International Journal of Innovative
Computing, Information and Control
Volume 11, Number 1, February 2015
 ISSN 1349-4198
pp. 375–386

MODIFIED SINGLE PASS CLUSTERING WITH VARIABLE THRESHOLD APPROACH

Mamta Mittal[1,*], Rajendra Kumar Sharma[2] and Varinder Pal Singh[1]

[1]Computer Science and Engineering Department
[2]School of Mathematics and Computer Applications
Thapar University
Patiala, Punjab, India
*Corresponding author: mittalmamta@rediffmail.com

Received March 2014; revised July 2014

Abstract. *Data mining is the process of extracting hidden, interesting, non-trivial, potentially useful and previously unknown information from large databases. Clustering is one of the data mining techniques that aims to separate dissimilar objects and group similar objects in the database. There are a number of clustering methods available in literature. In this paper, authors have focused on partitioning based methods. Most popular partitioning based algorithms, k-means and k-medoid, require the number of clusters to be generated as an input parameter. Another partitioning based algorithm, Single Pass Clustering (SPC), requires a threshold similarity value as an input parameter for clustering. In this paper, a modified SPC algorithm is proposed which also uses threshold similarity value but it is not an input parameter, rather, it is a function of data objects to be clustered. To assess performance of proposed approach, several clustering validity measures have been applied on k-means, SPC and the modified SPC algorithms. The stimulated experiments described in this paper confirm good performance of the modified SPC. It is also observed that actual number of clusters is generated when modified SPC is applied on real datasets.*
Keywords: Data mining, Clustering, k-means, Single pass clustering, Validity measures

1. **Introduction.** Advances in storage technology and the development of information technology in the field of internet search, web mining, image processing, etc. have created very high volume of datasets. Now, the problem is how to find potentially useful information from these datasets. Data mining [4,8], an analytical tool, helps to solve this problem.

Clustering [8,11] is one of the tools of data mining which falls under the category of unsupervised learning. It is the process of separating and grouping the data objects in a way that similar objects are in common groups and dissimilar objects are in different groups. Clustering has been a widely studied problem in knowledge discovery [8], pattern recognition and pattern classification [4,23]. It can be achieved using different methods, namely, partitioning based methods, hierarchical based methods, density based methods, grid-based methods, fuzzy clustering, and probabilistic methods [4,8,11,23]. Partitioning based methods, k-means [16] and k-medoid [13], divide the data objects into k partitions where k is the number of clusters. However, finding the appropriate value of k is a complicated task without prior knowledge about databases. The value of k can also be known by applying hierarchical clustering as a pre-processing step. Another partitioning based clustering method, *Single pass clustering* (*SPC*) [18,21], does not require the number of clusters as an input parameter but it requires a threshold similarity value as an input

parameter. This threshold similarity value is the maximum similarity value between two objects.

Requirement of the user's prior knowledge about the number of clusters or threshold similarity value or hierarchical clustering as an additional pre-processing step motivates the authors of this paper to find an efficient clustering technique. Thus, a modified *SPC* algorithm is proposed that uses a threshold similarity value which is the function of data objects left to be clustered instead of being defined by the user.

In this work, clusters are generated using k-means, *SPC* and modified *SPC* methods. Further, these methods are evaluated and compared for validity measures (separation and compactness) available in literature [7].

Rest of this paper is organized into four sections. In the next section, related work in this field is presented. In the third section, a new partitioning based clustering algorithm is proposed. In the fourth section, performance evaluation of k-means, *SPC* and proposed clustering algorithm is carried out for popular validity measures. And, in the last section, the work carried out in this paper is concluded.

2. **Related Work.** Data mining is the process of analyzing a transactional database to forecast the future trends of a business organization. This is one of the essential activities before taking managerial decisions like introducing/modernizing/discontinuing or grouping the products to analyze the sales. To achieve this objective, organizations employ a procedure to mine their databases which is known as Knowledge Discovery in Databases (*KDD*) [8]. Data mining is one of the steps in this process. Data mining tools predict future trends and behaviors, allowing businesses to take profitable and knowledge-driven decisions.

In Data mining, data analysis techniques can be classified into two categories, namely, supervised methods and unsupervised methods [23]. In the supervised methods, system learns from example patterns. It involves only labeled data means training patterns with known category labels while the latter in volves only unlabelled data and optimizes the maximum similarity among similar objects and minimizes similarity among dissimilar objects by using an objective function.

