A COMBINED MULTIDIMENSIONAL SCALING + SELF-ORGANIZING MAPS METHOD FOR EXPLORATORY ANALYSIS OF QUALITATIVE DATA

E. Miret[(1)], F. García-Lagos[(2)], G. Joya[(2)], H. Arazoza[(1)] F. Sandoval[(2)]
[(1)]Dpt. Ecuaciones Diferenciales.
Facultad de Matemática y Computación. Universidad de La Habana.
Ciudad de La Habana. Cuba
elina@matcom.uh.cu, arazoza@matcom.uh.cu
[(2)]Dpt. Tecnología Electrónica
ETSI Telecomunicación. Universidad de Málaga
Málaga. Spain
gjoya@uma.es, lagos@dte.uma.es, sandoval@dte.uma.es

Abstract – *This paper describes a method that combines Multidimensional Scaling and Self-Organizing Maps (MDS+SOM). The main aim of the proposed method is classification and exploratory analysis of populations described by means of qualitative variables.* A*s a merely illustrative case of study, the method has been applied to the exploratory analysis of a population of HIV-AIDS infected anonymous individuals, in order to disclose factors that influence the detection of the disease.*

Key words – Multidimensional Scaling, Self-Organizing Maps, Qualitative variables, Exploratory Analysis

1 Introduction

The well known ability of Self-Organizing Maps by Kohonen [1] for projection and topology preservation has permitted their extensive use in classification and behaviour prediction of complex system [2]. The easy visualization of their output representation and their unsupervised nature make this algorithm a very advisable method for exploratory analysis, since it searches for relations among the variables describing a population of individuals, which are unknown a priori. This latter usage may be applied to model the behaviour of a complex system by means of pointing out the factors intervening in its evolution

The original definition of the SOM is based upon the Euclidean distance between input patterns and neuron weights as the criterion for the winning neuron, hence resulting in an appropriate treatment of quantitative variable vectors. The adaptation of the algorithm to dealing with both quantitative and qualitative data is described in [3]: The Map generates a classification from the quantitative variables and then analyses, for each obtained class, the distribution of the modalities of each qualitative variable. In order to populations only described by qualitative variables, several SOM adapted algorithms are described in [4]: KACM carries out a classification of the modalities by considering as input the rows of the Burt matrix; KACM-ind performs a classification of individuals by setting the input as the

rows of the Complete Disjunctive matrix; finally, KDISJ classifies both modalities and individuals by defining as inputs the rows of the Corrected Disjunctive matrix.
In this work we present a different approach for classifying individuals which are described by means of qualitative variables. This approach is based on the combination of Multidimensional Scaling (MDS) and Self-Organizing Maps by Kohonen (SOM) methods. At a first stage, the algorithm profits from the capacity of MDS for projecting a set of qualitative component vectors onto an Euclidean k-dimensional space. Thus, SOM algorithm is not applied to the original Disjunctive matrix but rather to the k-dimensional projected vectors. A remarkable feature of MDS methods is their applicability when a matrix of distances or dissimilarities between individuals is available, rather than a description of the individuals themselves. In this sense, the described method may extend the range of SOM applications to this kind of problems. This possibility is described in [5] for the Travelling Salesman Problem (TSP), where a table of cities distances is used as input information, instead of the conventional spatial coordinates of cities.
As an illustrative case of study, we apply the MDS+SOM method to the exploratory analysis of a HIV-AIDS population for detecting factors that influence the detection of infected individuals.
The rest of this paper is organized as follows: Section 2 describes the proposed MDS+SOM method, and analyses the particular features of each component, as well as their interaction possibilities; section 3 describes, as a merely illustrative case of study, the application to the classification of a HIV-AIDS infected population; section 4 analyses some results related to the infected detection time; finally, section 5 summarizes the main conclusions.

