ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

Analyzing surveys using the Kohonen algorithm

Marie Cottrell, Patrick Letrémy,

Université Paris I, SAMOS-MATISSE, UMR CNRS 8595
90 rue de Tolbiac,
F-75634 Paris Cedex 13, France
pley,cottrell@univ-paris1.fr

Abstract :
The Kohonen algorithm (SOM, Kohonen 1995) is a very powerful tool for data analysis. Most of the time, each observation is a *p*-vector of numerical values. But in many cases, for survey analysis for example, the observations are described by qualitative variables with a finite number of modalities. In that case, we define a specific algorithm (KDISJ) which provides a simultaneous classification of the observations and of the modalities.

1. Introduction

In many real cases, in data analysis, individuals are described by some qualitative variables. This is very frequent when the data are collected after a survey, in which people have to answer a number of questions, each of them having a finite number of possible modalities (i.e. sex, professional category, level of income, etc.). Sometimes the values (codes) that are used to represent the modalities of these variables are viewed as numerical values, but it is well known that this is not correct : the code values can be not comparable, the codes are neither necessarily ordered nor regularly spaced. Most of the time, using the codes of the modalities as quantitative (numerical) variables has no meaning.

So the qualitative data need a specific treatment. Two kinds of problems can be addressed : it can be of interest to only study the relations between modalities, or (it is the case of this paper), we want to classify both individuals and the modalities of the qualitative variables that describe them. In this paper, we are interested in a simultaneous representation of modalities as the same time as individuals.

In classical data analysis, one uses factorial correspondence analysis (FCA for two variables, MCA for more than 2), which is a variant of principal component analysis. See for example Benzécri, 1973, Lebart et al., 1984, for example.

It is a projection method, and it provides several two-dimensional maps, each of them represents a small percent of the global information. So it is necessary to look at several maps at once, the modalities and/or the individuals are more or less well represented, and it is not always easy to deduce pertinent conclusions about the proximity between modalities, between individuals, and between modalities and individuals. Their main property is that each modality is drawn as an approximate

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

center of gravity of the individuals which possess it. But the approximation can be very poor.

The main idea is to use the Kohonen algorithm (SOM) which is a very powerful tool for analyzing and visualizing numerical data. However in its classical version, it is not adapted to deal with qualitative data. Hence the frame of our work is to propose modifications of the genuine SOM algorithm, designed to deal with qualitative data. In previous papers (Cottrell et al. 1993, Cottrell et al. 1995, 1997, 1998, 1999, Ibbou, 1998), we defined two algorithms (CORRESP for two variables, KMCA for more than two variables). These algorithms deal with qualitative variables, proved their capability to give good results as well for the quality of the representation as for their small computing time, but ignore the individuals.

So, in the present paper, the goal is to simultaneously represent the individuals and the modalities, by classifying them on a Kohonen map.

2. A new algorithm for a simultaneous analysis of individuals together the modalities : KDISJ

Let us define the notations. We consider a sample of N individuals and a number K of qualitative variables (i.e. the questions of the survey). Each variable $k = 1,2, ..., K$ has m_k possible modalities. For each individual, there is one and only one modality. If M is the total number of modalities, each individual is represented by a row M-vector with values in {0, 1}. There is only one 1 between the 1st component and the m_1-th one, only one 1 between the (m_1+1)-th component and the (m_1+m_2)-th one and so on. The $(N \times M)$ data matrix is called the *complete disjunctive table* and is denoted by $D = (d_{ij})$, $i = 1, .., N$, $j = 1,..., M$. The term d_{ij} takes its values in {0,1}. This table D contains all the information about the individuals.

A fruitful method would then consist of processing the qualitative variables beforehand via a multiple correspondence analysis and preserving all of the co-ordinates. Once individuals have been represented by numerical variables, they can be classified using the Kohonen algorithm. We will however *have lost the modalities*.

