한국경영과학회/대한산업공학회 2003 춘계공동학술대회
2003년 5월 16일-17일 한동대학교(포항)

A Simple Tandem Method for Clustering of Multimodal Dataset

C. Cho, S.Y.Kim, J.W. Lee
Department of Industrial Engineering, POSTECH

Abstract

The presence of local features within clusters incurred by multi-modal nature of data prohibits many conventional clustering techniques from working properly. Especially, the clustering of datasets with non-Gaussian distributions within a cluster can be problematic when the technique with implicit assumption of Gaussian distribution is used. Current study proposes a simple tandem clustering method composed of k-means type algorithm and hierarchical method to solve such problems. The multi-modal dataset is first divided into many small pre-clusters by k-means or fuzzy k-means algorithm. The pre-clusters found from the first step are to be clustered again using agglomerative hierarchical clustering method with Kullback-Leibler divergence as the measure of dissimilarity. This method is not only effective at extracting the multi-modal clusters but also fast and easy in terms of computation complexity and relatively robust at the presence of outliers. The performance of the proposed method was evaluated on three generated datasets and six sets of publicly known real world data.

1. Introduction

Many clustering algorithms have been developed for applications in different areas to find more appropriate and meaningful clusters from given data. The numerous researches indicate the necessity of developing clustering methods specific for certain data characteristics. Our study proposes .

Our study proposes a method called tandem clustering process (TCP) designed for data with multi-modal or non-Gaussian distributions within clusters. The proposed TCP method is constitutedof conventional k-means and hierarchical algorithms with Kullback-Leibler divergence as a measure of dissimilarity. By using k-means algorithm in the first step of generating pre-clusters, the effect of outliers and the computational time is to be reduced compared to running hierarchical algorithm alone. Secondly, the implementation of Kullback-Leibler (K-L) divergence into hierarchical algorithm would extract the multi-modal clusters effectively.

Two major branches of the conventional clustering techniques are hierarchical clustering and non-hierarchical algorithms such as k-means and fuzzy k-means method. Since our proposed method tries to resolve some of the disadvantages that conventional hierarchical and k-means algorithm have and magnify their advantages, the following section of literature survey is devoted to a short survey on hierarchical and K-means clustering algorithms.

2. Previous Works

Hierarchical Agglomerative Clustering (HAC)

The basis of hierarchical clustering is a cluster hierarchy and the essence of the algorithms is to build a tree structure from the data having such hierarchy. Hierarchical agglomerative clustering (HAC) algorithms start with N number of single data-point clusters and merges a pair of clusters that are closest to one another (or most similar to one another) recursively. After a single merge, new distance (or dissimilarity or similarity) between all pairs of clusters are re-calculated before the next merge. The process is repeated until a stopping criterion is satisfied or all data points are merged into a single clusterand the hierarchical structure of the clusters are represented in the form of dendrogram.

One of the most critical issues in HAC is the measure of dissimilarity between pair of clusters called linkage metrics. The linkage metrics can be subdivided into graphic metrics and geometric metrics as described in Dash et al. (2003). Single, complete, average linkage are included in graphic methods considering each point in a cluster to be its representative and centroid and Ward's linkage are geometric metrics representing a cluster by its central point (Berkin, 2002).

The early algorithms proposed by Sibson implemented single linkage as a measure of dissimilarity (or distance) between clusters where the distance between two clusters is represented by the minimum distance between points in the two clusters (Sibson, 1973). The complete linkage used in the algorithm proposed by Defays (1977) uses the maximum distance to be the representative distance between two clusters. Voorhee's method (Voorhee, 1986) is based on the average link where the dissimilarity is measured as the average distance between points in the two clusters.

The geometric type of linkage method selects one representative point in each cluster and calculates the distance (dissimilarity) between them. Centroid method uses each cluster's centroid to be its representative (Dash

et al., 2003). Ward's minimum variance method (Ward, 1963) select a pair of clusters to give the minimum increase in the sum of squared errors Current study proposes a relatively new measure of dissimilarity called Kullback-Leibler divergence for clusters of distribution which is discussed in detail in section 3.2.

