Issues in joint dimension reduction and clustering methods

Metodi congiunti di riduzione della dimensionalità e classificazione automatica: alcuni aspetti applicativi

Michel van de Velden, Alfonso Iodice D'Enza and Angelos Markos

Abstract Joint data reduction (JDR) methods consist of a combination of well established unsupervised techniques such as dimension reduction and clustering. Distance-based clustering of high dimensional data sets can be problematic because of the well-known curse of dimensionality. To tackle this issue, practitioners use a principal component method first, in order to reduce dimensionality of the data, and then apply a clustering procedure on the obtained factor scores. JDR methods have proven to outperform such sequential (tandem) approaches, both in case of continuous and categorical data sets. Over time, several JDR methods followed by extensions, generalizations and modifications have been proposed, appraised both theoretically and empirically by researchers. Some aspects, however, are still worth further investigation, such as *i*) the presence of mixed continuous and categorical variables; *ii*) outliers undermining the identification of the clustering structure. In this paper, we propose a JDR method for mixed data: the method in question is built upon existing continuous-only and categorical-only JDR methods. Also, we appraise the sensitivity of theproposed method to the presence of outliers.

Abstract *Abstract in Italian* I metodi congiunti di sintesi dei dati (Joint data reduction, JDR) rappresentano una combinazione di approcci di analisi non supervisionata quali tecniche fattoriali e classificazione automatica. La classificazione automatica, quando basata sulla distanza tra le osservazioni, diventa di difficile applicazione quando queste siano descritte da un elevato numero di variabili, a causa della *maledizione della dimensionalità*. Per aggirare tale problema, è pratica comune ridurre la dimensionalità dei dati utilizzando tecniche fattoriali, e applicare successi-

Michel van de Velden
Erasmus University of Rotterdam, visiting Professor at Università della Campania Luigi Vanvitelli
e-mail: vandevelden@ese.eur.nl

Alfonso Iodice D'Enza
Università di Cassino e del Lazio Meridionale e-mail: iodicede@unicas.it

Angelos Markos
Democritus University of Thrace e-mail: amarkos@eled.duth.gr

vamente la classificazione automatica sui fattori ottenuti in precedenza. In letteratura è stato dimostrato empiricamente che a tale approccio sequenziale è preferibile utilizzare dei metodi JDR. Il filone di ricerca sui metodi JDR comprende varianti, generalizzazioni e confronti teorici ed empirici tra i diversi metodi proposti. Tuttavia, alcuni aspetti applicativi non sono ancora stati oggetto di studio: in particolare, si fa riferimento ai casi di dataset che contengano sia variabili quantitative che qualitative, e alla presenza di valori anomali. L'obiettivo del presente contributo è quello di presentare un metodo di JDR che sia applicabile ad insiemi di dati di tipo misto. Inoltre, si vuole valutare la robustezza delle soluzioni ottenute in presenza di valori anomali.

Key words: Dimension reduction, cluster analysis, mixed data sets

1 Joint dimension reduction and clustering methods

Distance-based clustering methods aim at defining groups such that observations that belong to the same group are similar to each other. The distance or dissimilarity measure being used depends on the nature of the considered variables. When the set of observations is described by a large number of variables, it becomes difficult to calculate meaningful pair-wise dissimilarities and, hence, to define clusters. To overcome this problem, methods that combine dimension reduction with cluster analysis have been proposed.

The most straightforward approach is to apply dimension reduction prior to clustering, the latter being therefore applied to the scores obtained in the first step. In this two-step approach, however, two different criteria are optimized: in particular, while dimension reduction aims at defining a reduced set of combinations of the original variables that maximize the original variability, cluster analysis aims at maximize the between-groups variability. This may lead to the cluster masking problem (e.g., van Buuren and Heiser, 1989; De Soete and Carroll, 1994; Vichi and Kiers, 2001) and several solutions have been proposed that proposed a combined optimization of the two steps. We refer to this class of methods as joint data reduction (JDR).

Methods for JDR have been proposed that deal with continuous and categorical data. In particular, for continuous (or, interval) data we consider reduced K-means (De Soete and Carroll, 1994), factorial K-means (Vichi and Kiers, 2001) as well as a compromise version of these two methods. For categorical data, cluster correspondence analysis (van de Velden et al, 2017), which, for the analysis of categorical data, is equivalent to GROUPALS (van Buuren and Heiser, 1989), multiple correspondence analysis and K-means (MCA K-means; Hwang et al, 2006), and iterative factorial clustering of binary variables (i-FCB; Iodice D'Enza and Palumbo, 2013) are considered.