It is also possible to classify the modalities using KMCA algorithm, and then to classify individuals as if they were additional data (once they have been properly normalized). Conversely, it is also possible to classify individuals alone and then classify as additional data the "virtual individuals" associated with the modalities (in that case also a normalization is needed.). Both alternatives were defined by Smaïl Ibbou in his PHD thesis (Ibbou, 1998). They give good results, but the *symmetry between individuals and modalities is broken.*

In order to surpass the drawbacks of both above propositions, we consider the disjunctive table D as a large contingency table, which crosses an extra-variable MODALITY with an extra-variable INDIVIDUAL. In fact it is an extension of the

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

algorithm KORRESP, (Cottrell et al., 1993) that has been introduced to analyze contingency tables which cross two qualitative variables.

Let us define the algorithm in a self-consistent way.

We take our inspiration in the definition of the classical correspondence analysis. It is well known that it is in fact a principal weighted component analysis that uses the Chi-Square distance simultaneously along the row and column profiles. So in order to use a χ^2-distance along the rows as well as down the columns, and to weight the modalities proportionately to the size of each sample, we preprocess the complete disjunctive table, and put:

$$d_{ij}^c = \frac{d_{ij}}{\sqrt{d_{i.} d_{.j}}}.$$

where $d_{i.} = \sum_{j=1}^{M} d_{ij}$ et $d_{.j} = \sum_{i=1}^{N} d_{ij}$.

Note that for a complete disjunctive table, $d_{i.}$ is equal to K, regardless of i. The term $d_{.j}$ represents the number of persons who are associated with the modality j.

When adjusted thusly, the table is called D^c (adjusted disjunctive table). This transformation is the same as the one that Smaïl Ibbou proposes in his thesis (Ibbou, 1998, Cottrell & Ibbou, 1995). Note that after this transformation, to use the Euclidean distance on D^c is equivalent to use the weighted χ^2-distance on D.

We then choose a Kohonen network, and associate with each unit u a code vector C_u that is comprised of $(M + N)$ components, with the M first components evolving in the space for individuals (represented by the rows of D^c) and the N final components in the space for modalities (represented by the columns of D^c). We denote

$$C_u = (C_M, C_N)_u = (C_{M,u}, C_{N,u})$$

to put in evidence the structure of the code-vector C_u. The Kohonen algorithm lends itself to a double learning process. At each step, we alternatively draw a D^c row (i.e. an individual i), or a D^c column (i.e. a modality j).

When we draw an individual i, we associate the modality $j(i)$ defined by

$$j(i) = Arg \max_j d_{ij}^c = Arg \max_j \frac{d_{ij}}{\sqrt{K d_{.j}}}$$

that maximizes the coefficient d_{ij}^c, i.e. the rarest modality out of all of the corresponding ones in the total population. We then create an extended individual vector $X = (i, j(i)) = (X_M, X_N)$, of dimension $(M + N)$. See Fig. 1. Subsequently, we look for the closest of all the code vectors, in terms of the Euclidean distance restricted to the M first components. Note u_0 the winning unit. Next we move the code vector of the unit u_0 and its neighbors closer to the extended vector $X = (i, j(i))$, as per the customary Kohonen law. Let us write down the formal definition :

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

$$\begin{cases} u_0 = Arg\min_u \left\| X_M - C_{M,u} \right\| \\ C_u^{new} = C_u^{old} + \varepsilon\, \sigma(u, u_0)\,(X - C_u^{old}) \end{cases}$$

where ε is the adaptation parameter (positive, decreasing with time), and σ is the neighborhood function, such that $\sigma(u,\ u_0)$= 1 if u and u_0 are neighbour in the Kohonen network, and = 0 if not.

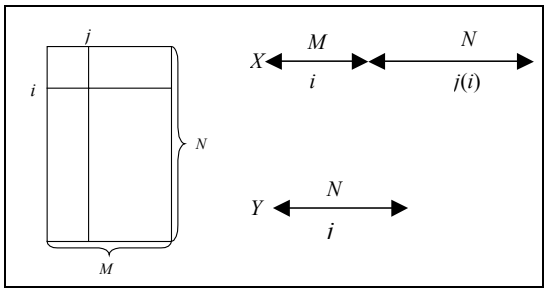

Fig 1 : The matrix D^c, vectors X and Y

When we draw a modality j with dimension N (a column of D^c), we do not associate an individual with it. Indeed, by construction, there are many equally placed individuals, and this would be an arbitrary choice. We then seek the code vector that is the closest, in terms of the Euclidean distance restricted to the N last components. Let v_0 be the winning unit. We then move the N last components of the winning code vector associated to v_0 and its neighbors closer to the corresponding components of the modality vector j, without modifying the M first components. For simplicity let us denote by Y (see Fig. 1) the N-column vector corresponding to modality j. This step can be written :

