Dynamic visualization of changes in association patterns

Palumbo, Francesco
Universià degli Studi di Napoli Federico II
Dipartimento di teoria e metodi per la ricerca umana e sociale,
via Porta di Massa, Napoli 80100, Italia
E-mail: fpalumbo@unina.it

Iodice D'Enza, Alfonso
Università di Cassino
Dipartimento di Scienze Economiche,
via S. Angelo a Folcara, Cassino 03043, Italia
E-mail: iodicede@unicas.it

Introduction

In his paper *The Analysis of Multivariate Binary Data* (1972) Cox foresaw the relevance of binary data analysis. Dealing with binary attributes, Cox pointed out, the aim of the analysis had to be the study of multiple association structures of the attributes rather than the one-to-one or one-to-many dependences. Cox's intuition turns out to be strictly relevant still nowadays: that is, the increasing attention towards binary data coding characterized the last decade. Such tendency depends on several reasons, some of them overcome what previewed by Sir Cox. Binary data are the most basic digital data coding and storing, and they represent the most straightforward coding to automatically collect and store information about studied phenomena. In terms of computational costs, binary coding of data ensures fast, flexible and low memory storage requirements. A binary data set consists of a collection of binary sequences, each one arranged in a p-dimensional binary row-vector, being p the number of considered attributes. Applications involving binary data structures represent a very large set: Market Basket Analysis (MBA), Web Mining, Microarray Data Analysis are some classical domains. Text Mining and Image Analysis are further frameworks of application dealing with binary data structures. In all of these contexts, the number of binary sequences tends usually to be very large.

More recent continuos monitoring systems permit to collect data without interruption over the time, so that in modern applications it is fair to consider binary data as split in several batches. However, the stratification in batches of a binary data mart can be also due to other different reasons: the amount of data in question is too large to be analyzed as a whole; the data batches refer to different occasions in time and/or space. In this framework the analysis aims to monitor and visualize possible changes in the association structure characterizing the considered data batches.

The present proposal deals with high-dimensional binary data collected in different occasions in time or space. Studying the associations of data collected at different occasions, a primary aim is to detect changes in the association structure from one occasion to another. A suitable exploratory technique for the analysis of multiple associations in high-dimensional data is the multiple correspondence analysis (MCA; Greenacre, 2007). However, the comparison of MCA factorial displays referring to different occasions is meaningless. A possible solution to link the association structures of different data batches is to start from an MCA display of a reference and incrementally update the solution with further batches (Iodice D'Enza and Greenacre, 2010). In case of sparse data, the co-occurrence of groups of binary attributes suggests a latent statistical unit-wise cluster structure.

Then the present proposal takes into account the cluster-structure of the units in order to monitor, describe and visualize the evolving association among the p attributes. The strategy consists in determining a latent categorical variable that assigns each statistical unit to a cluster. Such latent variable describes the cluster structure underlying the statistical units and it is updated at each subsequent data batch: in other words this variable is determined according to the association structure and represents the 'link' over the analyzed data batches.

The strategy is implemented through the combination of clustering and factorial techniques: the clustering phase leads to define the latent variable; the factorial technique leads to synthesize and visualize the multiple association structure of the attributes, given the latent variable. The analysis consistency of the facts that both the implemented steps satisfy the same statistical criterion. The combination of clustering and factorial techniques has been already proposed in the literature, yet with different specific aims. Vichi and Kiers (2001) propose a combination of principal component analysis (PCA) with k-means clustering method. In the framework of categorical data, another interesting approach combining clustering and multiple correspondence analysis (MCA) is proposed by Hwang et al. (2006). Similarly, yet dealing with binary data, Palumbo and Iodice D'Enza (2010) propose a suitable dimension reduction and clustering. However all the above proposals suitably apply to static data sets: they are based on iterative procedures and hence all the data being analyzed has to be processed at once. The proposal in question aims to a non-iterative comparative association analysis of subsequent binary data batches.
The paper is structured as follows: in section 2 the problem is described and formalized;

