Binary data flow visualization on factorial axes

Alfonso Iodice D'Enza[1], Francesco Palumbo[2]

[1] Dipartimento di Matematica e Statistica
Universitá di Napoli Federico II
Complesso Universitario di Monte S. Angelo, via Cinthia
I-80126, Napoli, Italy
iodicede@unina.it
[2] Dipartimento di Istituzioni Economiche e Finanziarie
Universitá di Macerata
via Crescimbeni, 20
I-62100, Macerata, Italy
palumbo@unimc.it

Abstract. *Data streams* are one of the most relevant new data sources, they refer to flows of data that come at a very high rate. Let us consider a stock-exchange market, where n different stocks with p considered attributes (e.g. price, quantity, seller/buyer id, ...) are negotiated all day long. The distinguishing feature in data streams analysis is that the focus is on *transient relations*. The present paper proposes a visualization tool exploiting Multidimensional Data Analisis (MDA) techniques to represent the evolving association structures among attributes over different time-frames. The general aim is to detect the stability of the deviation from indipendence in the occurrence of an observed set of attributes stored as binary stream.

1 Introduction

In recent years, enhancements in monitoring activities and collecting data determined the need for a different approach in knowledge extraction: new data are produced at a faster rate than the capability of analyzing them.
Information mining through traditional data mining systems becomes often inadequate. *Data streams* are one of the most relevant new data sources, they refer to flows of data that come at a very high rate. Let us consider a stock-exchange market, where n different stocks are negotiated: at each interval time-unit (seconds or minutes) a $n \times p$ array is added to the database, with p indicating the number of considered attributes (e.g. price, quantity, seller/buyer id, ...). These features make data streams leading to data structure unusual in the data analysis and statistical data mining framework. New and more appropriated techniques should be taken into account to usefully extract knowledge without storing the data for a long term [12].

The most relevant changing feature in data mining systems dealing with high-speed data streams is the necessity of analyzing data in a single pass: iterative procedures will lead to unfeasible solutions. Further ideal features of a data stream mining system are described in the proposal by [4].

Data streams mining, in a wide sense, can be considered an evolution of data mining. The development of data mining techniques has been feeded in the last decade by statisticians and computer scientists. Similarly, the data stream analysis should be enhanced with the contribution of researchers from different areas.

Data stream analysis techniques can be roughly divided into: *i*) data-based, indicating techniques aiming to summarize or reduce the amount of streams to be analyzed; *ii*) task-based techniques, facing the crucial problem of adapting existing algorithms to the new computational costs;

iii) mining techniques, the properly defined knowledge extraction techniques [5].
The present proposal aims at defining a dynamic association mining technique focusing on binary data streams.

Binary strings recording the presence/absence of a set of attributes (*items*) represent a very basic data structure characterizing a wide variety of applications. In particular the general aim of the paper is to monitor the co-occurrence relations among items through the visualization of the association structures on factorial maps.

Different contributions ([10],[8]) have shown as the multidimensional data analysis (MDA) graphical output can be very useful as exploratory pre-processing tool. In fact, graphical data display ease to discover well defined patterns of items highly co-occurring in the data. Note that *well defined* stands for patterns that strongly characterize even small subsets of the considered streams. We consider a set of streams to be a binary $n \times p$ matrix of n records storing the presence absence of to p attributes. Let us consider a set of streams that we assume as the *normal*, *starting* situation: these are used to define a graphical support where to plot streams as they arrive. Looking at the streams positions on the data plot, we are able to appreciate the system changes.

Firstly, we refer to a starting static data structure (regarding a fixed number of time-frames). Then new streams are processed with respect to the starting structure.

A factorial display of the co-occurrence structure of the starting data is obtained; on such display new streams are represented as supplementary information. In this way it is possible to detect the changes over time in the co-occurrence structure characterizing the considered items or attributes: representing the trajectories of the item/points on the factorial display it is possible to observe the evolutionary relations. Of course, the factorial display can be updated periodically, in order to privilege recent data.

The MDA technique exploited by the procedure is multiple correspondence analysis (MCA).

In addition, to take into account more than two axes the process, we propose to cluster the items according to their coordinates on the orthogonal factorial map. This taks is achieved through an agglomerative clustering on the items coordinates.

The aim of the application of MCA is two-folded: to visualize the co-occurrence structure characterizing items and to remove noise and redundancies in the structure underlying data. The items position change according to new data, but always with respect to the starting, reference situation.

