Validation of *k*-means and Threshold based Clustering Method

Mamta Mittal, R.K.Sharma, V.P.Singh
Thapar University,Patiala (Punjab) INDIA
Corresponding Author Email: mittalmamta@rediffmail.com

Abstract

Data mining isa process of extracting interested hidden information from large databases. It can be applied on many databases but kind of patterns to be found is specified by various data mining techniques.Clustering is one of the data mining techniques that partitions database into clusters such that data objects in same clusters are similar and data objects belonging to different cluster are differ.Researchers have developed many algorithms for clusteringbut this paper focus on well known partitioning based technique i.e k-means with threshold based clustering technique. k-means algorithm partition the database into k clusters where k is the user defined parameter, beside this it is sensitive to outliers and intial seed selection.Threshold based clustering is the another method which generates the clusters automatically based on threshold value. To assess quality of clustering obtained from both techniques several validity measures and validity indices have been applied on synthetic data.By the experimentations and comparisions of the clustering results, it has been obsereved that clusters obtained from the threshold based technique are more separated and compact which indicates good clustering.

Keywords: Data mining,Clustering, k-means, Validity measures, Validity indices.

1. Introduction

The exponential growth of data emerges the need of technique which can transform this huge amount of data into useful information. Data mining [8] is most suitable for this need. It analyzes databases and classifies/partition it so that any organization take decision based on this classification and can improve their future plans. There are many techniques of data mining exists by which we can detects the hidden patterns in the databases. These techniques are fall under two categories *i.e.* supervised and unsupervised learning.

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structurein a collection of unlabeled data. It separates data objects in different clusters with goal of maximizing intra and minimizinginter cluster similarity. For this a number of clustering methods are available in literature such as partitioning based clustering, hierarchical based clustering, density based , grid based, fuzzy based and probabilistic based methods[8]. In this paper well known partitioning based technique *i.e. k*-means is compared

with the threshold based technique.*K*-means algorithm is popular because of its simplicity[10]. It iteratively partition the dataset into *k*clusters where *k* clusters and *k* initial centroids are provided beforehand, beside thisit has many limitations due to which authors have discussed threshold based technique also known as single pass clustering [5]. It is not so much popular as *k*-means but authors have focused its advantages over k-means and evaluated quality of both clustering algorithm with the help of popular validity measures(separation and compactness) and validity indices(*DB* and*DUNN* index) on synthetic datasets.

The paper is organized as follows:In section 2, related work is discussed. In section 3, performance evaluation of both methods on well known validity measures and validity indices has been presented.And in last section, work carried out is concluded.

2. Related Work

Today many international organizations produce more information in a week than many people could read in a life time. This situation is even more effective due to widespread use of internet. Now we are drowning in information but starved for knowledge . We are in fact inundated with data in most fields but do not have enough trained humananalysts available who are skilled at translating all of this data into knowledge, and thence up the taxonomy tree into wisdom. So data mining is a knowledge discovery process, which is used to discover knowledge from large amount of data, stored either in databases, data warehouses, or other information repositories. Data mining involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, high performance computing, pattern recognition, neural network, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis. It becomes popular due to growing data volume, limitation of human analysis and low cost of machine learning. It automates prediction of trends and behaviour as well as discovery of previously unknow patterns through various techniques.

Clusteing a primary data analysis technique in machine learning, data mining, pattern recognition, image analysis and bioinformatics. It is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait.Generally the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or

more objects belong to the same cluster if one defines a concept commonto all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.Depending on the clustering technique, clusters can be expressed in different ways:

- Identified clusters may be exclusive, so that any object belongs to only one cluster.
- They may be overlapping; an object may belong to several clusters.
- They may be probabilistic, whereby an object belongs to each cluster with a certain probability.
- Clusters might have hierarchical structure, having crude division of objects at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

In this paper authors discussed well known partitioning based clustering algorithm known as *k*-means [11,12]with another partitioned based clusreing known as threshold based clusteing[5,13].