K-means and Fuzzy K-means Method

K-means method first proposed by Ball and Hall (1967) is one of the most popular clustering algorithms in many application areas. It directly assigns each observation to a cluster and each observation belongs to one and only one cluster. As the name "k-means" implicitly indicates, this method groups data points around k number of centroids by assigning an observation to the nearest centroid using greedy heuristics applied for iterative optimization.

An appropriate representation of the sum of difference between an observed data point and its centroid shown below is used as the objective function which is also equal to the total intra-cluster variance (Hastie, 2001). The objective function can be represented as

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} \left\| \mathbf{x}_j - \mu_i \right\|^2 ,$$

where the μ_i and $\mathbf{x}_j$ represent the centroids of cluster i and individual data respectively. The squared errors are summed over all n data points and c clusters. The clustering process begins with the initial k number of centroids and the data points are assigned to one of the k centroids in such a way to minimize the objective function. The initial cluster assignment is used to calculate a new set of centroids to minimize the total cluster variance and the data points are re-assigned to the new centoids minimizing the sum of squared error criterion. This process of re-calculating centroids and re-assignment of points are repeated until the convergence is achieved.

The major advantage of the k-means algorithm is the comparatively simple computation. The fast algorithm is especially favorable when the high dimensional data with large number of data points is dealt with. However, the solutions found from the conventional k-means process described are bound to be sub-optimal local minima and depend heavily on the location of initial centroids.

The original k-means algorithm was modified to more general fuzzy clusters by Dunn (1974) and Bezdek (1974). The fuzzy k-means clustering algorithm tries to minimize a heuristic global cost function of

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} \left[\hat{P}(\omega_i \mid \mathbf{x}_j, \hat{\theta}) \right]^b \left\| \mathbf{x}_j - \mu_i \right\|^2 ,$$

where $\hat{P}(\omega_i \mid \mathbf{x}_j, \hat{\theta})$ represents the probability of $\mathbf{x}_j$ to be in class i for a given set of parameters $\hat{\theta}$ and b is the fuzziness index which is greater than 1 (Duda, 2001).In fuzzy k-means algorithm, each point has probabilities of belonging to k different centroids instead of 0 or 1 in the case of k-means algorithm.

One critical disadvantage of the k-means and fuzzy k-means algorithms comes from their implicit assumption of Gaussian distribution of data points. They tend to group the data points in spherical clusters and often unsuccessful at detecting clusters with different shapes (Rousseeuw et al., 1996).The problem of missing multi-modal nature of datasets could be overcome by the proposed tandem clustering process.

3. Proposed Clustering Method

We propose a tandem clustering process (TCP) exhibiting exceptional performance for the multi-modal dataset and yet it is simple and relatively fast. The basic idea of TCP is to apply simple k-means or fuzzy k-means algorithm to the raw dataset including the multi-modal clusters. The first step would group the data into many small clusters called pre-clusters each having normal distribution. In the second step, hierarchical clustering method is applied to the pre-clusters using Kullback-Leibler divergence as a measure of distance.

Kullback-Leibler Divergence

Kullback-Leibler divergence, also called as relative entropy, is a measure of the difference between two arbitrary distributions (Ripley, 1996). The general Kullback-Leibler divergence is written as

$$D(f \mid g) = \int f \log \frac{f}{g} ,$$

where f and g represent the arbitrary distributions.

The above equation can be modified to give symmetric difference (or distance) between two distributions, $k1$ and $k2$, as

$$I(k_1, k_2) = \frac{1}{2} \int p(x \mid k_1) \log \frac{p(x \mid k_1)}{p(x \mid k_2)} dx + \frac{1}{2} \int p(x \mid k_2) \log \frac{p(x \mid k_2)}{p(x \mid k_1)} dx ,$$

where $p(\mathbf{x} \mid k1)$ and $p(\mathbf{x} \mid k2)$ are the conditional probability density of x for distribution $k1$ and $k2$ respectively (Ripley, 1996).

