Robust Two-mode clustering

Maurizio Vichi
Department of Statistics, Sapienza University of Rome, Italy,
maurizio.vichi@uniroma1.it

Abstract

Starting from an extension of standard *K*-means for simultaneously clustering observations and features, namely Double *K*-Means (DKM) (Vichi, 2001), the model is developed in a probabilistic framework with a robustification necessary to take into account a certain amount of outlying observations assumed included in the data, that generally lead to unsatisfactory clustering results. An efficient algorithm is proposed and the advantages of using this approach are discussed.

Key Words: Two-mode clustering, double k-means, disjoint principal component analysis, robustness.

1. Introduction

Two-mode clustering is the activity of clustering modes (e.g., objects, variables) of an observed two-mode data matrix, simultaneously. This task is required because objects, frequently, are homogeneous only within subsets of variables, while variables may be strongly associated only on subsets of objects. For example, in microarray data analysis groups of genes are generally co-regulated within subsets of samples and groups of samples share a common gene expression pattern only for some subsets of genes. In market basket analysis customers have similar preference patterns only on subsets of products and, vice-versa, classes of products are more frequently consumed and preferred by subgroups of customers. In these situations a classical cluster analysis would cluster one mode (e.g., objects) on the basis of the complete set of the other mode (e.g., variables), thus producing weak results, while this is avoided with a more appropriate two-mode clustering. For *big data,* represented by matrices with a huge number of rows and columns, frequently the main analysis is a two-mode clustering, trying to mine and synthesize the relevant information by reducing the size of the data to a matrix of compact dimensions formed by prototype objects and variables. This is achieved by the simultaneous grouping of rows and columns so that results are informative and easy to interpret, denoting compressed, but relevant representation of the big data, while trying to preserve most of the original information. The reduction is generally soft to obtain a light compression of the multivariate data in order to allow the successive application of other multivariate statistical methods that are computationally prohibitive for large data sets. The quality of big data is not always certifiable and frequently they are inflated by many outliers and influential data that have an high impact on the two-mode clustering and successive analyses. Therefore, robust multimode clustering techniques are needed for compressing large data set, while preserving the most relevant information.

A new robust asymmetrical two-mode clustering technique is proposed. The applications on both, synthetic and real datasets, validate the performance and applicability of the new algorithm.

2 Double K-means Model

Given a data set with I objects and J variables, and associated data matrix $\mathbf{X}$ of dimension $(I \times J)$, a *two-mode cluster* ${}_{OV}C_k$ *of objects and variables* (briefly 2-*mc* ${}_{OV}C_k$) is a set of ordered pairs $(o_i, v_j) \in (O \times V)$, with $o_i \in O$ and $v_j \in V$, where $(O \times V)$ is the *Cartesian product*, of the sets of objects O and variables V. Following the definition of *direct cluster* given by Hartigan, (1972), the 2-*mc* can be seen as a sub-matrix or *block* $\mathbf{X}_{pq}=[x_{ij} : (o_i, v_j), o_i \in {}_OC_p, v_j \in {}_VC_q]$ of the data matrix $\mathbf{X}$, with dimensions $(I_p \times J_q)$.

A *two-mode partition* $P_{OV} = \{{}_{OV}C_1,\ldots, {}_{OV}C_k,\ldots, {}_{OV}C_K\}$ is a set of K disjoint clusters ${}_{OV}C_K \in (O \times V)$, such that their union is $(O \times V)$ itself. Note that each pair (o_i, v_j) is completely assigned exactly to one of the K 2-*mc* and this guarantees that a partition is defined. Two-mode clusters forming a partition have associated blocks $\mathbf{X}_1, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K$. A unique observation x_{ij} is a *singleton block*, while matrix $\mathbf{X}$ is the *total bock*.

We now focus on two-mode partitions linked to partitions of sets O and V. Given marginal partitions $P_O=\{{}_OC_1,\ldots, {}_OC_p,\ldots, {}_OC_P\}$, $P_V=\{{}_VC_1,\ldots, {}_VC_q,\ldots, {}_VC_Q\}$ of O and V, respectively, the Cartesian product $(P_O \times P_V)$ defines a *two-mode single-partition* (briefly, 2-*msp*) that induces, one-to-one, a *two-mode partition of* $\mathbf{X}$ into blocks $\mathbf{X}_{11}, \ldots \mathbf{X}_{pq}, \ldots, \mathbf{X}_{PQ}$.