The *k*-Means partitional clustering algorithm is the simplest and most commonly used algorithm to cluster or to group the objects based on attributes/ features into *k* number of cluster,where *k* is positive integer number and defined by user beforehand. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The basic steps of *k*-means clustering are:

1. Take number of cluster *k* and we assume first *k* objects as the centroid or center of these clusters or randomly pick *k* centroids.

2. Iterate until *stable* (= no object move group)

a) Determine the distance of each object to the centroids

b) Group the object based on minimum distance and update the centroids

Thus the purpose of *k*-means clustering is to classify the data or partition the data.

But it has several limitation like sensitive to initial seed selection, outlier sensitivity [9] and finds a local optimum and may actually miss the global optimum. Its time complexity is O(*nkl*) where *n* is number of data objects and *k* is number of clusters and *l* is nunber of iterations. These limitations motivate

the reseacher and they developed number of variant of *k*-means algorithm[1,4,14].

A threshold based clustering algorithm is employed to improve the chances of finding the global optimum and not sensitive to outlier. Basic steps are:

1. Assign the first data object to the first cluster.

2. Iterate until all objects are selected

i)Select next object. Determine the minimum distance between selected object and centroid of existing clusters. Compare the distance with threshold value, group the object into existing cluster or form a new cluster if it is not within threshold limit.

Time complexity of this approach is O(*nk*) and loop terminates in finite steps means when *n* objects are selected.In case if outlier is there it will be considerd as a separated cluster because of threshold limit. Only first cenroid is randomly selected, rest clusters formation is dependent on object's distance with the exixting centroid so it do not trap in local optimum problem.

Both techniques clusters the dataset but which one is the better partitioning? To answer this several validity measures and validity indices exists in literature[6,7].In this paper separation and compactness are validity measures being used. Further *DB* index [2] and *Dunn* index [3] are used as the validity indices so that quality of both techniques can be judged accurately.

3. Performance

Clustering will be good if the clusters are maximum separated from each other and objects within clusters should be more and more close(compact) to the centroid. Experiment is carried out for the three methods of separation named as Single linkage, Complete linkage and Centroid linkage methods.Compactness can be measured by variance of the cluster. Further, two validity indices, Dunn index and*DB*index are used by the *k*-means and the threshold based algorithms for different values of *k*. The experimentation is based upon fifteen datasets each containing five hundred *2-D* data objects. These data objects are generated randomly in the range 100 to 499 in both dimensions. The numbers of clusters generated vary between four and forty nine. In *k*-means algorithm, the number of clusters (*k*) acts as an input parameter while in the threshold based algorithm, threshold value (T_{th}) acts as an input

parameter. In *k*-means algorithm, initially, randomly *k* data objects are taken as centroids of *k* clusters on the other hand, in the threshold based algorithm, an object is chosen randomly and is assigned to first cluster. This will also act as the centroid of that cluster. Fig. 1, Fig. 2 and Fig. 3 depict single linkage, complete linkage and centroid linkage separation methods respectively. From these figures it can be observed that threshold based algorithm gives a better separation than *k*-means algorithm as value of separation is large in each seprartion method.

