Indian Journal of Science and Technology, Vol 9(12), DOI: 10.17485/ijst/2016/v9i12/89282, March 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Comparative Study of Clustering Methods over Ill-Structured Datasets using Validity Indices

Sheik Faritha Begum[1]*, K. P. Kaliyamurthie[2] and A. Rajesh[3]

[1,2]Department of Computer Science and Engineering, Bharath University, Chennai, Tamil Nadu, India; sfaritha@gmail.com

[3]Department of Computer Science and Engineering, C. Abdul Hakeem College of Engineering and Technology, Vellore – 638052, Tamil Nadu, India; amrajesh73@gmail.com

Abstract

Objective: This paper discusses and compares the various clustering methods over Ill-structured datasets and the primary objective is to find the best clustering method and to fix the optimal number of clusters. **Methods:** The dataset used in this experiment has derived from the measures of sensors used in an urban waste water treatment plant. In this paper, clustering methods like hierarchical, K means and PAM have been compared and internal cluster validity indices like connectivity, Dunn index, and silhouette index have been used to validate the clusters and the optimization of clustering is expressed in terms of number of clusters. At the end, experiment is done by varying the number of clusters and optimal scores are calculated. **Findings:** Optimal score and optimal rank list are generated which reveals that the hierarchical clustering is the optimal clustering method. The optimum value of connectivity index should be minimum, silhouette should be maximum, dunn should be maximum. So by interpreting the results, the optimal number of clusters for the experimental dataset have been concluded as K=2 and the optimal method for clustering the given dataset is hierarchical. **Applications:** The experiment has been done over the dataset derived from the measures of sensors used in a urban waste water treatment plant.

Keywords: Clustering Methods, Ill-Structured Datasets, Optimization,Validity Indices

1. Introduction

Based on nature of domain attributes, Clustering methodology tends to identify homogeneous group of objects. The aim of clustering is to categorize or group the similar data items together in order to reduce the amount of data. All approaches of clustering face a common problem of interpreting the generated clusters. Some of the algorithms uses cluster shapes as a solution to the above mentioned problem, and those will assign the data to clusters of such shapes. Therefore, inferencing cluster shape attracts more attention rather than compressing the data set. So cluster analysis plays an important role in entire clustering process. Validation of the cluster analysis results must also be done. Hence the primary work of clustering process is to express the data patterns in the form of "meaningful" groups, which leads us to identify similarities and dissimilarities and also to derive some needed conclusions about them. The two basic questions which needs to be addressed in every typical clustering method are: a) The number of clusters originally present in the data and b) The quality of clusters formed, which means that the validation of clusters must be done while applying clustering technique[1]. Clustering is basically divided into two groups namely Partitional clustering and Hierarchical clustering. Hierarchical clustering iterates repeatedly by either dividing larger clusters into smaller ones, or by merging smaller clusters. The former one is termed as top down and latter is bottom up. The variation of clustering methods relies on the selection of larger cluster for splitting or in the

*Author for correspondence

selection of two small clusters for merging. Hierarchical clustering produces clusters in tree form which is so called dendrogram showing the relationship of clusters. Dendrogram can be cut at any level randomly, grouping of the data items into disjoint groups called clusters. On the other hand partitional clustering proceeds directly to divide the entire data into a set of dissimilar clusters. The most common criteria includes maximizing measure of similarity between the data patterns within each cluster, while maximizing the dissimilarity of objects resides in different clusters. Internal cluster validity indices like Dunn index[2], Connectivity index[3], Silhouette index[4] had been used to validate the resultant clusters.

2. Background and Related work

In paper[5] discussed and compared different validity measures of clustering on ill-structured datasets. A single clustering algorithm named K means has been employed and the result has been validated with many cluster validity indices and optimization of validation has been done. In our proposed approach we extend the above work by optimizing the clustering method suitable for clustering ill-Structured dataset. In paper[6] a specific distance measure is related with clusters produced by hierarchical approach. Two clustering methods have been developed based on the relationship above and proved that the methods produce clusters rapidly. Defining explicitly, one method produces optimal "connected clusters" and the next produces optimal "compacted clusters." In[7] Chameleon's key feature combines closeness and interconnectivity in order to find the atmost similar pair of clusters. It uses two-phase algorithm to find the clusters in the data set. In first phase, graph partitioning algorithm is used to cluster the data objects into many small relative subclusters. During second phase, original clusters are obtained by repeatedly combining these subclusters with the help of an algorithm.

