Binary attributes quantification with external information

Alfonso Iodice D'Enza[1*]

1. University of Cassino

* Contact author: iodicede@unina.it

Keywords: binary data, Non Symmetric Correspondence Analysis, association patterns

Binary data bases (DB) characterize several fields: transactional data, web-clickstream data, gene-expression data. Binary data mart are often characterized by two main features: high dimensionality and sparsity. This is the case, for example, of transactional data. In tabular form, a transaction data mart is a binary matrix: customers choices/DB records are stored on rows, products/attributes on columns. Cell values are '1' if a customer buys a product, '0' otherwise. It is then fair considering that the data matrix has a large number of columns,as many as the products available in a store (high dimensionality), and that most of the cells are null (sparsity). Both of these aspects make difficult to identify and describe relations among blocks of columns and blocks of rows. A well known exploratory approach to analyze this structures is frequent pattern mining which identifies groups of highly co-occurring attributes. Although this approach is computationally efficient, it produces huge output that hides away interesting relations and that is generally difficult to interpret. An alternative approach is to quantify the binary attributes on the basis of the underlying association structure using Multiple Correspondence Analysis (Greenacre, 2007). Dimensionality of data is drastically reduced: this aspect eases exploratory purposes such as the identification of groups of records or groups of co-occurring attributes. It also provides a visualization display of the association structure that eases user interpretation. In this paper it is proposed a quantification of binary attributes that takes into account external information coded as a categorical variable. In particular, the proposed quantification enphasize both the co-occurrences among attributes *and* the groups of records defined by the external categorical variable. The reference method is a reformulation of Nonsymmetric Correspondence Analysis (Lauro and D'Ambra, 1984). The general criterion beyond the quantification is to maximize the following quantity

$$\frac{1}{n}\sum_{k=1}^{K}\left(\sum_{j=1}^{p}\frac{f_{kj}^2}{f_{.j}}-\frac{f_{k.}^2}{n}\right)(k=1,\dots K, j=1\dots J) \tag{1}$$

where f_{kj} is the number of records that present the k^{th} category of the external variable and the j^{th} binary attribute; $f_{.j}$ is the number of records the j^{th} binary attribute; $f_{k.}$ is the number of records that present the k^{th} category of the external variable; n is the total number of records. Remark that the external categories can be refferred to a previous clustering of the records. This method can also be integrated in a two-step procedure together with a clustering step, as in Palumbo and Iodice D'Enza (2009). The paper illustrates an application of the proposed method to a binary data set representing a group of italian students of the University of Macerata. In particular, binary attributes refer to the exams passed by the students, while external information will refer to socio-demographic characteristics of the students.

References

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

N.C. Lauro and L. D'Ambra (1984). L'analyse non symmétrique des correspondances. In E. Diday et al., eds, *Data Analysis and Informatics, III.* North-Holland, 1984.

F.Palumbo and A.Iodice D'Enza (2009). Clustering and Dimensionality Reduction to Discover Interesting Patterns in Binary Data. In *Proceedings of GFKL08.* Hamburg, accepted, in press.