Cluster analysis is an important part of human life. It mainly focuses on distance based approach. It clusters the dataset of size n into k clusters ($k \leq n$) in a way that each data object d_i, $1 \leq i \leq n$, belongs to one cluster k_j, $1 \leq j \leq k$ and each cluster k_j has at least one data object. Each cluster is represented by a prototype. Depending on the kind of prototype, one can distinguish between k-means and k-medoid. In the k-means algorithm [16], the prototype, called the centroid, is the mean value of all objects belonging to a cluster. In the k-mediod algorithm [13], also called PAM (Partioning Around Medoid), the prototype, called the mediod, is the most centrally located object of a cluster. In both methods, the number of clusters (k) needs to be specified beforehand. Many researchers have improved the performance of k-means algorithm [2,3,5,9,14,24] so that it can be applied to a variety of databases of different sizes in an effective and efficient manner.

After that, A. K. Jain [10] has given a survey paper on the popularity of k-means algorithm since last 50 years and objected out some of emerging and useful research direction. Research is still going on how one can better decide initial cluster centers. Further, E. Murat *et al.* [19] proposed a method in which initial centroids are not taken randomly. In this firstly, mean object of the whole database is calculated. Then, a data object with the largest distance to mean object is selected as the first initial centroid; second initial centroid is a data object at largest distance from first initial centroid; similarly, k^{th} initial centroid is a data object at largest distance from $k - 1^{th}$ initial centroid. Objects, once, selected as an initial centroid is not considered again. D. Reddya

and K. J. Prasanta [20] have given a novel method to select the initial centroids with the help of Voronoi diagram constructed from the given set of data objects. The initial centroid will be those objects which lie on the boundary of Voronoi circles having highest radius.

Scope of research is not limited, currently, it is shifted on the similarity measures like multi-view object based similarity measure, clustering on uncertain data using probability distribution similarity and Mahalanobis distance similarity measure [1,12,15,17]. D. T. Nguyen *et al.* [15] presented a paper in which similarity between two documents d_i and d_j is determined by not only cosine similarity measure but also considered the direction and distance of vectors. Now, it is possible to use more than one object of reference and have more accurate assessment of how close or distant a pair of objects is.

The existing k-means techniques to cluster uncertain data heavily rely on geometric distances between objects and do not take into account the probability distribution of objects. So, B. Jiang *et al.* [12] used the well-known Kullback-Leibler divergence as the similarity measure which can capture distribution between uncertain objects in both the continuous and discrete cases.

I. Melnykov and V. Melnykov [17] worked on the Mahalanobis distance similarity measure in which initial estimation of covariance matrices is quite complicated. They developed an initialization procedure that aims to gather the information about covariance matrices by identifying a group of objects with a high concentration of neighbours that represent the core of the selected cluster. Thus, these objects provide a rough covariance matrix estimate. This estimate is improved further by updating the membership of objects in the cluster according to a probability coverage criterion. Then, it can be used for calculating Mahalanobis distance for the rest of the objects. The performance of entire strategy depends upon the quality of covariance matrix estimation.

Recently, R. Scitovski and K. Sabo [22] proposed a technique about what can be done in case the data object occurs on the border of two or more clusters. In this technique, unit weight is associated with all data objects, except the data object which belongs to two or more clusters. The weight of shared data object is uniformly divided in two or more clusters. Then, for each participating clusters centroid and objective function value is calculated. On the behalf of that shared object becomes a part of the cluster showing better clustering.

Thus, researchers have provided numerous variants of k-means as discussed above. However, in this paper authors are giving emphasis on existing technique, known as single pass clustering (threshold based clustering). The simplest and fastest one seems to be the "single pass" method proposed by G. Salton [21] for document clustering. In this method each document is processed once and is either assigned to one (or more, if overlap is allowed) of the existing clusters, or it creates a new cluster based on threshold value. Single pass method is named as Threshold based algorithm, in which Euclidean distance is used as a similarity measure. It is a good alternate method over k-means only if one opts right threshold value. Opting the right threshold value is the limitation of single pass clustering method but it has lots of advantages over k-means as k needs not to be specified beforehand, it is not sensitive to outliers, its running time is also less than running time of k-means and in this technique, and clusters are formed on the behalf of threshold value so it does not suffer from the local minima problem.

In the real world, there exists a constraint on the user knowledge of number of clusters to be generated, especially when it is being done first time on a dataset. In addition, k-means method does not remove the outlier; moreover, it places the outlier in nearest cluster owing to degradation of quality of clustering. However, *SPC* algorithm generates clusters automatically without previous knowledge of numbers of clusters to be generated; it uses

a threshold similarity value as an input parameter. In case of outliers, threshold similarity value helps to place them in a separate cluster instead of being the part of the nearest cluster. In this paper, a modified *SPC* algorithm is proposed that uses threshold similarity value which is a function of data objects left to be clustered, rather than provided by the user. This methodology will serve as a right technique in partitioning based clustering algorithms.