2 Problem statement

In this section we provide a description of the proposed approach to study the structure of associations in binary high-dimensional data.
Let n and p be respectively the number of statistical units and the number of binary attributes Z_j $(j = 1,\ldots,p)$; let K be the number of latent groups of statistical units. Assigning a single statistical unit to one of K groups is a single occurrence of a multinomial experiment with K possible outcomes. The group k, $k = 1,\ldots,K$, is coded via the indicator variable I_k, where $I_k = 1$ if the statistical unit is assigned to the k^{th} group, $I_k = 0$ otherwise. Considering n trials, the random vector $X = (X_1, X_2, \ldots, X_K)$ follows a multinomial distribution with parameters $(n; \pi_1, \pi_2, \ldots, \pi_K)$, with $\pi_k = Pr(I_k = 1)$: n is the only known parameter. Assuming the parameters $(\pi_1, \pi_2, \ldots, \pi_K)$ to be given, the optimal criterion for the allocation of the n statistical units into the K groups is to randomly assign units to groups proportionally to the corresponding π_k parameters.
In this contribution the aim is to estimate π_k (that is to optimally assign units to groups) in order to maximize the heterogeneity among groups with respect to the Z_j attributes. Each of the attributes Z_j is Bernoulli distributed (with z indicating success and $\bar{z}$ failure) distributed with parameter π_Z. According to the same criterion, new statistical units are processed in order to update both the clustering solution and the binary attribute quantifications.

The groups heterogeneity corresponds to the qualitative variance between the K levels of the X variable, then with n statistical units described by p binary attributes $Z_1, Z_2, \ldots, Z_j, \ldots, Z_p$, the quantity to maximize is

$$\sum_{j=1}^{p} (G(X) - G(X \mid Z_j)). \tag{1}$$

This expression represents the sum of variances explained by each of the attributes Z_j. Consider X as a quali-

tative variable with $k = 1, \ldots, K$ categories and Z a binary variable with attributes $\{z, \bar{z}\}$.

		Z		
		z	$\bar{z}$	
X	1	f_{11}	f_{12}	f_{1+}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	K	f_{K1}	f_{K2}	f_{K+}
		f_{+1}	f_{+2}	n

Let $\mathbf{F}$ be the cross-classification table (represented above) with general element f_{kh} being the co-occurence number of the categories k and h, with $h = 1, 2$; the row margin f_{k+} is the number of occurrences of the category k and the column margin f_{+h} is the number of occurrences of the category h, with $f_{++} = n$ being the grand total of the table.
The variation of X explained by the categories of Z is

$$\begin{aligned} G(X) - G(X \mid Z) &= 1 - \sum_{k=1}^{K} \frac{f_{k+}^2}{n^2} - \left(1 - \frac{1}{n} \sum_{k=1}^{K} \sum_{h=1}^{2} \frac{f_{kh}^2}{f_{+h}}\right) = \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{h=1}^{2} \frac{f_{kh}^2}{f_{+h}} - \frac{1}{n} \sum_{k=1}^{K} \frac{f_{k+}^2}{n} \end{aligned} \tag{2}$$

It is worth noting that the quantity in equation 2 can also be expressed in terms of proportional prediction that refers to the situation when statistical units are randomly assigned to the group k with probability $\frac{f_{k+}}{n}$. When additional information is provided by a variabie Z, the proportional prediction becomes $\frac{f_{kh}}{f_{+h}}$, $h = 1, 2$. The average proportion of correct prediction with and without additional information is $\sum_{k=1}^{K} \frac{f_{k+}^2}{n^2}$ and $\sum_{k=1}^{K} \sum_{h=1}^{2} \frac{f_{kh}^2}{f_{+h}}$, respectively (Mirkin, 2001). The generalization of expression 2 to the $p-$attributes case gives the expression 1.

3 The procedure

This section illustrates the procedure to alternatively determine the factorial structure that better synthesizes the multiple associations among attributes and the latent variable that links the solutions of each data batch. It keeps the between groups heterogeneity maximized as new data batches are analysed; the procedure runs over the following steps:

1. determination of the latent variable X observed on the starting data batch: it is obtained by performing a K-means clustering of the starting data;
2. determination of the factorial synthesis of the association structure given the X variable;
3. update of X according to new data.