2 Combined MDS+SOM for qualitative data analysis.

Let us consider a population of N individuals which are described by means of P qualitative variables, each having a particular number of modalities m_i, where $M = \sum_{i=1}^{P} m_i$.

The MDS+SOM method proceeds through the following stages:

1.- *Generate the Complete Disjunctive matrix* $D = (d_{ij})$; $i = 1,..,N$, $j = 1,...,M$. In this matrix, d_{ij} equals 1 if the i-th individual presents the j-th modality, and 0 otherwise. Thus, for each row i, only one d_{ij}=1 in each interval $\sum_{h=0}^{q-1} m_h < j \le \sum_{h=0}^{q} m_h$, where m_q is number of modalities of variable q (q=1...P), and m_0 =0 by definition.

2.- *Apply MDS to the matrix D* [6].

2.1.- Generate the dissimilarity matrix $\Delta = (\delta_{ij})_N$ and the matrix $A = (a_{ij})_N$ with $a_{ij} = -(1/2)\delta_{ij}^2$.

2.2.- Obtain $B = (b_{ij})$ by means of $B = HAH$, where $H = I_N - (1/N)1_N 1_N^t$, 1_N^t is a column vector of N ones, and I_N is the Nth order identity matrix.

2.3.- Extract the k largest strictly positive singular values $\sigma_1 \ge ... \ge \sigma_k$ from the singular decomposition of B given by $B = V\Sigma V^t$ and its corresponding normalized singular vectors. Let $V_{(k)}=(V_1,...,V_k)$ denote the matrix of the k first columns of V and $\Sigma_{(k)}^{1/2} = \mathrm{diag}(\sigma_1^{1/2},\ldots,\sigma_k^{1/2})$, then:

$$Y_{(k)} = B_{(k)}^{1/2} = V_{(k)}\Sigma_{(k)}^{1/2} = (Y^{(1)},...,Y^{(n)})^t \quad (1)$$

The algebraic solution of equation (1) is the optimum of the STRAI function of Carrol [7]. This STRAIN may be described as following:

$$STRAIN = \left\| B - B_{(k)} \right\|^2 = tr(B - B_{(k)})^2 = \sum_{i=k+1}^{n} \sigma_i^{1/2} \quad (2)$$

Where $\sigma_{k+1} \geq ... \geq .\sigma_n$ are the n-k smallest singular values of B.

3.- *Apply SOM to* $(Y^{(1)},...,Y^{(n)})^t$

The description of the conventional SOM algorithm is omitted for brevity and the reader is referred to [8]. The SOM is applied by assuming a 10x10 neuron grid, and the learning rate is given by equation (3) [9],

$$l_r(t) = l_{r0} / (1 + \frac{ct}{nn}) \quad (3)$$

where l_{r0} is the initial learning rate (0.3 in our experiments), c is a constant (0.2), t is the current iteration and nn is the number of neurons.

Regarding the above described algorithm, some advantages and limitations of MDS and SOM methods can be brought to light. The benefits of their cooperation can be justified, too.

On one hand, MDS can generate a set of points in a k dimensional space as a projection of the row vectors of the matrix D. It can even generate these k-dimensional vectors from a matrix of distances or dissimilarities. Its computational cost depends on the number of individuals rather than their dimension, so it can be used to face those problems having a high number of variables with a high number of modalities. However, this method only allows the visualization on two dimensions (three at most) of the vectors of matrix $Y_{(k)}$, hence one can only visualize the possible relations between pairs of variables, but no global structure is available in order to analyze all possible relations.

On the other hand, the computational cost of SOM depends on the dimension of the input vectors. Thus, it can be applied to a continuous variable vector rather than a (binary) row of the matrix D. Its self-organization after learning results in a single structure where every set of variables can be analyzed, so that the exploratory analysis of the population is considerably easier. Consequently, the sequential MDS-SOM application seems reasonable, because MDS produces the adequate inputs to SOM, whereas this latter provides an output that is more suitable for the data analysis.