Any two-mode single-partition can be can be defined by modeling matrix $\mathbf{X}$ according to

$$\mathbf{X} = \mathbf{U}\overline{\mathbf{X}}\mathbf{V}' + \mathbf{E} \tag{1}$$

which has been called *double k-means model* (Vichi, 2001), where unknown partitions P_O, P_V, specified by membership matrices $\mathbf{U}$ and $\mathbf{V}$, need to be identified in order to best reconstruct matrix $\mathbf{X}$.

Matrix $\overline{\mathbf{X}}$ is the centroid matrix, with generic element $\bar{x}_{pq}$ representing the centroid of the 2-*mc* ${}_{OV}C_{pq}$, $p=1,\ldots,P$, $q=1,\ldots,Q$. The centroid matrix represents the relevant information in the data matrix $\mathbf{X}$ and can be seen as the reduced data matrix of dimension $(P \times Q)$, corresponding to P non-observable prototype objects described by Q non-observable prototype variables. The name double K-means model is justified by the property that if matrix $\mathbf{V}$ degenerates into the identity matrix of order J ($\mathbf{V}=\mathbf{I}_J$), model (1) becomes: $\mathbf{X} = \mathbf{U}\overline{\mathbf{X}} + \mathbf{E}$, that is, the clustering model implicitly associated to the K-means algorithm for partitioning objects (Ball et al., 1967 MacQueen, 1967). On the other hand when $\mathbf{U}$ degenerates into the identity matrix of order I ($\mathbf{U}=\mathbf{I}_I$),model (1) is $\mathbf{X} = \overline{\mathbf{X}}\mathbf{V}' + \mathbf{E}$, i.e., the clustering model implicitly associated to the K-means algorithm for partitioning variables. Thus, in general two interconnected K-means problems have to be considered in model (1). Model (1) for similarity data was introduced by Hartigan (1975) and generalized by De Sarbo (1982), who started from the ADCLUS model: $\mathbf{S}=\mathbf{U}\mathbf{W}\mathbf{U}'+\mathbf{C}+\mathbf{E}$, given by Shepard and Arabie, (1979), where $\mathbf{S}$ is the observed similarity matrix, $\mathbf{W}$ is a diagonal matrix weighting P clusters identified by the membership matrix $\mathbf{U}$ and $\mathbf{C}$ a constant matrix. In particular, GENNCLUS model: $\mathbf{S}=\mathbf{U}\mathbf{W}\mathbf{V}'+\mathbf{C}+\mathbf{E}$, was introduced for an asymmetric similarity matrix $\mathbf{S}$, where $\mathbf{W}$ (the centroid matrix) is supposed symmetric with zero diagonal elements and representing the association between P derived clusters for rows and Q derived clusters for columns of the asymmetric similarity matrix $\mathbf{S}$.

Busygin et al., (2008) noted some connections of the *double k-means model* with the *Singular Value Decomposition* (SVD) of a rectangular matrix $\mathbf{X}$ which generalizes spectral decomposition of a square symmetric matrix. SVD is applicable to any

rectangular matrix $\mathbf{X}$ and specifies orthogonal matrices $\ddot{\mathbf{U}}$ and $\ddot{\mathbf{V}}$ such that: $\ddot{\mathbf{U}}'\mathbf{X}\ddot{\mathbf{V}} = \mathrm{diag}(\sigma_1, \ldots, \sigma_K)$, with $K = \min(n, J)$, where $\sigma_1 \geq \cdots \geq \sigma_K$ are the *singular values* and $\ddot{\mathbf{U}}$, $\ddot{\mathbf{V}}$ are the *left* and *right singular vectors* of $\mathbf{X}$, respectively. Bi-clustering and SVD may be related, when matrix $\overline{\mathbf{X}}$ =diag($\mu_1, \ldots, \mu_K$); therefore, $\overline{\mathbf{X}}$ is constrained to be square of order K and diagonal. The corresponding reconstructed (ideal) matrix $\mathbf{X}$ is the block diagonal matrix of the form: $\mathbf{X}$ = blockdiag($\mathbf{X}_1$,…,$\mathbf{X}_K$), which is generally appropriate for (dis)similarity rectangular asymmetric data.
Note that any block matrix $\mathbf{X}_{pq}$, associated to 2-*mc* ${}_{OV}C_{pq}$ in model (1) can be re-parameterized as a ($I \times J$) extended block matrix