The latter two phases are repeated for each new data batch to analyse. The following algebraic notation is adopted:

- $\mathbf{Z}$ $(n \times 2p)$ disjunctive coded binary data matrix, with two columns per attribute (presence-absence);
- $\mathbf{z}_j$ $(j = 1, \ldots, 2p)$ the general column vector of $\mathbf{Z}$.

- $\mathbf{X}$ $(n \times K)$ matrix that assigns each statistical unit to one of the K modalities of X.
- $\mathbb{F} = \mathbf{X}^\mathsf{T}\mathbf{Z} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_j, \ldots, \mathbf{F}_p]$ is a block matrix, the j^{th} block corresponds to the cross-tabulation of the categorical variable X with the attribute Z_j, it is the same as table previously described.

The $\mathbf{X}$ matrix is defined by performing a K-means on $\mathbf{Z}$. The otho-normal basis $\mathbf{U}$ of the factorial sub-space of the multiple association structure is determined by maximizing the criterion in expression 1, which is in algebraic form is

$$\begin{aligned} (3) \qquad & \operatorname{tr}\left[\mathbb{F}(\Delta)^{-1}\mathbb{F}^\mathsf{T} - \tfrac{p}{n^2}\left(\mathbf{X}^\mathsf{T}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{X}\right)\right] \equiv \\ \equiv \quad & \operatorname{tr}\left[\mathbf{X}^\mathsf{T}\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^\mathsf{T}\mathbf{X} - \tfrac{p}{n^2}\left(\mathbf{X}^\mathsf{T}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{X}\right)\right] \end{aligned}$$

where $\Delta = diag(\mathbf{Z}^\mathsf{T}\mathbf{Z})$ and $\mathbf{1}$ is a n-dimensional vector of ones.
The solution to the problem lies in maximizing the trace of the above matrix, and the least square solution consists in the eigen-analysis of

$$(4) \qquad \frac{1}{n}\left[\mathbf{X}^\mathsf{T}\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^\mathsf{T}\mathbf{X} - \frac{p}{n}\left(\mathbf{X}^\mathsf{T}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{X}\right)\right]\mathbf{U} = \Lambda\mathbf{U}.$$

The statistical unit coordinates on the factorial space are in the matrix Ψ, such that

$$(5) \qquad \Psi = \left(\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^\mathsf{T} - \frac{p}{n}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\mathbf{X}\mathbf{U}\Lambda^{\frac{1}{2}};$$

The centroid matrix $\mathbf{G}$ is such that the k^{th} row-vector contains the average coordinates of the statistical units belonging to the k^{th} group; in particular,

$$(6) \qquad \mathbf{G} = \mathbf{X}^\mathsf{T}\Psi\mathbf{C}$$

where $\mathbf{C} = \mathbf{c}^\mathsf{T}\mathbf{I}$: $\mathbf{c}$ is a K-dimensional vector with elements representing the reciprocal of the size of each group; $\mathbf{I}$ is a K-dimensional identity matrix.
Let $\mathbf{Z}^+$ be a $(n^+ \times 2p)$ matrix of a new data batch consisting of n^+ statistical units. The new data are projected on the factorial plan defined by the orthonormal basis $\mathbf{U}$ according to the following formula

$$(7) \qquad \Psi^+ = \left(\mathbf{Z}^+(\Delta)^{-1}\mathbf{Z}^\mathsf{T} - \frac{p}{n}\mathbf{1}^+\mathbf{1}^\mathsf{T}\right)\mathbf{X}\mathbf{U}\Lambda^{\frac{1}{2}}.$$

The data points in Ψ^+ are then assigned to the closest of the centroids in $\mathbf{G}$: then a $(n^+ \times K)$ allocation matrix $\mathbf{X}^+$ for the new data is obtained. The update process requires the orthonormal basis $\mathbf{U}$ to be updated, too. Thus, all the following quantities are updated according to the new available information:

- $n^* = n + n^+$ is the total number of statistical units;
- $\mathbb{F}^* = \mathbb{F} + \mathbb{F}^+$, with $\mathbb{F}^+ = \mathbf{Z}^{+\mathsf{T}}\mathbf{X}^\mathsf{T}$;
- $\Delta^* = \Delta^+ + \Delta$, where $\Delta^+ = diag(\mathbf{Z}^\mathsf{T}\mathbf{Z})$ is the diagonal matrix of the attributes occurrences.