3 Classification of a HIV-AIDS infected population.

For testing purposes, the combined MDS+SOM method has been applied to finding the possible groups in a database of 999 patterns containing information about an anonymous population of HIV-AIDS infected individuals, whose infection was detected from 1990 to 1996. The considered qualitative variables are described in table 1.

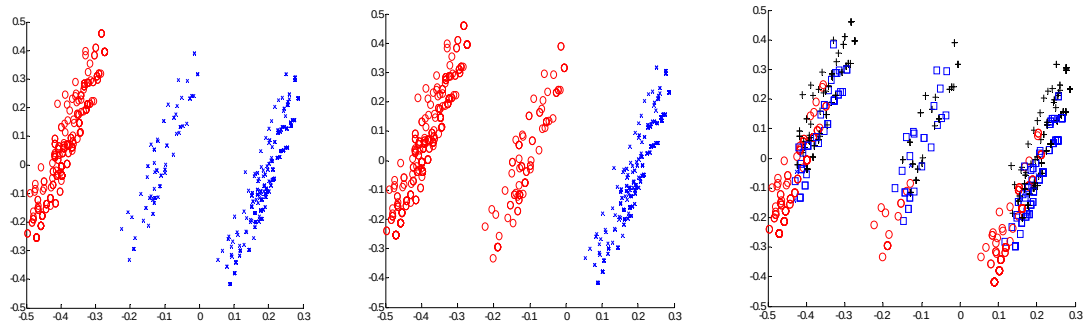
Five variables have been considered with a total of 14 modalities. In this case, the Complete Disjunctive matrix will be $D = (d_{ij})_{999x14}$, where the element d_{ij} equals 1 if the individual i presents the modality j, and equals 0 otherwise. At the MDS phase, the coefficient of Sokal was used to obtain the dissimilarity matrix Δ_{999}. The MDS output is the matrix $Y_{(5)}$, which presents the 88.41% of data significance. This 999x5 dimensional matrix represents each one of the population individuals by means of only five variables, those corresponding to the largest five singular values

obtained at the step 2.3 of the above described algorithm. The rows of this matrix will be the input patterns to the SOM.

Table 1. Qualitative variables and their respective modalities in a HIV-AIDS database.

Variable	*Modalities*
Age	$mod_1 \equiv (13,20]$, $mod_2 \equiv (20,30]$, $mod_3 \equiv (30, \rightarrow)$
Sex	$mod_4 \equiv$Male, $mod_5 \equiv$Female
Civil State	$mod_6 \equiv$Married, $mod_7 \equiv$Single,
Academic Level	$mod_7 \equiv$ Illiterate, $mod_8 \equiv$Primary level, $mod_9 \equiv$Secondary level, $mod_{10} \equiv$Bachelor, $mod_{11} \equiv$Technichal$_{12} \equiv$University
Sexual Orientation	$mod_{13} \equiv$Homosexual, $mod_{14} \equiv$Heterosexual