$$ {}_{e}\mathbf{X}_{pq} = \mathbf{u}_p \, \bar{x}_{pq} \mathbf{v}_q', \tag{2} $$

that takes the values 0 for all objects and variables that do not belong to the 2-*mc* ${}_{OV}C_{pq}$, while has value $\bar{x}_{pq}$ if the corresponding object and variable belongs to ${}_{OV}C_{pq}$. Since clusters are disjoint, we have

$$ \mathbf{U}\overline{\mathbf{X}}\mathbf{V}' = \sum_{p=1}^{P} \sum_{q=1}^{Q} \mathbf{u}_p \, \bar{x}_{pq} \mathbf{v}_q' \, . \tag{3} $$

Long et al., (2005) propose model (1) for nonnegative $\mathbf{X}$; thus, requiring nonnegative $\mathbf{U}$, $\mathbf{V}$ and $\overline{\mathbf{X}}$ and relaxing the binary form of $\mathbf{U}$ and $\mathbf{V}$.
The least-squares assessment of model (1) leads to the formulation of the following quadratic optimization problem that has to be solved with respect to variables u_{ip}, v_{jq} and $\bar{x}_{pq}$, where constraints (6) and (7) are necessary to specify an objects (rows) partition P_O, while constraints (8) and (9) are needed to ensure a variables (columns) partition P_V, (Vichi, 2001),

$$ J_{DKM}(\mathbf{U},\mathbf{V},\overline{\mathbf{X}}) = \min_{\mathbf{U},\mathbf{V},\overline{\mathbf{X}}} \| \mathbf{X} - \mathbf{U}\overline{\mathbf{X}}\mathbf{V}' \|^2 \tag{4} $$

$$ = \min_{\mathbf{H_U},\mathbf{H_V},} \| \mathbf{X} - \mathbf{H_U}\mathbf{X}\mathbf{H_V} \|^2 \tag{5} $$

subject to [P1]

$$ u_{ip} \in \{0, 1\} \qquad i=1,\ldots,I;\ p=1,\ldots,P; \tag{6} $$

$$ \sum_{p=1}^{P} u_{ip} = 1 \qquad i=1,\ldots,I; \tag{7} $$

$$ v_{jq} \in \{0, 1\} \qquad j=1,\ldots,J;\ q=1,\ldots,Q; \tag{8} $$

$$ \sum_{q=1}^{Q} v_{jq} = 1 \qquad j=1,\ldots,J; \tag{9} $$

Form the previous formulation of DKM an iterative relocation algorithms can be developed, which turns out to be a coordinate descent algorithm

ALGORITHM 1 *DKM Coordinate Descent Algorithm for objective function* (4)

Step 0 Initialization. Generate a random membership matrix $\mathbf{U}$ and $\mathbf{V}$.
Step 1 Update $\overline{\mathbf{X}}$ given $\mathbf{U}$ and $\mathbf{V}$ by: $\overline{\mathbf{X}} = \left(\hat{\mathbf{U}}'\hat{\mathbf{U}}\right)^{-1}\hat{\mathbf{U}}'\mathbf{X}\hat{\mathbf{V}}\left(\hat{\mathbf{V}}'\hat{\mathbf{V}}\right)^{-1}$;

Step 2 Update $\mathbf{U}$, given $\hat{\overline{\mathbf{X}}}$ and $\hat{\mathbf{V}}$

$\forall i$: u_{ip}=1 *if* $\|\mathbf{x}_i - \hat{\mathbf{V}}\hat{\overline{\mathbf{X}}}'\mathbf{u}_k\|^2 = min\{ \|\mathbf{x}_i - \hat{\mathbf{V}}\hat{\overline{\mathbf{X}}}'\mathbf{u}_l\|^2 : l=1,\ldots,P\}$,
u_{ip}= 0 *otherwise* ;
where $\mathbf{x}_i$ is the ith row of matrix $\mathbf{X}$