$$\begin{cases} v_0 = Arg\min_u \left\| Y - C_{N,u} \right\| \\ C_{N,u}^{new} = C_{N,u}^{old} + \varepsilon\, \sigma(u, u_0)\,(Y - C_{N,u}^{old}) \end{cases}$$

while the M first components are not modified.

By so doing, we are carrying out a classical Kohonen classification of individuals, plus a classification of modalities, maintaining all the while their association with one another. After the convergence, the individuals and the modalities are classified into Kohonen classes. “Neighboring” individuals or modalities are classified in the same class or in neighboring classes. We call this algorithm just defined KDISJ. Its computing time is small, (the number of iterations is about 15 times the data base). In the following we present two real data examples.

3. Example : Part-time employees

The data are extracted from a large INSEE 1998-1999 *Timetables* survey. See Letrémy et al, 2002, for the complete study, Cottrell and Letrémy, 2001 for a first study about part-time employees. We only consider 207 part-time employees working

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

on either an open-ended or a fixed term contract. They are described by 8 qualitative variables and 23 modalities according to the following table :

Heading	Modalities	Name
Type of employment contract	Open-ended / fixed term contract	OEC,FTC
Age	<25, [25, 40], [40,50], ≥50	AGE1,AGE2,AGE3,AGE4
Daily working schedules	Identical, Posted, Variable	HOR1,HOR2,HOR3
Saturday shifts	Never, sometimes, usually	SAT1,SAT2,SAT3
Sunday shifts	Never, sometimes, usually	SUN1,SUN2,SUN3
Able to take time off	Yes, yes under certain conditions, no	ABS1,ABS2,ABS3
Part-time status chosen	Yes, no	CHO1,CHO2
Possibility of carrying over working hours	Not applicable, yes, no	REC0,REC1,REC2

Simple cross analysis of the variables shows that the OEC contracts represent 83.57 % of all the population, while forced (and therefore involuntary) part-time work (CHO2) accounts for 46%. The goal is then to simultaneously represent all the modalities and the individuals, by realizing a classification of the (207 + 23) items. The KDISJ algorithm provides this classification, and as it is built by a SOM technique, the main associations and proximity are visible on the map.

On the map (Fig 2), a 5 by 5 grid, using the KDISJ algorithm, we display findings from a simultaneous classification of individuals and variables. In order to simplify the representation, we have in each case displayed the modalities, the number of individuals (and not the complete list), the number of persons working on a chosen or not chosen part-time work. The starred units correspond to the classes where the number of fixed term contract (FTC) is greater than the mean value in the whole population (16.43 %).

We can see immediately that the bottom of the map correspond to the unpleasant working conditions (involuntary part-time status, fixed term contract, Saturday and Sunday shift, no chance to take any time off, etc... They are the youngest persons. The relatively favorable situations are displayed in the center of the map (with identical daily working schedules, no work on Saturday, nor Sunday, chance to take time off, chosen part-time status, open end contract, etc...).

It is possible to control the good position of the modalities with respect to the individuals, by computing the deviations[1]. They are all positive, except for the modality ABS3, which should be classified in the class just above.

[1] The deviation for a modality m (shared by n_m individuals) and for a class k (with n^k, individuals) can be calculated as the difference between the number of individuals who possess this modality and belong to the class k and the "theoretical " number $n_m\ n_k\ /\ n$ which would correspond to a distribution of the modality m in the class k that matches its distribution throughout the total population.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

If we want to reduce the number of classes, it is possible to realize a clustering, by using any clustering method for the 25 code vectors. For example, we use a one-dimensional Kohonen algorithm and cluster the 25 classes into 7 macro-classes, easier to describe, if the goal is to build a typology of all the individuals together with their modalities. The SOM one-dimensional map classify the units in an ordered scale from light gray (good conditions of work) to very dark gray (bad conditions of work).