The symmetric Kullback-Leibler distance can be simplified when the two distributions for which the distance is to be measured are assumed to be Gaussian. The simplified expression of the distance between two guassian distributions derived by Larsen et al.(2002) is

$$I(k_1, k_2) = \frac{1}{4} (\mu_{k_1} - \mu_{k_2})^T (\Sigma_{k_1}^{-1} + \Sigma_{k_2}^{-1})(\mu_{k_1} - \mu_{k_2}) + \frac{1}{4} \mathrm{tr}[\Sigma_{k_1} \Sigma_{k_2}^{-1} + \Sigma_{k_2} \Sigma_{k_1}^{-1} - 2I] ,$$

The μ_{ki} and Σ_{ki} represent the mean and covariance matrix of cluster i respectively and a simple Euclidean distance is expressed as d. Since all clusters in the first level of hierarchical clustering can be assumed to be

Gaussian in our TCP method, above equations can be used in the first step of hierarchical clustering without modification.

However, the K-L(Kullback-Leibler) distances weighted by the mixing proportions was substituted beyond the first level in hierarchical clustering process since the distributions within some clusters are not Gaussian from the second level of merge. The simplified representation of the weighted K-L distance for non-Gaussian distributions derived by Larsen et al.(2002) can be written as

$$D_{j+1}(k,k_3) = \frac{(P_j(k_1)D_j(k_1,k_3) + (P_j(k_2)D_j(k_2,k_3)}{P_j(k_1) + P_j(k_2)} .$$

This equation can approximate the distance from cluster k and k_3 where k is the resultant of the previous merge of clusters k_1 and k_2. Also, $P_j(k_i)$ represents the priors of the cluster k_j at j-th level.

Since the resultant from step one is a set of many, small Gaussian-like distributions, Kullback-Leibler divergence used in the current process could be far more effective than any of the linkage metrics mentioned in the previous section.

Tandem Clustering Process

The first step is to run k-means algorithm with cluster number greater than the expected number (between twice to six times the expected number) of clusters. This step of producing k' pre-clusters segregates the small Gaussian distributions within multi-modal clusters effectively. Also, this step could reduce the effect of outliers which can be magnified by running one step hierarchical or k-means algorithms alone.

The second step of TCP process re-groups the k' pre-clusters generated from step one to capture the multi-modal nature of dataset. The modified Kullback-Leibler divergence adapted to basic hierarchical clustering algorithm was used in this step. The usage of K-L distance is especially meaningful since many of the existing measure of distance used in hierarchical method are not suitable for measuring dissimilarity between distributions of clusters. Especially, the modified K-L distance (Larsen, 2002) derived under the assumption of Gaussian distributions of data within clusters matches with the implicit assumption of k-means algorithm used to generate the pre-clusters.

4. Simulation

The simulation of the proposed algorithm was implemented in Matlab 6.1 along with some conventional algorithms for comparison purpose.

Evaluated Datasets

In order to evaluate the effectiveness of our proposed TCP, three simulated datasets and six publicly known datasets were evaluated. The detailed information of the nine datasets is listed in Table 1 including the number of observations and clusters as well as the number of dimension. Since the first step of our TCP implements k-means type algorithms, the datasets with only continuous variables were used in the simulation.

Table1. Evaluated Datasets

DataName	Variable Dimension	Observation Number	Cluster Number
Triangle	2	600	3
Xours	2	600	3
Taeguk	2	800	2
Balance scale	4	625	3
Iris	4	150	3
Liver disorder	6	345	2
Sonar	60	208	2
Tokyo1	44	959	2
Waveform 21	21	500	3

The six datasets (Balance Scale, Iris, Liver Disorder, Sonar, Tokyo1, and Wave form 21) were obtained from publicly open sources of researches. The Triangle, Xours, and Taeguk data were generated to have the problematic distribution shapes such as multi-modal and non-gaussian distribution. The three generated datasets plotted in Figure 1 have dimension of two to allow the visual evaluation of the clustering process.