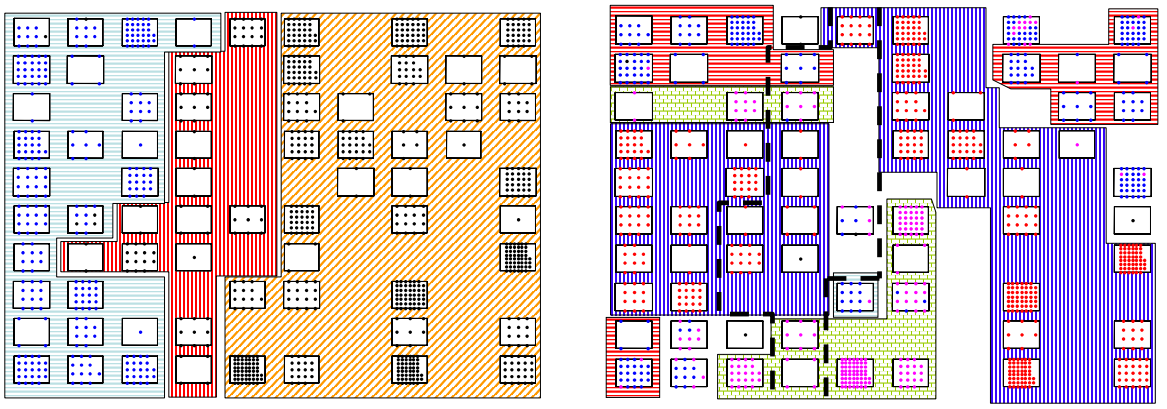
Figure 1 represents the obtained distribution of individuals, by showing the first two dimensions of $Y_{(5)}$, so that the classification of the population can be viewed as a function of the *Sex* (figure 1.a) or the S*exual Orientation* (figure 1.b). However, it is impossible to obtain information about every other criterion, so in figure 1.c, representing the population distribution as a function of the *Age*, all modalities appear merged. This is the main MDS limitation, because it is necessary to visualize multiple two (or three) dimensional projections of $Y_{(k)}$ in order to analyze the influence of each variable.



a) b) c)

Figure 1. Graphical representation of the first two dimensions of $Y_{(5)}$. a) Classification by sex (Circles (o) represent female and crosses (x) represent male). b) Classification by sexual orientation (Circles (o) represent heterosexual and crosses (x) homosexual). c) Classification by age (Circles (o) represent $13 \leq age < 20$, plus signs (+) represent $20 \leq age < 30$ and squares (□) represent $30 \leq$ age classes).

On the other hand, figure 2 represents different classifications on the final organized SOM. Remarkably, a single map allows to visualize both the *Sex* or S*exual Orientation* classification (figure 2.a) and the *Age* classification (figure 2.b). The exploratory analysis of the relation between every population feature and variables used for classification will be clearly boosted starting from the SOM output.



a) b)

Figure 2. Final organization of Kohonen map. a) Classification according to *Sex* and *Sexual Orientation* variables (represents heterosexual female, represent heterosexual male and represents homosexual male). b) Classification according to *Age* (represents 13 ≤ age < 20, 20 ≤ age < 30, and 30 ≤ age).

4 Exploratory Analysis of Disease Detection Factors.

The available database provides both date of detection of the infection (*Ddet*) and date of development of the disease (*Daids)* of each individual, but it does not provide the date of infection. Consequently, the individual detection time can not be directly determined. We start from the hypothesis that once the infection is detected, palliative cares are uniformly applied to the infected population, thus the latency time (t_{lat}=*Daids – Ddet*), may give an approximate idea about the detection time: the greater t_{lat}, the earlier the detection, and viceversa.

A visual analysis of the resulting Kohonen map provides an estimation, for each class, of the proportion of individuals whose latency time t_{lat} belongs to a certain range. Thus, in figure 3, dots (•) represent individuals whose time $t_{lat} \leq 3$, crosses (x) represent individuals with $3 < t_{lat} \leq 7$, and circles (o) represent individuals with $7 < t_{lat}$. At first sight, the proportion of crosses and circles is larger for female than for male. Also, for the female class, it seems that there is a larger proportion of these symbols for the *Age>30* class. With respect to the male class, crosses and circles seems to appear more often within the *Homosexual* class.

These qualitative perceptions can be numerically confirmed by means of the bar graphs in figure 4, which have been obtained from the database a posteriori. Indeed, figure 4.a shows a larger proportion of $t_{lat} < 3$ individuals in the Heterosexual-Male class with respect to the Homosexual-Male class. This fact suggests a larger detection delay for the former class. Figure 4.b shows a behaviour for the *Age>30-Female* class, which is specially interesting: the proportion of individuals with $t_{lat} < 3$ is extraordinarily low with respect to the rest of *Age* classes.