Points/items are projected on the factorial map in different colours according to the cluster they belong to: if the items in different colors result well-separeated on the factorial map, then the first two dimensions discriminate the groups; if the items in different colors result close, this means that their distance depends mostly on the third and fourth factor. The paper is structured as follows: in section 2 basic notation and definition are provided; the section ends with a brief overview of the related literature; section 3 contains the description of the steps of the proposed strategy; the following section shows an application of the proposed strategy to synthetic data; the paper ends with a section of conclusion and perspectives.

2 Data structures and related work

The starting data structure $\mathbf{Z}$ is a $(n \times p)$ presence/absence matrix characterized by n binary vectors considered with respect to p Boolean variables (e.g.: *presence* or *absence* of an attribute). The MCA is then applied on $\tilde{\mathbf{Z}}$ that is the disjunctive coded version of $\mathbf{Z}$. The cells of the matrix $\mathbf{B} = \tilde{\mathbf{Z}}^{\mathsf{T}}\tilde{\mathbf{Z}}$ represent the co-occurrence of each possible pair of items. The correspondence analysis of $\mathbf{B}$ determines the starting support dysplay. Let us indicate with $\tilde{\mathbf{Z}}_k$ the indicator matrix of the stream produced in the k^{th} time-frame. Then at each time-frame a Burt table $\mathbf{B}_k$ is obtained through

$$\mathbf{B}_k - \tilde{\mathbf{Z}}^{\mathsf{T}}\tilde{\mathbf{Z}}_k, \tag{1}$$

The correspondence analysis of $\mathbf{B}_k$ determines new coordinates of the items on the starting map: the starting configuration of points is updated according to the new streams.

We define $\mathbf{S} = n^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ to be the co-occurrence matrix of all the possible item pairs. S is a symmetric contingency table, whose general extra-diagonal term $s_{jj'} = s_{j'j}$ represents the degree of co-occurrence of the items (attributes) j and j'. Indexes j and j' vary in $1, \dots, p$. On the main diagonal of $\mathbf{S}$ there are the degree of occurrence of each single item.

The data matrices $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_K$ contain the new records produced in the considered K time-frames. For each $\mathbf{Z}_k$, the corresponding $\mathbf{S}_k$ matrix can be computed in the same way of $\mathbf{S}$.
The hierarchical clustering of the items projected in a reduced-dimensional space defines q item-set that will be plotted in different colors.
In the literature different proposals aim to explore the relationship structure characterizing a data set through the combination of clustering procedures and factorial techniques. An example of such combined approach is the *tandem analysis* proposed by [2]: following such approach, a principal correspondence analysis is first applyed on data and a clustering procedure is then performed on the object scores of a reduced number of components. Other procedures combining clustering with MDA techniques have been proposed: [14] propose a combination of principal component analysis (PCA) with K-means clustering method. Another really interesting approach combining clustering and multiple correspondence analysis, is proposed by [7].

3 Multidimensional analysis of binary streams

The presented strategy recalls the multi-phase approach proposed by [9] and [8] for identifying potentially interesting simple AR. The difference is in the nature of the studied relations: in the present paper we only focus on the inter-dependency structure characterizing items

The general task is to monitor and represent transient co-occurrence relations among considered items or attributes. The aim is then to identify association patterns resulting stable over time as well as itemsets with increasing/decreasing degree of co-occurrence.

The strategy implemented to obtain this task consists of different phases:

1. multiple correspondence analysis (MCA) of indicator matrix $\tilde{\mathbf{Z}}$ that is $n \times 2p$, being the disjunctive coded version of $\mathbf{Z}$: the n considered streams determine the *starting* (*normal*) association structure represented on the map
2. representation of items as supplementary points projected on the starting factorial map (see step 1)
3. after a user-defined number of time-frames, the step 1 is repeated and the starting map is updated

The multiple correspondence analysis (MCA) capabilities are exploited in order to synthesize and represent the association structure characterizing items. In [8] and [9] it has been used a simple CA on the support matrix $\mathbf{S} = n^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$. However the matrix $\mathbf{S}$ is square and the application of a simple CA may not be suitable to represent both items and transactions. Indeed, analyzing the matrix $\tilde{\mathbf{Z}}$ through an MCA does not determine crucial changes in the computations for searching the sub-space of approximation; anyway it represents a classical application of the MCA on a set of binary variables (items) with respect to a set of respondents (transactions).