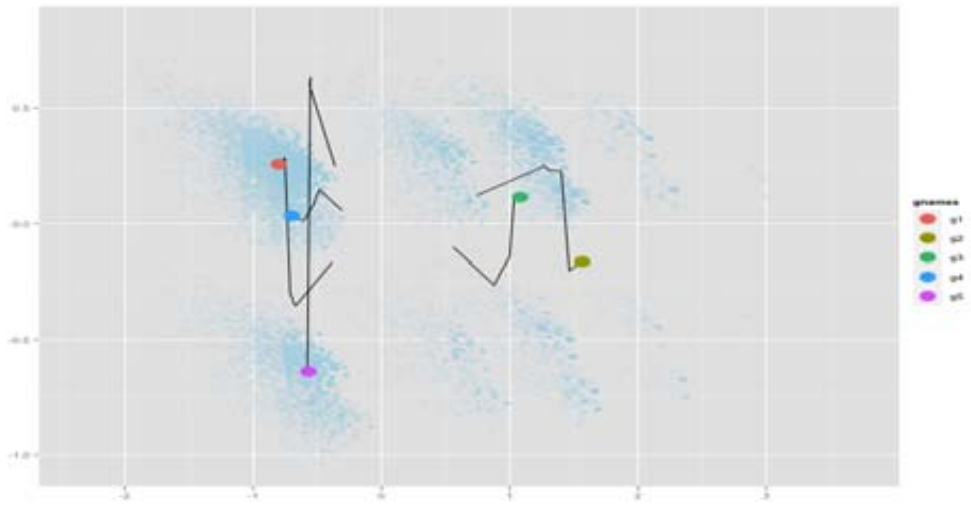

Figure 1: Statistical units and batch-wise centroid path

The updated orthonormal basis $\mathbf{U}^*$ is obtained via the eigen-analysis of the following quantity

$$\frac{1}{n^*}\left[\mathbb{F}^*(\Delta^*)^{-1}\mathbb{F}^{*\mathsf{T}} - \frac{p}{n^*}\left(\mathbf{f}^*\mathbf{f}^{*\mathsf{T}}\right)\right]\mathbf{U}^* = \Lambda^*\mathbf{U}^* \tag{8}$$

where $\mathbf{f}^*$ is the row-margin vector of the $\mathbb{F}^*$ matrix.

4 Example on real data

In this section the proposed procedure is applied to the 'retail' data set (Brijs *et al.*, 1999). The retail market basket data set is supplied by an anonymous Belgian retail supermarket store. The data are collected over three non-consecutive periods, for a time range of approximately 5 months of data. The total amount of receipts (statistical units) being collected equals $n = 88163$, whereas the number of products (binary attributes) $p = 28549$. The data set is very sparse, and the analysis aim is to study the association structures: in a pre-processing phase the attributes occurring in less than 1% of statistical units are discarded, as well as the receipts with less than three products. In summary, the analysis considers 60 attributes observed on 20000 statistical units. The first 4000 statistical units are the starting batch, then the solution is updated according to the further process. The remaining flow of 16000 units is progressively processed such that each upcoming batch is the 20% of the whole amount of analyzed data. Then the number of analyzed batches is eight. The proposed strategy leads to display the multiple association structure of attributes. Results are adaptively updated as new data batches are processed. For sake of space the choice of the number of groups underlying the statistical units K is not discussed and it is set $K = 5$ on the basis of previous studies. Figure 1 shows the final factorial display of statistical units. The five points represent the final position of the centroids; the path of each centroid along the batches is also represented.