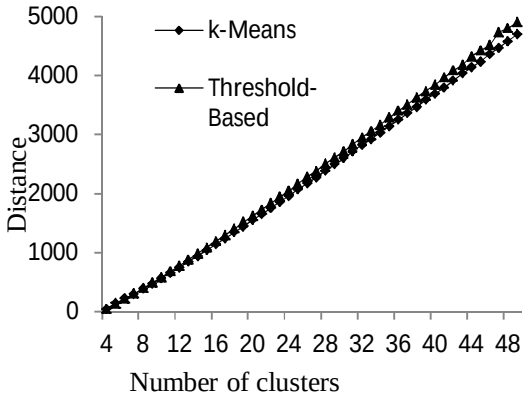

Fig. 1 Separation using single linkage method

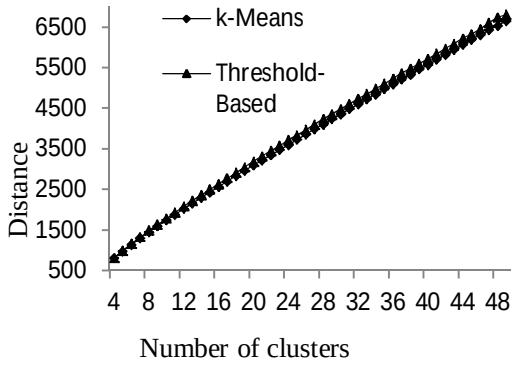

Fig. 2 Separation using complete linkage method

Compactness of clusters is also an important characteristic of cluster validation. Here, it is calculated using the variance of the clusters. The algorithm with maximum compactness (minimum average distance) is considered as better clustering approach. Compactness of *k*-means and threshold based algorithms are shown in Fig. 4. depicts that threshold based algorithm generates compact clusters. Further to draw conclusion which is better approach, *k*-means and threshold based algorithms are evaluated on *Dunn* index and *DB* index .*Dunn* index is defined as the ratio of the separation to compactness which indicates that if value of *Dunn* index is

large means clusters are well separeted. The results of this comparision is shown in Fig. 5. *DB* index is defined as the ratio of compactness to separation and depicted in Fig. 6. Here, small value of *DB* index depicts that clusters are more compact. Thus threshold based algorithm provides good clustering results instead of *k*-means algorithm when compared with *Dunn* and *DB* indexs . Thus threshold based technique will be better clustering approach if one may carefully choose the threshold value.

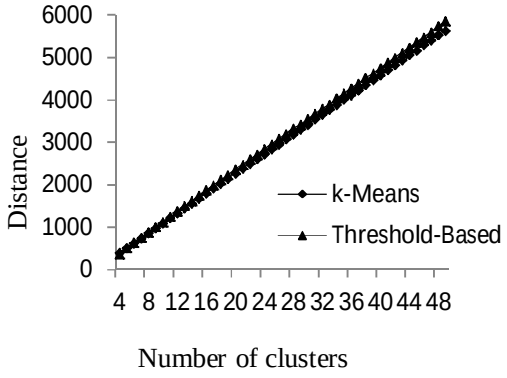

Fig. 3 Separation using centroid linkage method

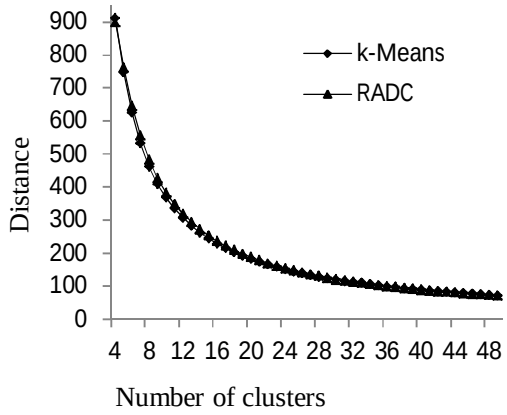

Fig. 4 Compactness using variance of clusters

Researcher can focus on this methodogy also as it has manyadvantages over *k*-means except in this threshold value shoud be given as an input parameter.