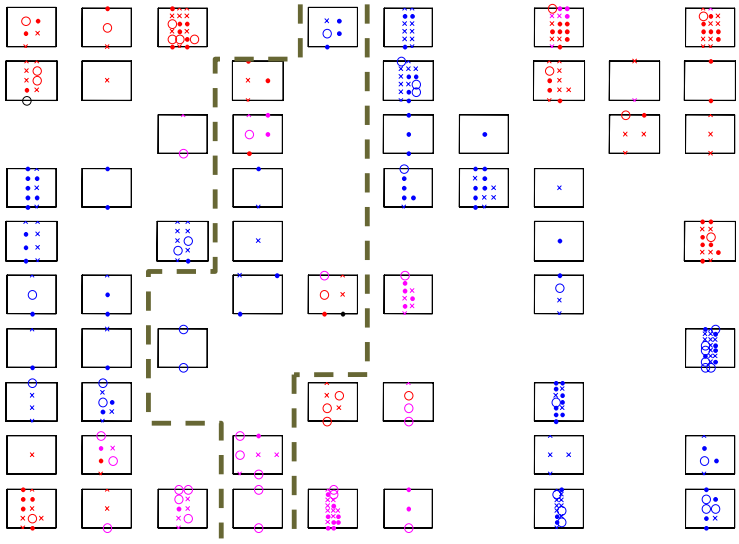


Figure 3. Representation of Kohonen classification according to latency time. Dots (•) represent individuals with $t_{lat} \leq 3$, crosses (x) represent $3 < t_{lat} \leq 7$, and circles represent $t_{lat} > 7$.

The presented exploratory analysis encourages us to outline the following hypothesis: on the one hand, the current detection policies seem more effective for the homosexual class; on the other hand, detection seems to be specially effective for the *Age >30-Female* class. Eventually, this 30 year frontier could probably be refined by means of a more careful set of *Age* modalities in the original database.

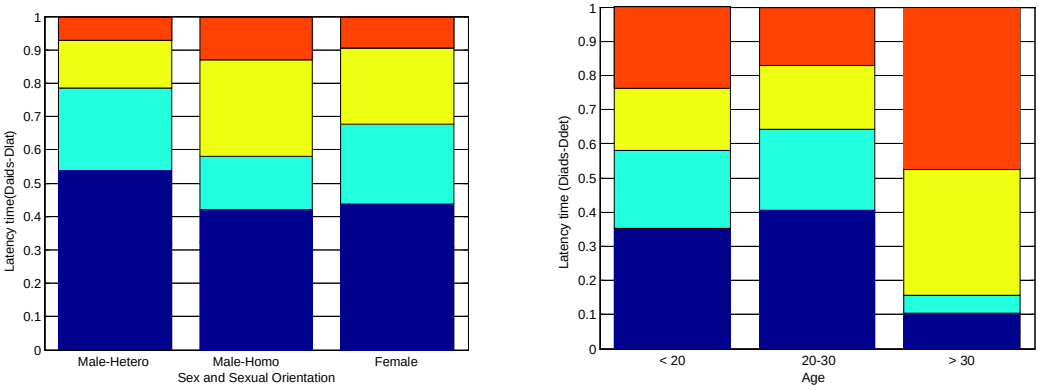


Figure 4. Population distribution according to latency time. From bottom to top, blocks represent population proportions with latency time $t \leq 3$, $3 < t \leq 5$, $5 < t \leq 7$ and $t > 7$, respectively. a) Comparison between heterosexual male, homosexual male and female (from left to right). b) Distribution for female 13≤ age< 20, 20≤ age< 30, and 30≤ age.

Numerical analyses of several epidemic models [10] confirm the intuitive idea that an early detection is a fundamental factor for keeping infection rates controlled, as well as improving the quality of life of infected population. Thus, exploratory studies such as the one outlined here could be useful to medical experts in order to search new detection strategies and improve the currently existing ones.

