New K-means clustering methods with dimension reduction

Ryo Takahashi

Cluster analysis, also called data clustering, taxonomy analysis, or unsupervised classification, is a method of creating groups of objects in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct. Cluster analysis has been developed in various disciplines: computer science, economics, engineering, psychology, biology, medical science, and so on. Particularly, K-means clustering, the most popular and the simplest partitional algorithm, is used extensively in real world and studied in various scientific fields. K-means clustering is designed to partition n objects into K classes. The number of clusters K is specified by users.

When the number of variables becomes large, or a cluster structure lies in a low-dimensional subspace of data, it is widely known that the conventional K-means algorithm does not work well. This problem is called *"Curse of Dimensionality"* (Bellman, 1957). In this case, researchers often apply the following two-steps procedure; [1] carry out principal component analysis (PCA) for dimension reduction, and [2] apply the conventional K-means algorithm to the principal scores on the first few principal components. This approach is called *"tandem clustering"* and widely used as common subspace clustering. However, it has been criticized by several authors, because PCA may identify dimensions that do not necessarily contribute much to capturing an underlying cluster structure in a data set. In other words, each step in this approach aims to optimize a different optimization criterion.

To avoid this problem, De Soete & Carroll (1994) proposed a method simultaneously partitioning the n objects in $\mathbf{X}$ into K clusters and finding the q components that summarize p variables, with $K < n$ and $q < p$. This method, called *"Reduced K-means"* (RKM), is formulated by combining the objective function of K-means clustering and that for principal component analysis (PCA) into a single criterion to be optimized. Therefore, RKM could identify the low-dimensional space that keeps the information about the cluster structure underlying a data set. Let $\mathbf{X}$ be an n-objects $\times$ p-variables data matrix. The model of RKM is written as

$$f_{RKM}(\mathbf{U},\mathbf{C},\mathbf{A}) = \|\mathbf{X}-\mathbf{UCA'}\|^2$$

where $\mathbf{U}$ is an $n \times K$ binary indicator matrix, $\mathbf{C}$ is a $K \times q$ matrix of cluster centroids, and $\mathbf{A}$ is a $p \times q$ loading matrix. The constraint $\mathbf{A'A}=\mathbf{I}$ is imposed on the loading matrix, where $\mathbf{I}$ represent an identity matrix.

The purpose of our research is to modify this RKM model so that loading matrix $\mathbf{A}$ is easier to interpret. If the matrix is sparse, i.e., includes a number of zero elements, its interpretation is facilitated. The elements of loadings indicate how strongly each of the observed variables (rows) contributes to the principal components (columns). If the principal components are constructed by a small subset of the observed variables, they are said to be interpretable. However, this is not often the case. Each principal component is represented by a linear combination of all observed variables. Especially, when the number of variables becomes large, this complicates the interpretation of the low-dimensional space extracted by RKM. In RKM or other clustering methods, one of the most important purposes is to capture the

characteristics of each cluster. Then, we need to obtain a loading matrix which indicates an interpretable low-dimensional subspace in which every cluster is embedded.

Thus, we solved this problem by employing some penalty constraints on the loadings obtained by RKM. Such penalty constraints make the loading matrix sparse mathematically. In our paper, we took two approaches to achieve it; [1] cardinality constraint approach, and [2] *LASSO* type penalized approach.

In first approach, we consider the following minimization problem:

$$\mathrm{Min}_{\mathbf{U,C,A}} \|\mathbf{X}-\mathbf{UCA'}\|^2 \quad \text{s.t. } \mathbf{C'U'UC}=m\mathbf{I},\ \mathrm{card}(\mathbf{A})=m\ .$$

In this equation, the second constraint card(**A**) indicates the cardinality of **A** and the number of cardinality m should be specified by users. The cardinality is the number of non-zero elements in an arbitrary matrix. Thus, the cardinality parameter m controls the sparsity of **A**. We can solve this minimization problem by using an alternating least-squares algorithm.

In second proposal, we consider the following minimization problem:

$$\mathrm{Min}_{\mathbf{U,C,A}} \|\mathbf{X}-\mathbf{UCA'}\|^2 \quad \text{s.t. } \mathbf{C'U'UC}=\mathbf{I},\ \|\mathbf{a}_j\|_1 \leqq t,\ \mathrm{diag}(\mathbf{A'A})=\mathbf{I}\ ,$$

where $\mathbf{a}_j$ ($j = 1,\cdots,q$) are column vectors of **A** and t is the tuning parameter which controls the sparsity of loadings. In this equation, the second constraint $\|\mathbf{a}_j\|_1 \leqq t$ is called L_1-norm penalty function, or *LASSO* type penalty, and widely used in the area of statistics. With fixed $\mathbf{F}=\mathbf{UC}$, this equation can be reformulated as following maximization problem:

$$\mathrm{Max}_{\mathbf{a}_j}\ \mathbf{f}_j'\mathbf{X}\mathbf{a}_j \text{ s.t. } \|\mathbf{a}_j\|_1 \leqq t,\ \mathbf{a}_j'\mathbf{a}_j \leqq 1\ (j = 1,\cdots,q)\ .$$

This reformulation takes the key role in an algorithm.

In both proposed methods, we must choose the sparsity parameter (m and t) in advance. Generally, the appropriate values of these parameters are unknown. In practical situation, we recommend to use some cross-validation (CV) methods for selecting these parameters (e.g. Wang, 2010). However, the existing CV methods choose the optimal values based only on clustering instability, not on interpretability of the loadings. Thus, we should prove that the validity of using these CV methods for our proposals. It remains as our future work.

References

[1] Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

[2] De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. *New Approaches in Classification and Data Analysis*, pp. 212-219. Heidelberg: Springer.

[3] Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, **97**, pp. 893-904.

(行動統計科学)