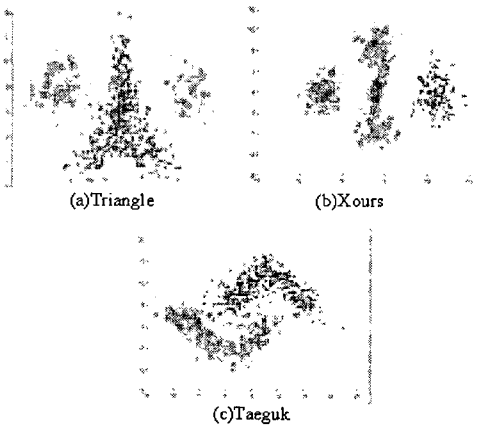


Figure 1. Three Generated Datasets

Rand Index

The performance of given clustering algorithm can be assessed by measuring the difference between the

clustering result and the actual answer of the cluster membership. Rand index was adapted as the measure of our TCP's performance. The simple criterion of Rand index tests the effectiveness of clustering algorithm when the actual target is known (Berkhin, 2002). The Rand index proposed by Rand (1971) selects a pair of objects and evaluates each object's class membership. The value a is the number of pairs of objects clustered to be in the same group and in fact, exist in the same clusters in the actual target answers. The value b is the number of pairs placed in the same class in the answer but clustered to be in different classes and c represents the number of pairs placed in the reverse way. Finally, d is the number of pairs of data points that are in different classes for both clustering result and the actual answers. The Rand index is then calculated as

$$RI\ (\textbf{Rand index}) = \frac{a+d}{a+b+c+d}.$$

The Rand index lies between 0 and 1 and has the value of 1 when the two sets of partitions agree perfectly. Obviously, the value closed to 1 would represent the better performed algorithm for the specific dataset.

Simulation Results

The proposed tandem clustering process was applied to nine datasets and the performance was measured using Rand index as tabulated in Table 2 and 3. The RI's of a conventional algorithm exhibiting the best performance were included as well for the purpose of comparison. The three generated datasets with distinctive multi-modal distributions produced the dramatic improvement in the algorithm performance. Also, the TCP quite effectively clustered the several real world datasets such as Sonar, Tokyo1 and Waveform1 increasing the RI value.

Table2. Simulation Result of three generated datasets

	Method giving the best result			TCP	
	k	Method Used	Rand Index	k'	Rand Index
Triangle	3	Hierarchical-Averag	0.7920	8	1
Xous	3	Hierarchical-Averag	0.7543	8	0.9947
Taeguk	2	K-mean	0.8118	10	0.9728

The proposed method was quite effective in reducing the time and complexity of computation as well especially when the large real world data was analyzed. Even though there was only slight increase for Balance Scale and no improvement for Iris and Liver Disorder in terms of RI value, the computation time was greatly reduced from the hierarchical methods obtaining the same clustering performance. The algorithm is robust for the datasets with large number of dimension and observations for which the hierarchical algorithms often fail to analyze due to the computation complexity.