5 Conclusions

This work describes a combined Multidimensional Scaling and Self-Organizing Map (MDS+SOM) method for the analysis of qualitative data. This method may be also applicable when only a matrix of distances or dissimilarities between individuals is available. Summarizing, starting from the complete disjunctive matrix D, the MDS algorithm produces the matrix of dissimilarities and generates a projection of data on an Euclidean k dimensional space. The new Euclidean vectors, each one representing an individual of the original population, are used as inputs to the SOM algorithm.

Preliminary experiments suggest that the cooperation between both methods allows to overcome their own limitations. On one hand, MDS is able to deal with a population defined by qualitative data, which is a major limitation of SOM. Besides, a dissimilarity matrix can be used as input. On the other hand, the analysis of the population classification, considering different sets of variables and criteria, is easily attained by visual inspection of the output of a single Kohonen map.

In order to illustrate the method viability, the classification of a population of anonymous IHV-AIDS infected individuals is chosen as a case of study. Therefore, an exploratory analysis is performed that suggest the factors that influence the latency time (from infection detection to disease development). With respect to this problem, a deeper analysis and discussion with medical experts, as well as a more refined data preprocessing, are necessary to obtain conclusions with a really practical value. As a remarkable preliminary result, we claim that the current detection policies are more effective among the homosexual population, whereas a specially effective detection factor seems to exist for women that are older than 30 years old. A refinement of this threshold would require a more detailed set of modalities for the *Age* variable.

Acknowledges

This work has been partially supported by the Agencia Española de Cooperación Internacional, Project No. A/2051/0, and the Spanish Ministerio de Educación y Ciencia, Project No. TIN2004-05961.

References

[1] T. Kohonen (1990), The Self-Organizing Map, *Proceedings of the IEEE* , **vol. 9, no.** 78, pp. 1464-1480.

[2] F. García-Lagos, G. Joya, F. J. Marín and Francisco Sandoval (2002) Self-Organizing Maps for Contingency Analysis: Visual Classification and Temporal Evolution, *Conference of the IEEE Industrial Electronics Society (IECON'2002)*, Seville, pp. 1451-1456.

[3] M. Cotrell, P. Gaubert, P. Letrémy, and P. Rousset (1999), Analyzing and representing multidimensional quantitative and qualitativedata: Demographic study of the Rhöne valley. The domestic consumption of the Canadian families, in *Kohonen Maps*, E. Oja and S. Kaski (Eds.), Elselvier, Amsterdam, pp. 1-14.

[4] P. Letrémy (2004), Traitements de données qualitatives par des algorithmes fondés sur l'algorithme de Kohonen, invited to *ACSE' 2004*, Lilles, 33 pages, ftp://samos.univ-paris1.fr/pub/SAMOS/preprints/samos204.pdf

[5] E. Miret (2005), Un enfoque unificado para técnicas de representación euclidiana *Ph. D. Thesis*, University of La Habana, (Cuba).

[6] I. Borg and P. Groenen (1997), *Modern Multidimensional Scaling. Theory and Applications*, Ed. Springer-Verlag, New York.

[7] J. D. Carroll and P. Arabie (1980), Multidimensional Scaling, *Annual Rewiew of Psychology*, 31, pp. 607-649.

[8] S. Haykin (1994), *Neural Networks. A comprehensive Foundation*, Ed. Macmillan, New York, chap. 10, pp. 411-412.

[9] M. Cottrel and P. Letrémy (1995), Classification et analyse des correspondances au moyen de L'Algoritme de Kohonen: Application à l'ètude de donnèes socio-économiques, *Prépublication du SAMOS*, (42), University of Paris I, France, 10 pages, ftp://samos.univ-paris1.fr/pub/SAMOS/

[10] H. Arazoza y R. Lounes (2002) A non linear model for sexually transmitted disease with contact tracing, *Mathematical Medicine and Biology, A Journal of the IMA*, **vol. 19** no. 3, pp. 221-234.