SUN2 15 (13, 2)	HOR2 5 (4, 1)	REC1 16 (15, 1)	 8 (8, 0)	ABS2 13 (9, 4)
 8 (3, 5)	SAT2 3 (1, 2)	 2 (2, 0)	AGE3 13 (8, 5)	 4 (2, 2)
HOR3 REC2 12 (3, 9)	 6 (4, 2)	AGE4 17 (13, 4)	OEC CHO1 3 (3, 0)	AGE2 16 (9, 7)
 	AGE1 * 6 (2, 4)	 10 (1, 9)	HOR1, SAT1, SUN1, ABS1,REC0	* 5 (1, 4)
SUN3 13 (9, 4)	 	SAT3 ABS3 11 (0, 11)	CHO2 * 3 (0, 3)	FTC * 18 (1, 17)

Fig 2: The map of modalities and individuals

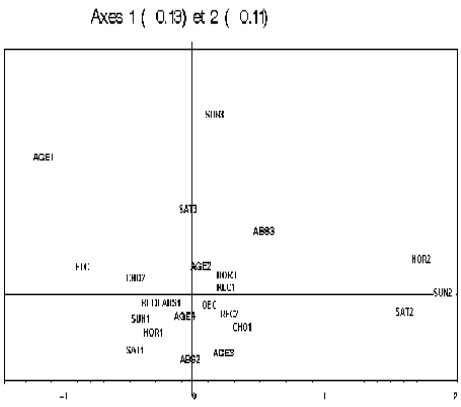

Fig 3: Correspondence analysis, axes 1 and 2 (24%)

It can be interesting to compare the map with the first projection that we get using a Multiple Correspondence Analysis, see Fig 3. We observe the same associations. The KDISJ technique do not imply several successive projections, but only one, which is very useful in the practice.

4. Example : Recurring unemployed workers

The initial data is the complete register of the unemployed held by the ANPE; information on unemployment benefits and compensations in latest job is added from the data collected by UNEDIC. The studied period goes from July 1993 to August 1996. The population is constituted of all the unemployed who were looking for a job at the beginning of this period, or who became unemployed later (but before the end of August 1996); at the end of the period they are either unemployed or their status has changed in some way. In a previous study, (Gaubert and Cottrell, 2000), we use a 1% sample of the unemployed registered in the administrative region of Ile-de-France (Paris and suburbs) having two or more spells of unemployment (590 000 individuals on a population of more than 2 millions 167 000), those named recurring unemployed. We thank ANPE and UNEDIC for the permission to use the data.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 85-92

In order to illustrate the use of the KDISJ algorithm, we restrict ourselves in this paper to a sample of 204 individuals, described only by 8 qualitative attributes and 28 modalities. They are presented in the following table.

AGE	Sub-categories of age: <25, 25-35, 35-45, 45-55, >55
BEN	Daily benefits: <60 F, 60-100, 100-150, >150
EDU	Level of education: > bac (post secondary school level), bac level (secondary school completed), <bac (secondary school not completed)
DUR	Cumulated duration of unemployment: <12 months, 12-24, >24
OCC	Monthly hours of occasional work OW : 0, 0-39, 39-78, 78-117, >117
POCC	Proportion of cumulated duration of unemployment doing OW: 0, 0-0.1, 0.1-0.3, >0.3
EXIT	Types of exits from unemployment (2 categories detailed below)
REG	Reasons for unemployment (2 categories detailed below)

The different types of *exits from unemployment* have been grouped in 2 categories:
1. job found by the individual himself or with the help of ANPE services;
2. training program, withdrawal from the labor market, other exit.

Similarly, the causes of *registration at the ANPE* have been coded in 2 modalities:
1. lay off, end of fixed-term contracts, voluntary quit;
2. first job search.

For this example, we use a one dimensional Kohonen map, with 6 units. We train it with the KDISJ algorithm, in order to simultaneously classify the modalities and the individuals. We get the following map, where the number at the bottom is the number of individuals in each class.