4. Conclusions

Validation of clustering algorithms is one of the most important issues in cluster analysis in order to justify the selection of right candidate algorithm for clustering. It aims identification of the method that best fits the underlying data. In this paper, *k*-means and threshold based methods are \evaluated on synthetic data with popular validity measures and validity indices. The experimentation carried out in this work uncovers thatthreshold

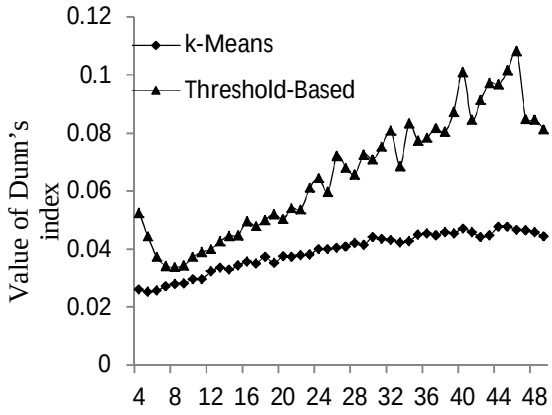

Fig. 5 Evaluation of Dunn Index

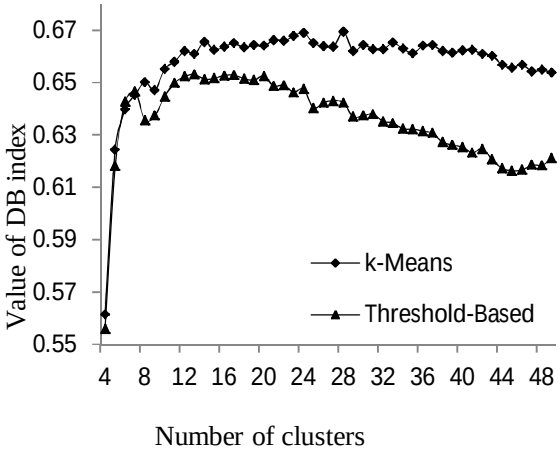

Fig. 6 Evaluation of DB Index

based algorithm perform better than *k*-means algorithm for popular validity measures and validity indices. In future, it can be considered as a popular methodology of partitioning based clustering methods if more research work is carried out in the direction of threshold value.

References

[1] Cheung, Y.M., (2003), A New Generalized K-Means Clustering Algorithm, *Pattern Recognition Letters*, Elsevier, vol. 24(15), 2883–2893.

[2] Davies, DL, Bouldin, D.W. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*vol. 1, No2. pp. 224-227.

[3] Dunn, J. C. 1974. Well separated clusters and optimal fuzzy partitions, *J. Cybern,*vol.4, pp. 95-104.

[4] Fahim, A.M., Salem, A.M., Tokey, F.A., & Ramadan, M., (2006), An efficient enhanced k-means clustering algorithm, *Journal of Zhejiang University Science A*, vol 7(10), pp. 1626-1633.

[5] G. Salton. The SMART Retrieval System. *Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.*

[6] Halkidi, M., Vazirgiannis, M., &Batistakis,I., (2000),Quality scheme assessment in the clustering process, *Proceedings of PKDD, Lyon, France.*

[7] Halkidi, M., Vazirgiannis, M., &Batistakis, Y,(2001), OnClustering Validation Techniques, *Journal of Intelligent Information Systems,* vol. 17, No. 2-3, pp. 107-145

[8] Han, J., &Kamber, M., (2006), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

[9] Hautamaeki, V., Cherednichenko, S., Kaerkkaeinen, I., Kinnunen, T., &Fraenti, P., (2005), Improving *k*-means by outlier Removal, *SCIA, LNCS 3540*, pp. 978-987.

[10] Jain, A. K.,(2010), Data clustering: 50 years beyond *k*-means, *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666.

[11] Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28, 129–137. Originally as an unpublished Bell laboratories Technical Note (1957).

[12] MacQueen, J.B., (1967), Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of 5th BerkleySymposium on Mathematical Statistics and Probability*, vol. 1: Statistics, pp281-297.

[13] Mittal, M., Singh, V. P., & Sharma, R. K., (2011), Random automatic detection of clusters. *In Image Information Processing (ICIIP),IEEE International Conference*, pp 1-6.

[14] M. EmreCelebi, Hassan A. K., Patricio A. V., (2013), A comparative study of efficient initializationmethods for the k-means clustering algorithm. *Expert Syst. Appl.* 40(1), pp. 200-210.