Step 3 Update $\mathbf{V}$ given $\hat{\overline{\mathbf{X}}}$ and $\hat{\mathbf{U}}$

$\forall i$: v_{jq}=1 *if* $\|\mathbf{x}^j - \hat{\mathbf{U}}\hat{\overline{\mathbf{X}}}\mathbf{v}_q\|^2 = \min\{ \|\mathbf{x}^j - \mathbf{U}\overline{\mathbf{X}}\mathbf{v}_l\|^2 : l=1,\ldots,Q\}$,
v_{jq}= 0 *otherwise* .

where $\mathbf{x}_j$ is the jth columns of matrix $\mathbf{X}$

Stopping Rule: *If* $J_{DKM}(\mathbf{U},\mathbf{V},\bar{\mathbf{X}})$ decreases less than an arbitrary small constant $\varepsilon>0$ the algorithm has converged.

The algorithm iterates steps 1-3 until stopping role applies, which generally follows after few steps. At each step the algorithm decreases the loss function (4) and since it is bounded below by zero, it stops to a stationary point which is at least a local minimum of the problem. Problem [P1] is NP-hard because it includes an NP-hard problem like *K*-means. Therefore to increase the chance to detect to global optimal solution it is advised to use a *multistart* procedure and retain the best result. Note that in Step 1 $\left(\hat{\mathbf{U}}'\hat{\mathbf{U}}\right)^{-1}$ and $\left(\hat{\mathbf{V}}'\hat{\mathbf{V}}\right)^{-1}$ are diagonal matrices therefore their inverse are obtained by simply inverting the elements on the diagonal.
It can be observed that the centroid matrix $\bar{\mathbf{X}}$ in the LS estimation of model (1) is a mean matrix influenced by the presence of a certain amount of outlying observations. Two approaches can be used to take into account the presence of *outliers*. As an alternative to compute the centroid matrix in step 1 of the algorithm, the matrix of *medoids* can be estimated by choosing for each cluster the closest observation in a LS sense. A second, and more elaborated approach consists in the introduction of the *concentration* step in the algorithm above, as described in García-Escudero et al. (2008). In each concentration step, the proportion α of the most remote observations (considering Euclidean distances) to previous *K* centers are discarded and, then, *K* new centers are obtained by computing means of the non-discarded observations. In this papers these two methodologies are discussed.

References

Ball G. H., Hall D. J., (1967) "A clustering technique for summarizing multivariate data", *Behavioral Science*, 12, 153-5.
Busygin S., Prokopyevb O., Pardalosa P. M., (2008) "Biclustering in data mining", *Computers & Operation Research*, 35, 2964–2987.
DeSarbo W. S., (1982) "GENNCLUS: new models for general nonhierarchical clustering analysis", *Psychometrika*, 47, 449–75.
García-Escudero, L., Gordaliza, A., Matran, C., Mayo-Iscar, A., (2008) "A general trimming approach to robust cluster analysis", Annals of Statistics 36, 1324–1345.
Hartigan J. A., (1972) "Direct clustering of a data matrix", *JASA*, 67, 337, 123-129.
Hartigan J. A., (1975) "*Clustering Algorithms*", Wiley, New York.
Long B., Zhang Z., YU P. S., (2005). Co-clustering by block value decomposition, in *Proceedings of SIGKDD'05*, 2005.
MacQueen J. (1967) "Some methods for classification and analysis of multivariate observations", in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-97.
Martella F. and Vichi M. (2012) "Clustering microarray data using model based double K-means", *Journal of Applied Statistics*,
Vichi M, (2001) "Double k-means clustering for simultaneous classification of objects and variables", in *Advances in Classification and Data Analysis*, S. Borra, R. Rocci, M. Vichi, and M. Schader, eds., Springer, Heidelberg, 43–52.
Shepard R. N., Arabie P., (1979) "Additive clustering: Representation of similarities as combinations of discrete overlapping properties", *Psychological Review*, 86, 86-123.
Vichi M. and Saporta G. (2009) "Clustering and Disjoint Principal Component", *Computational Statistics & Data Analysis*, vol. 53, 8, 3194-3208.