OCC4 OCC5 POCC3, POCC4 45	OCC2 BEN2 OCC3 POCC2 21	AGE5 AGE4 DUR3 22	AGE3 BEN3 BEN4 EDU3 40	AGE2, DUR2 OCC1, EXIT2 REG1 34	AGE1, BEN1 EDU1, EDU2 DUR1,POCC1 EXIT1, REG2 42

Fig 4: One-dimensional Kohonen map, modalities and individuals.

We immediately see that the classes are ordered. According to the proportion of cumulated duration of occasional work, from left to right, according to the age. We see that contrarily to a very common idea, the duration of the unemployment is not associated to the daily benefit, but with the age.
Class 1 and 2 comprise people highly involved in occasional work. The benefits obtained are slightly above the average. Class 3 and 4 are characterized by the very long duration of unemployment. People are older than the average with benefits from unemployment slightly above the average. Most of them are not exerting any occasional work. Class 5 is constituted with young people, with no occasional work, and a short seniority in past employment. Their situation seems a good illustration of a typical trajectory with successive periods of fixed-term contracts and unemployment. Class 6 is made up of very young people, having a very short seniority, a duration of unemployment lower than the whole population mean, an average amount of unemployment benefits close to 0. Most of them are still looking

for their first job. They leave unemployment by finding a job. We retrieve in a very easy way the main conclusions of the complete study, (Cottrell and Gaubert, 2000).

5. Conclusion

It is possible to extend the basic SOM techniques to many other frames, like the study of individuals described by qualitative variable. Il is also very important to keep in mind that the classical and Kohonen-based methodologies can be mixed. For example, it is possible to previously realize a classification of the observations based on quantitative variables, and then to use the number of the class as a new qualitative variable which is added to the other ones, in order to apply the KDISJ algorithm. Conversely the KDISJ algorithm can be applied to the qualitative variables to get a partition of the data, in order to look for a specific model for each class. It is each day more and more evident that the Kohonen-based methods are a part of the numerous tools that the statistician have at his disposal to analyze, represent, visualize data.

6. References

Benzécri J.P., *L'analyse des données, T2, l'analyse des correspondances,* Dunod, Paris, 1973.

Cottrell M., Letrémy P., Roy E., Analysing a contingency table with Kohonen maps: a Factorial Correspondence Analysis, *Proc. IWANN'93*, J. Cabestany, J. Mary, A. Prieto Eds., Lecture Notes in Computer Science, Springer-Verlag, 305-311, 1993.

Cottrell M., Ibbou S., Multiple correspondence analysis of a cross-tabulation matrix using the Kohonen algorithm, *Proc. ESANN'95*, M. Verleysen Ed., D Facto, Bruxelles, 27-32, 1995.

Cottrell M. and Rousset P., The Kohonen algorithm: a powerful tool for analysing and representing multidimensional quantitative et qualitative data, *Proc. IWANN'97*, Lanzarote, Springer, 861-871, 1997.

Cottrell M., Fort J.C., Pagès G., Theoretical aspects of the SOM Algorithm, *Neurocomputing,* **21**, 119-138, 1998.

Cottrell M., Gaubert P., Letrémy P., Rousset P., Analyzing and representing multidimensional quantitative and qualitative data: Demographic study of the Rhöne valley. The domestic consumption of the Canadian families, *WSOM'99,* In: Oja E., Kaski S. (Eds), *Kohonen Maps*, Elsevier, Amsterdam, 1-14, 1999.

Cottrell M., Gaubert P., Classification of recurring unemployed workers and unemployment exits, *Conf. ACSEG 1999*, *EJESS*, 14(1), 59-68, 2000.

Cottrell M. and Letrémy P., Working times in atypical forms of employment : the special case of part-time work, *Conf. ACSEG 2001*, Rennes, 2001, to appear, Kluwer, 2003.

Ibbou S., Classification, analyse des correspondances et méthodes neuronales, Doctoral thesis, Université Paris 1, 1998.

Kohonen T, *Self-Organizing Maps*, Springer Series in Information Sciences Vol 30, Springer, 1995.

Lebart L., Morineau A., Warwick K.M., *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices,* Wiley, 1984.

Letrémy P., Cottrell M., Macaire S., Meilland C., Michon F., Le temps de travail des formes particulières d'emploi, Rapport final, IRES, Noisy-le-Grand, February 2001, *Economie et Statistique*, Octobre 2002.