In the following of this section we first provide a very basic definition of MCA; then we describe the items projection for the upcoming time-frames.
Multiple correspondence analysis can be interpreted as the application of CA to the Burt matrix

$$\mathbf{B} = \tilde{\mathbf{Z}}^{\mathsf{T}}\tilde{\mathbf{Z}} \tag{2}$$

[6]. The correspondence matrix $\mathbf{P}$ is obtained by dividing $\mathbf{B}$ by its *grand total*; vectors $\mathbf{r}$ and $\mathbf{c}$ are the same and they represent the row and column margins, respectively, of $\mathbf{P}$. The reduced rank

matrix approximation of $\mathbf{P}$ can be obtained by performing a singular value decomposition of the matrix $\mathbf{Q}$, whose general element is

$$\mathbf{Q} = \{q_{ij}\} = \frac{(p_{ij} - r_i r_j)}{\sqrt{r_i r_j}}. \tag{3}$$

Then the singular value decomposition of $\mathbf{Q}$

$$\mathbf{Q} - \mathbf{U}\Lambda\mathbf{U}^{\mathsf{T}}, \tag{4}$$

determines eigenvector matrix $\mathbf{U}$, with general element $u_{i,s}$, with $i = 1, \ldots, 2p$ -since we operate on $\tilde{\mathbf{Z}}$- and the eigenvalue diagonal matrix Λ that has general element λ_s. The *principal co-ordinate* of the i^{th} row (column) point on the s^{th} dimension is obtained through

$$f_{is} = a_{is}\lambda_s, \tag{5}$$

with a_{is} being the corresponding *standard co-ordinate*, that is $a_{is} - \frac{u_{is}}{\sqrt{r_i}}$.
The variation of the frequency table $\mathbf{P}$ is measured by the *total inertia*, the weighted sum of squares of the centered matrix that is being approximated:

$$total\ inertia = \sum_i \sum_j \frac{(p_{ij} - r_i r_j)^2}{r_i r_j}$$

The inertia explained by the first q dimensions is given by $\sum_{s=1}^{q} \lambda_s$. As a consequence, the proportion of inertia explained by each considered factor s is given by $\lambda_s / \sum_{s=1}^{q} \lambda_s$. However, such quantity is a pessimistic measure of the explanatory power of the factors: in MCA it is common to observe low percentages of inertia even on the first factors. This is due to the disjunctive coding that force orthogonality among the modalities of a same variable [11]. In order to limit the spherical shape due to the disjunctive coding, the inertia of the first factors is adjusted, according to [3], in the following way:

$$\lambda^* = \left(\frac{p}{p-1}\right)^2 \cdot \left(\lambda - \frac{1}{p}\right)^2, \ \lambda > \frac{1}{p}.$$

We use the results obtained through the decomposition of $\mathbf{Q}$ (see formula 4) to obtain a factorial display of items for the binary streams. We project as supplemantary information the items related to each new time-frame . In particular, we consider for each new time-frame a Burt table $\mathbf{B}_k$, $1 \leq k \leq K$, obtained through

$$\mathbf{B}_k - \tilde{\mathbf{Z}}^{\mathsf{T}}\tilde{\mathbf{Z}}_k, \tag{6}$$

with $\tilde{\mathbf{Z}}_k$ containing on the rows the new records produced in the considered time-frame.

The hierarchical clustering is performed on the principal coordinates f_{is} of the items. The cutoff point of the hierarchy determines H groups of items which are close on the MCA-based global factorial map.
For each timeframe, we define f^*_{is} to be the principal coordinate of the i^{th} item of the s^{th} dimension, resulting from the MCA of the matrix B_k (see equation 6). Each item belongs to one of the H groups: it is represented in a different color according to the group it belongs to. A further remarkable aspect is that the hierarchy is determined taking into account q dimensions of the factorial space ($q > 3$): it is then possible to find on the factorial map very close points represented in different colors. The interpretation of such a situation is that the two points present larger differences in the coordinates on the dimensions following the first two.

4 Example

In this section the visualization strategy is applied to a synthetic data flow in order to show the suitability of the procedure in detecting changes in association patterns. The choice of using synthetic data in spite of real data is due to the need of keeping under control the characteristic of the analyzed data: the aim is to check the effectiveness of visualization in detecting changes in the starting association structure of the upcoming streams .

