Factor PD-Co-clustering in Official Statistics

Marina Marino, Germana Scepi, and Cristina Tortora

Abstract In this paper we propose an extension of Factor PD-clustering for simultaneous classification of rows and columns of frequency matrices extracted from large textual datasets. The aim is to extract information from documents usually produced and that are not used because of their special nature. The work is carried out within the European project BLUE-ETS, which aims to provide tools for the construction of robust and high quality official statistics for businesses.

Key words: Two-mode clustering, Factor PD-clustering, official statistics

1 Introduction

This work is part of European project BLUE-ETS[1], it aims at providing tools for producing high quality and robust statistical information in official business statistics. Our starting point is the development of quantitative tools for extracting useful information on firms activities by mining into documents produced by the firms. We consider textual data as a source of information "naturally available" and not sufficiently explored. In literature, some authors ([2], [1]) suggested to employ co-clustering when dealing with sparse and high-dimensional data with the aim of extracting knowledge from large database.
In this paper we propose to extend Factor PD-clustering (FPDC) algorithm [7] to a two-mode clustering. In particular, we compare the performance of this algorithm

Marina Marino
Dip. di Matematica e Statistica, Università di Napoli Federico II e-mail: marina.marino@unina.it

Germana Scepi
Dip. di Matematica e Statistica, Università di Napoli Federico II e-mail: scepi@unina.it

Cristina Tortora
Stazione Zoologia Anton Dhorn, Napoli e-mail: cristina.tortora@unina.it

[1] This paper is financially supported by European project BLUE-ETS

with the co-clustering algorithm proposed by Dhillon and others in 2003 [2] on a management commentaries dataset in order to detect the most suitable approach to identify interesting sub-structure.

2 Factor PD-clustering

PD-clustering is based on the assumption that the product between the probability of any point to belong to each class and the distance from the centers of the clusters is a constant, called Join Distance Function (JDF), depending on the point. The value of this constant is a measure of the classifiability of the point. The aim of the algorithm is to maximize the classifiability of all points that is equivalent to minimize the value of the JDF [3]. Formalizing, being given some random centers, the probability of any point to belong to each class is assumed to be inversely proportional to the distance from the centers of the clusters. Being given an X data matrix with n units and J variables, given K clusters that are assumed not empty, PD-Clustering is based on two quantities: the distance of each data point x_i from the K clusters centres c_k, $d_k(x_i)$, and the probabilities for each point to belong to a cluster, p_{ik} with $i = 1,\dots,n$ and $k = 1,\dots,K$. The relation between them is the basic assumption of the method: $p_{ik}d_k(x_i) = JDF(x_i)$, for a given value of x_i and for all $k = 1,\dots,K$. The quantity $JDF(x_i)$ is a measure of the closeness of x_i from all clusters' centers. The whole clustering problem consists in the identification of the centers that minimize the JDF. Basing on this assumption cluster centers can be computed as follow: $c_k = \sum_{i=1,\dots,N}\left(\frac{u_k(x_i)}{\sum_{j=1,\dots,N}u_k(x_j)}\right)x_i$, where $u_k(x_i) = \frac{p_{ik}^2}{d_k(x_i)}$ [3].
When the number of variables is large and variables are correlated non-hierarchical clustering methods, including PD-Clustering, become very unstable and the correlation between variables can hide the real number of clusters [9]. A linear transformation of the original variables into a reduced number of orthogonal ones can significantly improve the algorithm performance. The linear transformation of the variables and the clustering method must optimize a consistent criterion. In [6] it is demonstrated that a Tucker3 decomposition [4] of a distance matrix G minimize the JDF, or equivalently optimize the same criterion of PD-clustering. The matrix G is a distance matrix, of general elements $g_{ijk} = |x_{ij} - c_{kj}|$, where $i = 1,\dots,n$ indicates the units, $j = 1,\dots,J$ the variables and $k = 1,\dots,K$ the clusters. The method consists in finding the transformation of original data $x^*_{iq} = x_{ij}b_{jq}$, where b_{iq} is a weighting system, and cluster centers c_{kq} such that the JDF is minimized: $JDF^* = argmin_{C,B}\sum_{i=1}^{n}\sum_{q=1}^{Q}\sum_{k=1}^{K}(x^*_{iq} - c^*_{kq})^2 p_{ik}$, x^*_{iq} and c^*_{kj} indicate the projection of points and centers in the factorial space. The two quantities can not be found at the same time, so an iterative method is applied. At first a PD-clustering is applied on the original data matrix, starting on this first partition the distance matrix G is computed. On the distance matrix G a tucker 3 decompositions is applied in order to compute x^*_{iq}. On the transformed data PD-clustering is applied and the whole algorithm is iterate until the convergence is reached. The convergence of the method

is empirically demonstrated. The integration of the PD-Clustering and the Tucker3 factorial step makes the clustering more stable and permits to consider datasets with large number of variables and having not elliptical form [7].