Paper[8] discusses about the parallel algorithms and sequential algorithms for hierarchical clustering. Many distance metrics for parallel algorithms to execute hierarchical clustering has also been discussed. Various optimal algorithms are provided for various versions of hierarchical clustering's like the complete link, average link, median, centroid. In[9] discussed the fundamental concepts of clustering by surveying the well known clustering algorithms. It also focused on quality assessment of clustering results. It concerns about the features of the data set which are inherent. Cluster validity measures and approaches are reviewed. Paper[10] focused on surveying the different clustering techniques and classification of clustering algorithms into different categories have also been done. Importance of similarity measures have been proved. In[11] the authors studied and compared various clustering techniques. In[12] used two approaches namely Self Organizing map and Fuzzy C means to segment color images and compared the results with K-means Clustering.

3 Different Clustering over Waste Water Treatment Patterns

Most of the clustering algorithms had been employed over structured data sets. In this model we have performed different clustering over ill-Structured datasets.

3.1 Dataset Description

This dataset[13] derived from the measures of sensors used in a urban Waste Water Treatment Plant (WWTP). This domain was stated as an ill-structured domain. Since this proposed work can be extended to analyse the operational characteristics of WWTP, only the output values among those attributes are taken for clustering process.

Number of instances: 527
Number of Attributes: 07
Attribute Information:
All attributes are numeric and continuous
N. Attributes.

1 PH-S	(output pH)
2 DBO-S	(output Biological demand of oxygen)
3 DQO-S	(output chemical demand of oxygen)
4 SS-S	(output suspended solids)
5 SSV-S	(output volatile suspended solids)
6 SED-S	(output sediments)
7 COND-S	(output conductivity)

3.2 Architecture

The daily measures of sensors in a waste water treatment plant is retrieved and loaded as dataset. Data preprocessing has been done to remove missing values. Although hierarchical algorithm can be applied on all type of attributes, here only numerical attributes are considered so that we will be able to compare with Kmeans algorithm

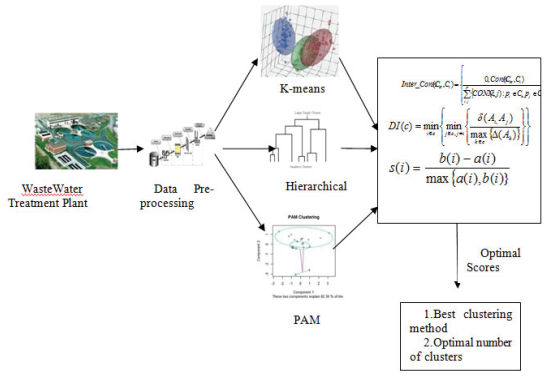

Figure 1. Architecture diagram for clustering applied over waste water treatment patterns.

which runs only on numerical data. So Filters are used to remove the attributes other than numeric. Three types of clustering are performed over the pre-processed dataset to obtain the clustered model and the clustered set by fixing the parameters which includes number of clusters, similarity measures. In order to measure the performance, the ratio of intra cluster distance and inter cluster distance is measured along with calculation of dunn index and silhouette index value. Finally optimal scores are retrieved which reveals the best clustering method and optimum number of clusters for the given dataset.

3.3 Hierarchical

Hierarchical method finds the similarity of all the data objects in a cluster to the cluster centroid which is given by the $Sim(C)=\Sigma_{d\in C}$ cosine(d,c) where d is a Data object in cluster C and c is the centroid of cluster C. The selection of pair of clusters to be merged is done by identifying pair of clusters which gives small difference in similarity.

The algorithm possesses the following steps:

1. Initializing each item as a cluster forming N clusters for N items.The distances between the items in each cluster is termed as the distances between the clusters
2. Determine the clusters which are closer and merge them into a single cluster, decreasing the cluster count by one.
3. Calcuate distances between the newly formed cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster, so that size of the final cluster is N.

3.4 K-Means

The following explains how the k-means algorithm works. First we introduce the term called centroid which is generally the center of a cluster. Using similarity measures like Euclidean distance or other similar measure, the centroid is defined as a point where average value of each attribute is taken as a point for those corresponding attribute for a cluster.