The simulation scheme we used to generate synthetic streams consists of the following steps:

1. generate a set of $n = 500$ binary records with respect to $p = 32$ attributes;
2. generate T_i, $i = 1, \dots, 4$, different data set $(n \times p)$ to be treated as consecutive time-frames with 'stable-association';
3. generate T_i, $i = 5, \dots, 8$, different data set to be treated as consecutive time-frames with 'unstable-association'.

The considered normal/starting situation is characterized by a well defined association structure: there are two groups of highly co-occurring items, a first pattern of twenty-four items, a second pattern involving the remaining eight items. The simulated incoming streams of the first four time frames are characterized by the same association structure of the starting situation. The following time frames $T_5, \dots, T_8$ present changing association features; in particular, in T_5 two items of the first pattern are replaced by two items of the second pattern and *vice-versa*. Similarly in the following blocks of streams (T_6, T_7, T_8) the number of exchanged patterns grows up to 4, 6 and 8.

Of course, things are not so well defined when dealing with real data, but a well defined situation makes easier to illustrate the MDA visualization of binary streams.

The first graphical output of the procedure consists of a factorial representation of the starting association structure and a dendrogram representing the agglomerative clustering of the starting configuration of points/items: clustering is performed with respect to the first q factors, with q beeing used defined.

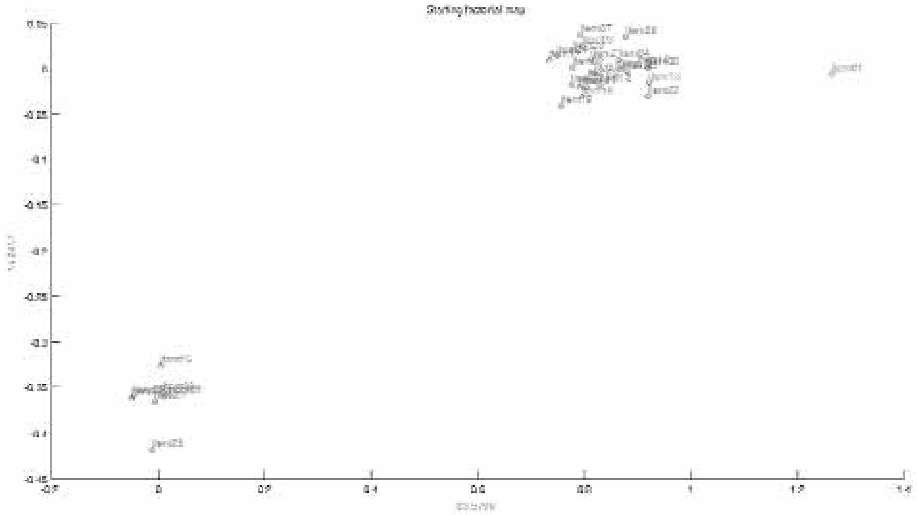

Fig. 1. *Factorial display of the starting situation*

Figure 1 represent the starting association structure and the *inertia* of the map explaine the 97.62% of the total variability of the starting data structure: this is due to the well defined association structure charcterizing the synthetic data.

The dendrogram in figure 2 clearly shows the two group structure of the starting data. The first 24 four items are in the first group, the remaing items in the second group.

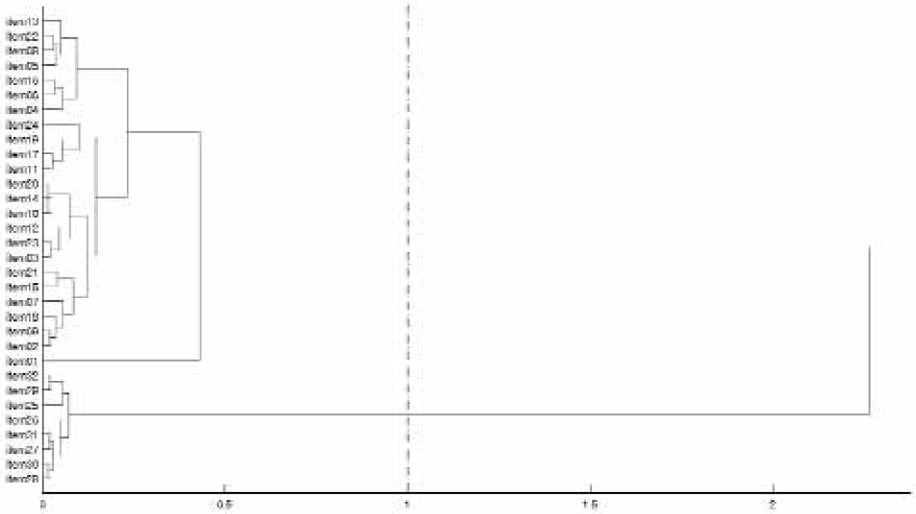

Fig. 2. *Dendrogram representation of the points/ items in four-dimensional space*

