, 1995.
22. M. Roux. Algorithmes de classification, 1985. Paris: Masson, Paris, France.
23. E. H. Shortlife. *MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection.* PhD thesis, Standford.
24. M. Volle. Analyse des données, 1985. Ed. Economica, Paris, France.