The objective function[4]

$$J = \sum_{j=1}^{k}\sum_{i=1}^{x}\left\|x_i^{(j)} - c_j\right\|^2 \quad (1)$$

The algorithm consists of the following steps:

Input: D (instance set), K (number of clusters)
Output: K number of clusters

1: Randomly select K cluster centers among the data objects.
2. Compute the mean of data items with respect to center initialized
3: until termination condition is not satisfied do
4: Reassign the data objects to the closest cluster center.
5: Recalculate the mean and change the cluster centers based on re-assignment.
6: end while

3.5 PAM

In many clustering algorithms, clusters are always characterized by different kinds of structural computes k representative objects, called medoids. Medoids are nothing but the objects of a cluster in the case of minimal average dissimilarity to all the objects. After identifying *the* set of medoids, all the objects are iteratively assigned to the nearest medoid. That is, object b is put into the cluster $v_{b,}$ while medoid mv_b is nearer than any other medoid $m_{c.}$

The objective function should be minimized by k representative objects, by calculating the total sum of dissimilarities of all objects:

$$\text{Objective function} = \Sigma d(b, mv_b) \quad (2)$$

4. Cluster Validity Indices used for Validation

4.1 Connectivity Index[3]

We define connectivity index by two quantities: Intracluster connectivity and intercluster connectivity. Intracluster connectivity is termed as within-cluster scattter and Intercluster connectivity is termed as between-cluster separation measure.

$$Intra\_Connectivity = \sum_{k}^{K} Intra\_Connectivity(C_K) / K \tag{3}$$

$$Intra\_Connectivity(C_K) = \frac{\sum_{a,b}^{P} \{CADJ(a,b) : p_a, p_b \in C_k\}}{\sum_{a,b}^{P} \{CADJ(a,b) : p_a \in C_k\}} \tag{4}$$

$$Inter\_Conn(C_K, C_l) = \left\{ \frac{0, Conn(C_K, C_l)}{\sum_{a,b}^{P} \{CONN(a,b) : p_a \in C_k, p_b \in C_l\}} . \right\} \tag{5}$$

where K is the number of clusters, P is the number of prototypes, Intra Conn(C_K) is the ratio of number of those data samples in C_k, CADJ(a,b) is cumulative adjacency matrix and it is nonsymmetric, and CONN(a,b) is the symmetric matrix. Increasing number of clusters will decrease Interconnectivity when the clusters are not divided over normal cluster boundaries .Always Interconnectivity depends at the cluster boundaries particularly on the connections of the prototypes.

4.2 Dunn Index[2]

The Dunn index is defined as

$$DI(c) = \min_{i \in c} \left\{ \min_{j \in c, j \neq i} \left\{ \frac{\delta(S_i, S_j)}{\max_{k \in c} \{\Delta(S_k)\}} \right\} \right\} \tag{6}$$

Where

$\delta(S_i, S_j) = \min\{d(x_i, x_j) \mid x_i \in S_i, x_j \in S_j\}$

$\Delta(S_k) = \max\{ d(x_i, x_j) \mid x_i, x_j \in S_j\}$

S_j is the set containing the data points assigned to the i_{th} cluster and d is a distance function.

4.3 Silhouette Index[4]

This index uses cohesion measure. The cohesion is calculated based on the distance between all the data objects within the same cluster and the separation is calculated based on the nearest neighbor distance. It is calculated as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{7}$$

$$a(i) = \frac{1}{n_k - 1} \sum_{i' \in I_k, i' \neq i} d(M_i, M_{i'}) \tag{8}$$

$$b(i) = \min_{k' \neq k} \partial(M_i, C_{k'}) \tag{9}$$

$$\partial(M_i, C_{k'}) = \frac{1}{n_k - 1} \sum_{i' \in I_k, i' \neq i} d(M_i, M_{i'}) \tag{10}$$

5. Results

5.1 Hierarchical Clustering: Result and Discussion

In this type of clustering first initialize each cluster to be a singleton. When more than one cluster exists then, find the two closest cluster by finding between those two. The average similarity between all pairs considering single member from each of the clusters gives the similarity between clusters. Similarity matrix gives the similarity between any two elements. Now replace those two clusters by their union. Repeat the above steps until no similarities exists between clusters.

Table 1. Statistics of generated clusters using Hierarchical

Internal Indices	K=2	K=3	K=3	K=4	K=5
Connectivity	2.9290	4.5079	7.4369	12.7413	17.9841
Dunn	0.5079	0.0575	0.0575	0.0212	0.0212
Silhouette	0.8191	0.6006	0.3861	0.5057	0.5000

5.2 K-Means

Defining k centroids for each cluster is the basic idea in k means. These centroids will not reside in same cluster during iteration since different location causes different result. Next the objects are associated to the nearest centroid. Early grouping is completed when no object is pending. Then K new centroids are calculated based on the grouping formed on preceding step. Again find the objects closes to the nearest centroids and relocate the

Table 2. Statistics of generated clusters using K-means

Internal Indices	K=2	K=3	K=3	K=4	K=5
Connectivity	9.9389	35.8881	32.5437	28.6583	49.4861
Dunn	0.0103	0.0057	0.0066	0.0158	0.0193
Silhouette	0.5602	0.5198	0.5149	0.5311	0.4724

objects between clusters. Finally, the K-means clustering algorithm is mainly used for minimizing the distance measure called sum of squared distances. Generally the value of *k* is a smaller integer, such as 2, 3, 4 or 5 Table 2.