The evolving association structure of the eight time-frames is represented in figure 3. The windows on the first row refer to time-frames with data characterized by a stable associatin structure, that is the same of the starting structure. In the time-frame T_5 the two exchanged items deviate from the starting group and they are highlighted on the map.The following windows represent the evolving association structure characterizing items in the time-frames T_6, T_7, T_8. In the last frame the two group structure is recovered, but the items within the groups are different from the starting situation. This is a small example to show how the procedure works, however it seems to be promising in monitoring the association structures in the attributes of a process, of a stock market title or on images provided by a geostationary satellite. The detection of deviations from a normal structure is a common tasks of monitoring activities, then a visualization tool of the evolving structures is very suitable as it ease the user interaction.

5 Conclusion and perspectives

The analysis of flows of information available as data streams is a challenging feature for data miners: it requires fast computations and highly interpretable recults. The user needs to detect transient evolutionary relations, visualization of structures underlying data is then a suitable solution to ease and speed up the user interpretation of the results. MDA techniques represent a relevant tool for dimensionality reduction and data visualization. The present proposal seems to be promising in detecting transient deviations from a reference situation: this can be visually represented on factorial display, but it can even be automated by defining a rule that states if the system is stable or not. A possible threshold can be set according to the distance between two configurations of points of consecutive time-frames. The computational costs of the procedure are not particularly relevant. The starting set of p attributes is considered just to obtain the reference sub-space: upcoming time-frames are processed according to this lower dimensional reference space. The exploitation of MDA techniques for data streams analysis is, in conclusion, a way to obtain a visualization support and increase the user/process integration: the addition of some interactive elements in the representations is a crucial task to ensure an on-the-fly analysis of the data flow.