To appreciate the changes in the attribute associations, refer to figure 2. The factorial representation in figure 2 represents a common visualization support of the attribute associations for each of the subsequent batch. In particular, such common support is obtained by performing a multi-way multidimensional scaling (Borg and Groenen, 2005) on the chi-square distances characterizing the attributes. Figure 2 displays the paths of the attribute points batch after batch. The longer the path, the larger the change in the association structure of the corresponding attribute. In order to increase the readability of the display we plotted only the 20% of the longest trajectories, so that the most changing attributes are highlighted.

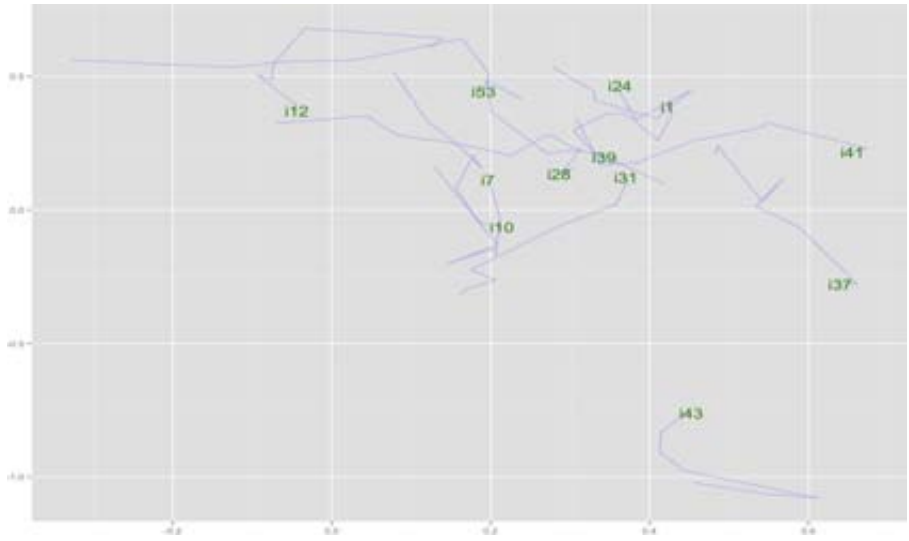

Figure 2: Attributes common sub-space batch-wise path

The results confirm that, in general, the buying behavior of customers does not radically change from a month to another. However, some przales do change over time and a common graphical display of the attributes (figure 2) turns out to be helpful in quickly identifying which attributes changed more. The proposed strategy leads to display the multiple association structure of attributes. Results are updated adaptively as new data batches are processed. Furthermore, the procedure does not require all the data batches are permanently stored in memory, but the last one.

REFERENCES

Borg I., and Groenen P., (2005) '*Modern multidimensional scaling*'. *Springer*.

Brijs T., Swinnen G., Vanhoof K. and Wets G., (1999) 'Using association rules for product assortment decisions: a case study'. *KDDM*. 254–260.

Greenacre M. J., (2007) '*Correspondence Analysis in Practice*', second edition.*Chapman and Hall/CR*.

Hwang H., Dillon W. R., and Takane Y., (2006) 'An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents'. *Psychometrika*. 71, 161–171.

Iodice D'Enza A., and Greenacre M.J., (2010) 'Multiple correspondence analysis for the quantification and visualization of large categorical data sets'. In proc. of SIS09 *Statistical Methods for the analysis of large data-sets*. (in press).

Iodice D'Enza, A., and Palumbo F., (2007) 'Binary data flow visualization on factorial maps'. *Revue Modulad*, 36.

Mirkin B., (2001) 'Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables'. *The American Statistician*, 55, 2:111-120.

Palumbo F., and Iodice D'Enza A., (2010). 'A two-step iterative procedure for clustering of binary sequences'. In *Data Analysis And Classification* F. Palumbo, CN Lauro and MJ Greenacre eds., Springer, 33–40, doi: 10.1007/978-3-642-03739-9_4

Vichi M. and Kiers H., (2001) 'Factorial k-means analysis for two way data'. *Computational Statistics and Data Analysis* 37(1): 49–64.