3 Factor PD-Co-clustering

Dealing with textual data, it is meaningful to define simultaneously clusters of words and clusters of documents. Co-clustering or two-mode clustering are methods where the rows and columns of the data matrix are clustered simultaneously [8]. This implies two main advantages. The first is that there is an overall objective function that cannot be reduced to a simple combination of constituent row and column objective functions. The second is that a simultaneous clustering can emphasize the association between units and variables clusterings that appear as linked clusterings from the data analysis, and it allows the researcher to characterize the nature of the interaction or of the dependence structure between units and variables, as implied by the data. In literature, there are very different simultaneous clustering methods, in particular Information-theoretic Co-clustering, proposed by Dhillon and others in 2003, [2]. FPDC can be easily extended to a two-mode clustering. Tucker 3 decomposition allows to compute project of units x^* and variables y^* on the factorial space: $x^*_{iq} = x_{ij}b_{jq}$, $y^*_{jr} = y_{ji}u_{ir}$, with $i = 1,\dots,n$, $j = 1,\dots,J$, where b_{iq} and u_{ir} are a weighting system. The indices $q = 1,\dots,Q$, $r = 1,\dots,R$ refer to the number of components Q,R of respectively B,U. On the projection obtained, variables y^* can be grouped using PD-clustering. The factorial space has been obtained by a Tucker 3 decomposition of distance matrix G, it is the space that maximize the classifiability of units, variables clusters are affected by the units clusters. Starting from this clustering structure a distance matrix of variables G' can be computed. Basing on the same procedure clusters of units, on the space that maximize the classifiability of variables, can be obtained. The entire process is iterated until the convergence is reached. Two-mode FPDC can be summarized in the following steps: 1. Random initialization of cluster of units and computation of distance matrix G; 2. Three-way decomposition of distance matrix G; 3. Projection of units and variables on the factorial space; 4. PD-clustering of variables on the factorial space y^*; 5. Computation of distance matrix of variables G'; 6. Three-way decomposition of distance matrix G'; 7. Projection of units and variables on the factorial space; 8. PD-clustering of units on the factorial space x^*. Steps from 2 to 8 are iterated until convergence to the solution.

4 Management commentaries dataset

An important part of the official budget of listed companies is the management commentary. Among the 406 italian listed company in 2009, 25 companies have been

selected using a sample technique. The management commentaries of the selected companies have been processed using ad hoc techniques, interested readers can refer to [5]. The matrix obtained is a frequency matrix (25 companies on rows, 81 words on columns). We compare the results obtained applying on this matrix both the Information-theoretic Co-clustering algorithm, proposed by Dhillon, and the Factor PD-Co-clustering. Information-theoretic Co-clustering converge very fast and allows to obtain a block structure. Results can be interpreted as clusters of firms, of words and simultaneously. However, on this dataset, the method founds some clusters of few elements and one big cluster. Two-mode FPDC allows to obtain a block structure, clusters can be interpreted simultaneously or not and the number of elements in each cluster is balanced. The resulting clustering structure is made by four clusters. Clusters of words can be labeled as: positive aspects, adverbs-adjectives, market-results, financial aspects. Firms are grouped according to which groups of words have used. Words concerned with market and results are the most used. The first cluster of firms is composed by: KME, Mondadori, Prima industry, Recordati and York villa. These firms have the richest language, they use all groups words. Elica, Montefibre, Enel, Esprinet and Juventus, on the contrary have the poorest language, they use few words only concerning market and results. Mediaset, Effegi, Gas plus, Centrale del latte di Torino, Carraro, S.E.I and Mediacontech use many words concerning market and results and few adverbs and financial words. The last cluster, composed by: Azimut, Kinexia, Acque Potabili, RCF, Autostrade Torino, Isagro and Meridiana fly use few adverbs and some words concerning marketing and results.

References

[1] Balbi, S., Miele, R., Scepi, G. (2010). Clustering of documents from a two-way viewpoint. Bolasco S. *et al.* eds. *10th International Conference on Statistical Analysis of Textual Data*.

[2] Dhillon, I., Mallela, S., Modha, D. (2003). Information-theoretic co-clustering. , *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[3] Iyigun, C. (2007). *Probabilistic Distance Clustering*. Ph.D. thesis at, New Brunswick Rutgers, The State University of New Jersey.

[4] Kroonenberg, P. (2008). *Applied multiway data analysis*. Ebooks Corporation, Hoboken, New Jersey.

[5] Spano, M. Triunfo, N. (2012) La relazione sulla gestione delle società italiane quotate sul mercato regolamentato in *JADT 2012 : 11es Journées internationales d'Analyse statistique des Données Textuelles* (to appear)

[6] Tortora, C. (2011). *Non-hierarchical clustering methods on factorial subspaces*. PhD thesis, Università di Napoli Federico II.

[7] Tortora, C., Gettler Summa, M., Palumbo, F. (2011). Factorial pd-clustering. Submitted in *Proceedings of the Symposium of the German Classification Society*.

[8] Van Mechelen, I., Bock, H., Boeck, P. D. (2004). Two-mode clustering methods: a structured overview. *Statistical methods in medical research*, 13(5):363–394.

[9] Vichi, M. Kiers, H. (2001). Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis*, 37:29–64.