2 Clustering mixed data

Data sets with observations being described by both continuous and categorical variables are common in real applications. Since most clustering procedures are designed to deal with variables on a same scale, a simple strategy is to homogenize the variables in a pre-processing phase. That is, either re-coding the continuous variables or the categorical ones. Recoding of continuous variables is achieved via discretization, the range of each continuous variable is split into a set of intervals, and all the values falling in a same interval are labeled with a same category. Of course, such kind of discretization leads to a loss of the original information. To overcome this issue, an alternative transformation through discretization is to code the original values of a continuous variable into a pre-specified number of fuzzy categories, i.e. to a set of k nonnegative values that sum to 1, quantifying the *possibility* of the variable to be in each category. These *pseudo-categorical* values represent each value of a continuous variable uniquely and exactly, i.e. the numerical information of the original variable is preserved (Aşan and Greenacre, 2011). Recoding of categorical data variables aims to put them on the same scale as the continuous. A rather general pre-processing and standardization approach is described in Mirkin (2012). Another approach, described by Everitt et al (2011), consists of clustering the data by type of variable, and then merge the obtained clustering solutions; the obvious drawback of this approach is that any relation between the two sets of variables is ignored.

A more direct approach is to use a dissimilarity measure designed for mixed data. The most popular one is Gower's dissimilarity coefficient, which takes into account the different nature of the variables (e.g., see Everitt et al (2011)). Once pairwise distances are obtained, a clustering procedure such as partitioning around medoids (PAM) can be applied on the distance matrix. Further partitioning methods for mixed data consist of extensions of K-means: examples are the K-prototypes (Huang, 1998), and the K-means Modha-Spangler weighting (Modha and Spangler, 2003), among others. In all of these approaches, assigning weights to variables is a sensitive task.

Probabilistic or model-based clustering is also a very popular way of clustering mixed-type data. Such methods typically assume the observations to follow a normal-multinomial finite mixture model and proved to be effective when the parametric assumptions are met. In this paper, however, we focus on distance-based clustering approaches.

3 JDR for mixed data

Let $\mathbf{X}$ denote a centered and standardized $n \times Q$ data matrix,, $\mathbf{B}$ is a $Q \times d$ columnwise orthonormal loadings matrix, i.e. $\mathbf{B}'\mathbf{B} = \mathbf{I}_d$, where d is the user supplied dimensionality of the reduced space. Furthermore, $\mathbf{Z}_K$ is the $n \times K$ binary matrix indicating cluster memberships of the n observations into the K clusters. Finally, we use $\mathbf{G}$ to

denote the $K \times d$ cluster centroid matrix.
A JDR method for continuous data is reduced K-means clustering (RKM) De Soete and Carroll (1994): both the dimension reduction and cluster analysis aim at maximizing the *between* variance of the clusters in the reduced space. The RKM objective function is

$$\min \phi_{\text{RKM}}(\mathbf{B}, \mathbf{Z}_K, \mathbf{G}) = \left\| \mathbf{X} - \mathbf{Z}_K \mathbf{G} \mathbf{B}' \right\|^2, \tag{1}$$

where $\|\cdot\|$ denotes the Frobenius norm. It can be shown, that the above minimization problem is equivalent to

$$\max \phi'_{\text{RKM}}(\mathbf{Z}_K, \mathbf{B}) = trace \mathbf{B}' \mathbf{X}' \mathbf{P} \mathbf{X} \mathbf{B} \tag{2}$$

Similar to RKM is cluster CA, a JDR method for categorical data: CCA aim is also to maximize the between cluster variation in reduced space. Let $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$ be the superindicator matrix of dummy coded categorical data and $\mathbf{M}$ the corresponding centering operator. The objective of cluster CA can be expressed as

$$\max \phi_{\text{clusca}}(\mathbf{Z}_K, \mathbf{B}) = trace \mathbf{B}' \mathbf{Z}' \mathbf{M} \mathbf{P} \mathbf{M} \mathbf{Z} \mathbf{B} \quad \text{s.t.} \quad \frac{1}{np} \mathbf{B}' \mathbf{D}_z \mathbf{B} = \mathbf{I}_k. \tag{3}$$

subject to

Comparing this equation to (2), we see that the methods are closely related. Furthermore, letting $\mathbf{B}^* = \frac{1}{\sqrt{np}} \mathbf{D}_z^{1/2} \mathbf{B}$ it is possible to re-write the clusterCA optimization problem as

$$\min \phi_{\text{CCA}}(\mathbf{B}^*, \mathbf{Z}_K, \mathbf{G}) = \left\| \mathbf{D}_z^{-1/2} \mathbf{M} \mathbf{Z} - \mathbf{Z}_K \mathbf{G} \mathbf{B}^{*\prime} \right\|^2, \tag{4}$$

subject to

$$\mathbf{B}^{*\prime} \mathbf{B}^* = \mathbf{I}_k$$

which is closely related to the RKM problem in Equation 1, and therefore the two equations can be combined to define the problem for mixed data. In particular from Equations (1) and (4) we can formulate as objective for a joint analysis of mixed data:

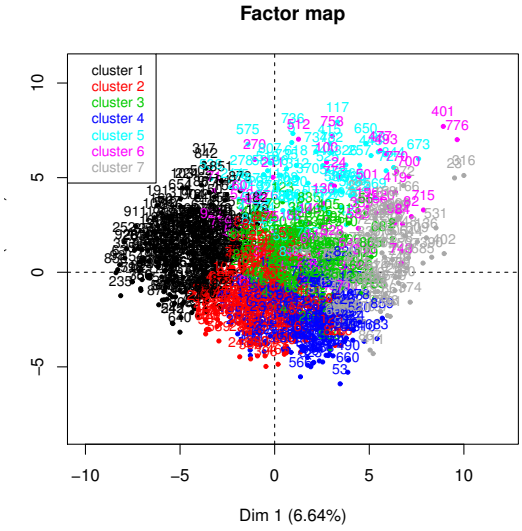
$$\min \phi_{\text{mixed RKM}}(\tilde{\mathbf{B}}, \mathbf{Z}_K, \mathbf{G}) = \left\| \begin{pmatrix} \mathbf{X} & \mathbf{D}_z^{-1/2} \mathbf{M} \mathbf{Z} \end{pmatrix} - \mathbf{Z}_K \mathbf{G} \tilde{\mathbf{B}}' \right\|^2, \tag{5}$$

where $\tilde{\mathbf{B}}' = [\mathbf{B}'_1 \quad \mathbf{B}'_2]$ and $\tilde{\mathbf{B}}' \tilde{\mathbf{B}} = \mathbf{I}_d$.
In a similar way, it is possible to combine FKM with clusterCA and define a further JDR method for mixed data.

4 Clustering patients with low back pain

In this section, we illustrate Mixed RKM on a real dataset. The dataset was part of a clustering challenge connected with the International Federation of Classification

Fig. 1 Mixed RKM factorial map of the patients on the first and second dimension. Clusters are indicated with different colours.



Societies (IFCS) 2017 conference. It contains baseline and outcome assessment of low back pain in 928 adult patients who were consulting chiropractors in Denmark. The clustering aim is to find a (semi-) automatic classification of the patients based on 112 pain history and work-related variables, in order to find clinically applicable and useful groups. 38 of the variables were treated as continuous and 84 as categorical. The data set along with associated meta-data and variable descriptions, can be downloaded at http://ifcs.boku.ac.at/repository/challenge2/.

Missing data were imputed using the regularised iterative FAMD algorithm (Audigier et al, 2016). Mixed RKM was then applied on the imputed dataset. The number of dimensions, 5, was determined on the basis of both empirical and statistical criteria. Solutions between 3 and 12 clusters were investigated. Average Silhouette Width and the Calinski-Harabasz index were used to assess cluster separation. A solution with 7 clusters was selected. Figure 1 depicts the object (patient) scores on the first and second dimension. The seven clusters are indicated with different colours. The description of each cluster in terms of the variables involved in clustering as well as a set of external (outcome) variables was perfomed using the function `catdes()` (package `FactoMineR` Lê et al (2008)). All clusters were significantly associated with external (outcome) variables, which provides evidence for external validity.

Cluster 1 (8.7%): acute suffering LBP, psychological effects of pain, mild improvement after 12m.

Cluster 2 (13.5%): acute suffering LBP, reduced activity, little time in pain, major improvement after 12m.

Cluster 3 (5.2%): leg pain only, showed mild improvement after 12m.
Cluster 4 (9.2%): acute suffering LBP, pain attributed to work, higher bmi, mild improvement after 12m.
Cluster 5 (16.4%): acute suffering LBP, age higher than average, reduced physical activity, little improvement after 12m.
Cluster 6 (12.2%): mild LBP intensity, age less than average, no psychological effects, major improvement after 12m.
Cluster 7 (34.6%): low LBP intensity, major improvement.
Finally, the application of mixed RKM to simulated mixed data, with and without presence of outliers, showed promising results both in terms of effectiveness in cluster structure identification and robustness to outliers.

References

Aşan Z, Greenacre M (2011) Biplots of fuzzy coded data. Fuzzy sets and Systems 183(1):57–71

Audigier V, Husson F, Josse J (2016) A principal component method to impute missing values for mixed data. Advances in Data Analysis and Classification 10(1):5–26

van Buuren S, Heiser W (1989) Clustering n objects into k groups under optimal scaling of variables. Psychometrika 54:699–706

De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional euclidean space. In: Diday E, Lechevallier Y, Schader M, Bertrand P, Burtschy B (eds) New Approaches in Classification and Data Analysis, Springer-Verlag, pp 212–219

Everitt BS, Stahl D, Leese M, Landau S (2011) Cluster analysis. John Wiley & Sons

Huang Z (1998) Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery 2(3):283–304

Hwang H, Dillon WR, Takane Y (2006) An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. Psychometrika 71:161–171

Iodice D'Enza A, Palumbo F (2013) Iterative factor clustering of binary data. Computational Statistics 28(2):789–807

Lê S, Josse J, Husson F, et al (2008) `FactoMineR`: an `R` package for multivariate analysis. Journal of statistical software 25(1):1–18

Mirkin B (2012) Clustering: a data recovery approach. CRC Press

Modha DS, Spangler WS (2003) Feature weighting in *k*-means clustering. Machine learning 52(3):217–237

van de Velden M, DEnza AI, Palumbo F (2017) Cluster correspondence analysis. Psychometrika 82(1):158–185

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