5.3 PAM

The main idea here is to define or initialize some set of objects randomly as medoids. Adding an object to the set is done for those sum of the distances to all other objects in the set is minimal. Next, select an object i which is to be included in the set, calculate the dissimilarity between the object j from the non-medoid set and closest object in the medoid set. If the distance $D_j > d(i,j)$,then object i is included in the set. The above step is done for improving the quality of the cluster Table 3.

Table 3. Statistics of generated clusters using PAM

Internal Indices	K=2	K=3	K=3	K=4	K=5
Connectivity	8.0956	29.1647	34.3698	55.6060	66.5337
Dunn	0.0072	0.0074	0.0056	0.0062	0.0051
Silhouette	0.5539	0.4674	0.5182	0.4713	0.4576

5.4 Optimal Scores

As we have discussed in the Section 4, the optimum value of connectivity index should be minimum, silhouette should be maximum, dunn should be maximum, so by interpreting the results, the optimal number of clusters for the given dataset has been concluded as **K**=2 and the optimal method for clustering the given dataset is hierarchical**.** Also Rank Aggregation has been done to prove the above result. R tool has been used efficiently both for clustering and validation. Finally optimal rank list have been proposed as follows Table 4.

Table 4. Optimal Scores of validity indices for varying number of clusters

	Score	Method	Clusters
Connectivity	2.9290	Hierarchical	2
Dunn	0.5079	Hierarchical	2
Silhouette	0.8191	Hierarchical	2

The optimal rank list is:

Hierarchical-2 Hierarchical-3 Hierarchical-4 Pam-2 K means-2
Algorithm: CE
Distance: Spearman
Score: 0.168816

The result of rank aggregation concludes that the hierarchical clustering with 2 numbers of clusters is the best clustering method followed by hierarchical with 3 clusters and hierarchical with 4 clusters.

6. Conclusion

In this paper we have used K-means, Hierarchical and Pam clustering to cluster the waste water treatment datasets. The three internal cluster validity indices are used as fitness value to find the optimal number of clusters. As a result optimal score and optimal rank list are generated which reveals that the hierarchical clustering is the optimal clustering method. In future other clustering algorithms can be used along with different validity measures.

7. References

1. Dubes RC, Jain AK. Clustering techniques: The user's dilemma. Pattern Recognition. 1976; 8(4): 247–60.
2. Baridam B, Barileé. More work on K -Means clustering algorithm: The dimensionality problem. International Journal of Computer Applications. 2012; 44(2): 23–30.
3. Demir KT, Merényi E. A validity index for prototype-based clustering of data sets with complex cluster structures. IEEE Transactions on systems, Man, and Cybernetics—Part B: Cybernetics. 2011; 41(4): 1039–53.
4. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J.Cybernetics. 1973; 3(3): 32–57.
5. Begum SF, Rajesh. Comparative study of validity indices for clustering over ill-structured datasets. IJAER. 2015; 10(1):1879–90.
6. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967; 2(3): 241–54.

7. Karypis G, Han EH, Kumar V. Chameleon:Hierarchical clustering using dynamic modeling .Computer. IEEE; 1999; 32(8): 68–75.
8. Olson CF. Parallel Algorithms for Hierarchical Clustering. Parallel computing. 1995; 21(8):1313–25.
9. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. Journal of Intelligent Information Systems. 2001; 17(2):107– 45.
10. Popat SK, Emmanuel M. Review and comparative study of clustering techniques. International Journal of Computer Science and Information Technologies. 2014; 5(1): 805–12.
11. Zafar MH, Ilyas M. A Clustering Based Study of Classification Algorithms. International Journal of Database Theory and Application. 2015; 8(1): 11–22.
12. Arumugadevi S, Seenivasagam V. Comparison of Clustering Methods for Segmenting Color Images. Indian Journal of Science and Technology. 2015; 8(7):670–77.
13. Bache K, Lichman M.UCI Machine Learning Repository, University of California. School of Information and Computer Science: Irvine, CA. Available from: http://archive.ics.uci.edu/ml.Date Accessed:11/12/2013.