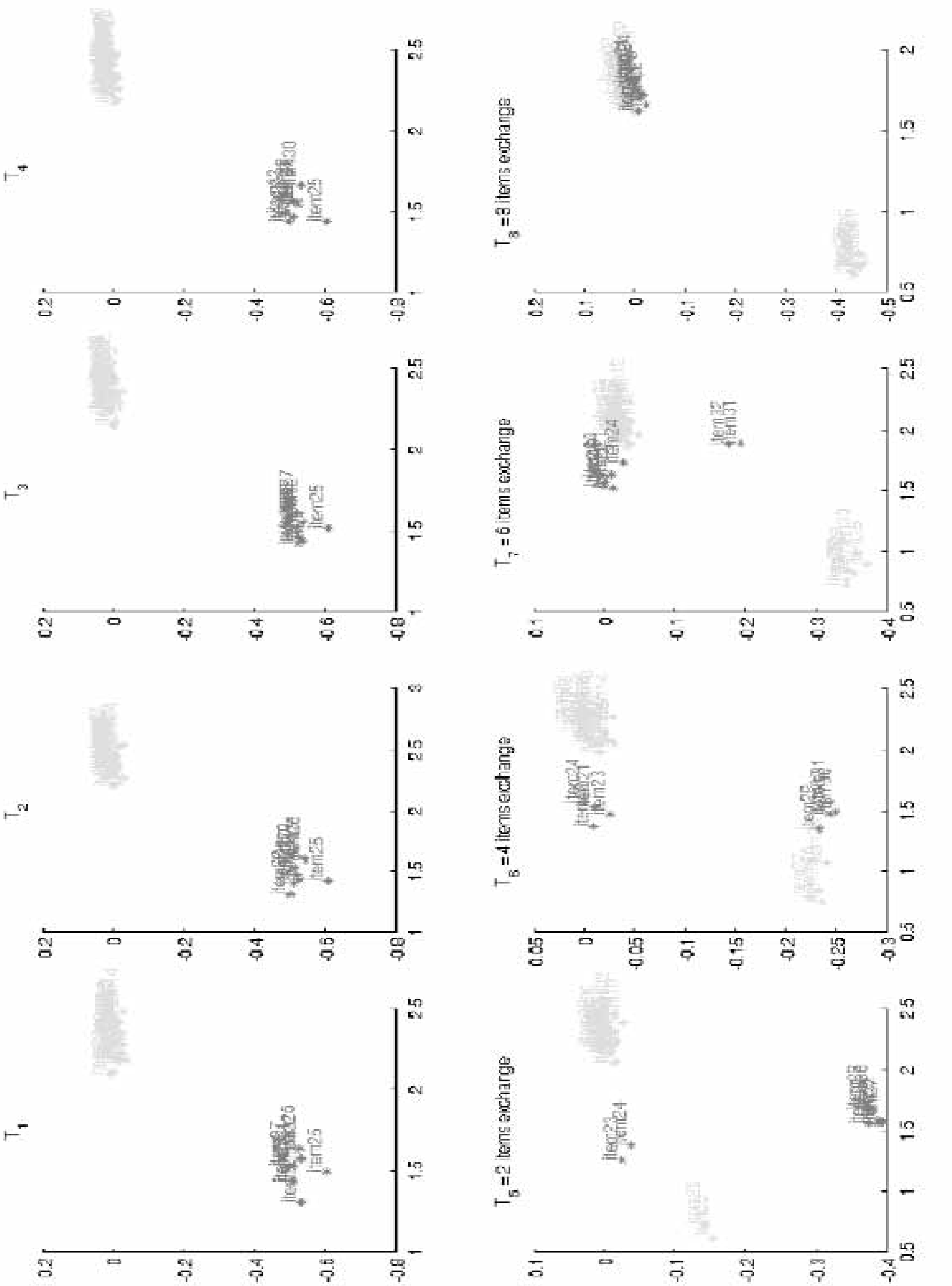

Fig. 3. *Mosaic representation of the evolving association structure of the items*

The adaption of different MDA techniques for data streams visualization and the introduction of interactive tools in the procedure are the future enhancements of the proposed approach.

References

1. Agrawal, R., Imielinski, T., and Swami, A.: 1993, Mining association rules between sets of items in large databases, *in* P. Buneman and S. Jajodia (eds), *ACM SIGMOD international conference on Management of data*, Vol. 22, ACM Press, Washington, D.C., pp. 207–216.
2. Arabie, P. and Hubert, L.: 1974, Cluster analysis in marketing research, *IEEE Trans. on Automatic Control* **AC–19**, 716–723.
3. Benzécri, J. P.: 1979, Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, *Le Cahiers de l'Analyse des Données* **4**(4), 377–378.
4. Domingos, P. and Hulten, G.: 2003, A general framework for mining massive data stream, citeseer.ist.psu.edu/domingos03general.html.
5. Gaber, M. M., Zaslavsky, A. and Krishnaswamy, S.: 2007, *Data Streams: Models and Algorithms*, Springer Verlag, chapter A Survey of Classification Methods in Data Streams.
6. Greenacre, M. J.: 1984, *Theory and Application of Correspondence Analysis*, Academic Press, London.
7. Hwang, H., Dillon, W. R. and Takane, Y.: in press, An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents, *Psychometrika* .
8. Iodice D'Enza, A.: 2006, *Exploratory Study of Association in Transaction Data Bases*, PhD thesis, Dipartimento di Matematica e Statistica Universitá degli Studi di Napoli Federico II, Napoli.
9. Iodice D'Enza, A., Palumbo, F. and Greenacre, M.: 2005, Exploratory data analysis leading towards the most interesting binary association rules, *Proceedings ASMDA 2005 Conference*, Brest, France.
10. Iodice D'Enza, A., Palumbo, F. and Greenacre, M.: 2006, Exploratory data analysis leading towards the most interesting simple association rules, *Computational Statistics and Data Analysis* **Accepted, in press.**
11. Lebart, L., Morineau, A. and Piron, M.: 1995, *Statistique exploratorie multidimensionelle*, Dunod, Paris.
12. Muthukrishnan, S.: 2003, Data streams: algorithms and applications, ACM-SIAM Symposium on Discrete Algorithms. citeseer.ist.psu.edu/article/muthukrishnan03data.html.
13. Plasse, M., Niang, N., Saporta, G. and Gauthier, D.: 2005, Combined use of association rules mining and clustering methods, *3rd world conference on Computational Statistics and Data Analysis*, Limassol, Cyprus.
14. Vichi, M. and Kiers, H.: 2001, Factorial k-means analysis for two way data, *Computational Statistics and Data Analysis* (37), 29–64.