Table3. Simulation Result of five real-world datasets

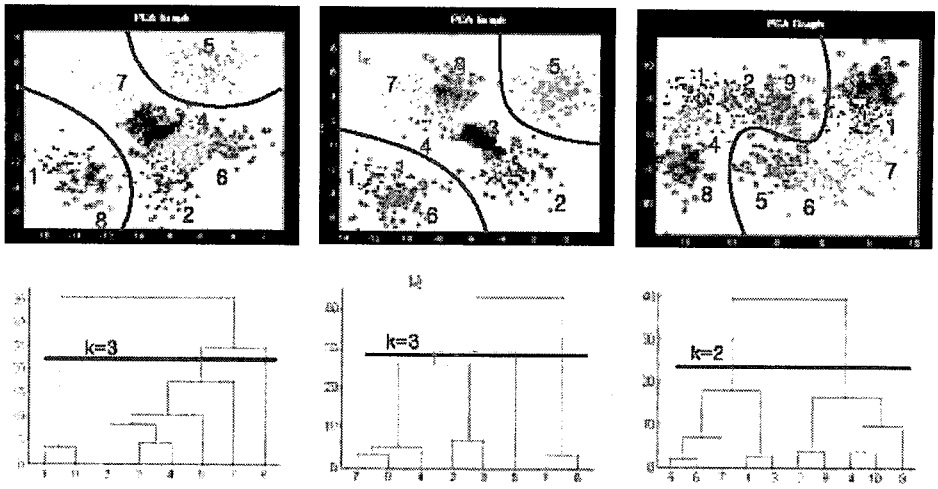
	Method giving the best result			TCP	
	k	Method Used	Rand Index	k'	Rand Index
Balance Scale	3	Hierarchical-Ward	0.6003	8	0.6138
Iris	3	Hierarchical-Averag	0.8923	6	0.8923
Liver disorder	2	Hierarchical-Singl	0.5104	12	0.5104
Sonar	2	Fuzzy k-mean	0.5032	12	0.6458
Tokyo1	2	Fuzzy k-mean	0.5987	4	0.7578
Waveform 21	3	Fuzzy k-mean	0.6842	8	0.7052

In addition, the clustering results of the three two-dimensional generated data are visualized in Figure 2. The pre-clusters generated are numbered on the PCA graphs and the corresponding dendrograms generated from the second step of hierarchical clustering are plotted as well. The numbers in the x-axis of the dendrogram matches the labels of pre-clusters on PCA plots and by horizontally cutting the dendrograms the datasets were successively clustered as indicated by the separated regions.

5. Conclusion

Simple and fast tandem clustering process suitable for clustering multi-modal distributions has been presented in this paper. Another important merit of this method beside its effectiveness at catching the multi-modal clusters would be the simplicity and the speed of analysis. The two major steps of the TCP constitutes of the most widely approved clustering algorithms such as k-means and hierarchical. This method tries to integrate the strengths and eliminate the weaknesses from two methods.

The testing on three generated multi-modal data and six real-world data demonstrates the superior performance of TCP. The clustering results were compared against those of several widely used clustering methods exhibiting an improvement either in the accuracy of cluster assignment or the time of computation.



(a) Taeguk (b) Triangle (c) Xours

Fig.7. PCA graphs of the clustering results using the propose TCP and corresponding dendograms

References

Ball, G.H., Hall, D.J., 1967.A clustering technique for summarizing multivariate data, Behavioral Sci. 12 153-155.

Berkin, P., 2002. Survey of Clustering Data Mining Techniques, Technical Paper, Accure Software, San Jose, CA.

Bezdek, J.C., 1974. Numereical taxonomy with fuzzy sets, J. Math Biol. 1 57-71.

Dash, M., Liu, H., Scheuermann, P., Tan K.L., 2003. Fast hierarchical clustering and its validataion, Data & Knowledge Engineering 44 109-138.

Defays, D., 1977. Efficient algorithm for a complete link method, The Computer Journal 20 364-366.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, Wiley-Interscience, New York, pp. 550-557.

Dunn, J.C., 1974. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybernet 3 32-57.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Verlag, pp. 461-466.

Larsen, J, Hansen, L.K., Have, A.S., Christiansen, T., Kolenda, T., 2002. Webmining: learning from the world wide web, Computational Statistics & Data Analysis 38 517-532.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 864-850.

Ripley, B.D., 1996. Pattern Recognition and Neural Network, Cambridge University Press, Cambridge.

Rousseeuw, P.J., Kaufman, L., Trauwaert, E., 1996. Fuzzy clustering using scatter matrices, Computational Statistics & Data Analysis 23 135-151.

Sibson, R., 1973. SLINK: An optimally efficient algorithm for the single link cluster method, Computer Journal 16 30-34.

Voorhees, E.M., 1986. Implementing agglomerative hierarchical clustering algorithms for use in document retrival, Information Processing and Management 22 (6) 465-476.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function, Journal Amer. Stat, Assoc. 58 235-244.