The outcomes of three clustering algorithms, namely, k-means, *SPC* and modified *SPC* algorithm can differ from each other for the same dataset. Quality of these clusters can be judged by the popular validity measures [6]. The objective of cluster validation is to find the partitioning that best fits the underlying data. Broadly, validity measures are of two types – Separation and Compactness. Separation is the measure of dissimilarity of objects of one cluster to the objects of another cluster which should be maximum and Compactness is the measure of dissimilarity among objects of a cluster which should be minimum. Separation can be measured mainly by three methods – single linkage (closest distance between two objects of two different clusters), complete linkage (farthest distance between two objects of two different clusters) and centroid linkage (distance between centroids of two clusters). Compactness can be measured by two methods – centroid based (average distance of all the objects to the centroid of the cluster) and averaged paired distance (average distance of all pairs of objects of the cluster). In this paper, these validity measures have been used to evaluate clustering obtained by k-means, *SPC* and the proposed modified *SPC* algorithms.

3. **Proposed Method.** In this section, modified *SPC* partitioning based clustering algorithm is proposed. The algorithm proposed in this work revolves around the proposition of a threshold similarity value which is not the user defined parameter; instead, it is the function of data objects left to be clustered. In this algorithm, a data object is selected randomly and assigned to first cluster. Rest of the data objects are selected randomly which will belong to either one of the existing clusters or will form a new cluster. For this, the distance between selected object and centroid of nearest cluster is determined using Eculidean distance formula. Subsequently, the selected object will be the part of an existing cluster or form a new cluster based on the comparison between calculated distance and the threshold similarity value. In our work, threshold similarity value is taken as average paired distance of all data objects left to be clustered as shown in Equation (1).

$$T_{th} = f(A) = \frac{2}{(n) * (n+1)} * \sum\nolimits_{i,j=1,j>i}^{n} a_{ij} \tag{1}$$

where, A, the set containing paired distance, a_{ij}, the distance between objects i and j; n is the number of objects left to be clustered. Threshold similarity value keeps on changing till all objects are clustered.

The proposed algorithm consists of the following general steps:

a) Initially, set the threshold similarity value, say, T_{th}, as a function of data objects to be clustered using Equation (1).

b) Select a data object randomly and assign it to first cluster.

c) Select the next data object again randomly. This data object will belong to either one of the existing clusters or to a new cluster. The data object is assigned to a cluster with the help of distance between the data object and centroid of already formed clusters; and the threshold similarity value. If the distance between the data object and the centroids is more than the threshold similarity value, a new cluster is created and data object is assigned to this cluster; otherwise the data object is assigned to one of the already formed clusters whose centroid has minimum distance from the data object.

d) If any existing cluster is updated due to entrance of new data object then update its centroid. Otherwise, assign the new data object as a centroid of new cluster.

e) Update threshold similarity value T_{th} as a function of data objects left to be clustered using Equation (1).

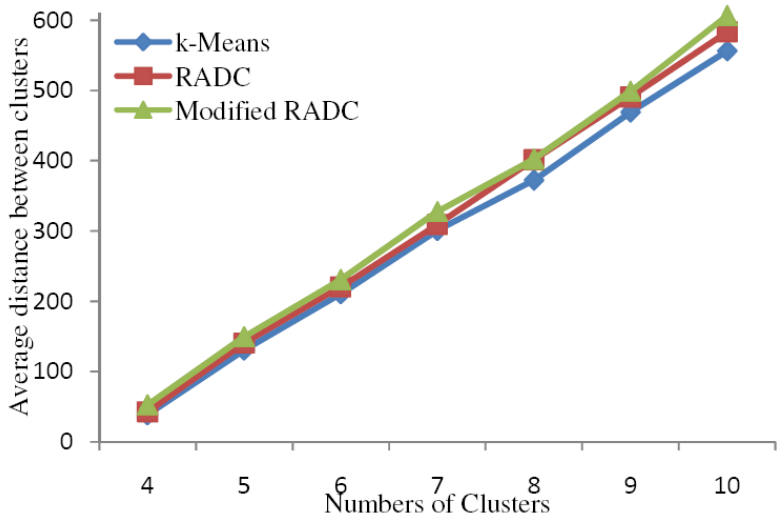
f) Repeat steps c), d) and e) until all the data objects are clustered.

Figure 1 consists of proposed clustering algorithm. In this algorithm, threshold similarity value (T_{th}) is used to determine that objects will be a part of an existing cluster or form a new cluster. This algorithm overcomes the drawback of specifying the number of clusters (k) and threshold similarity value for k-means and SPC algorithms, respectively. Like SPC, it also considers outliers in the separate clusters.

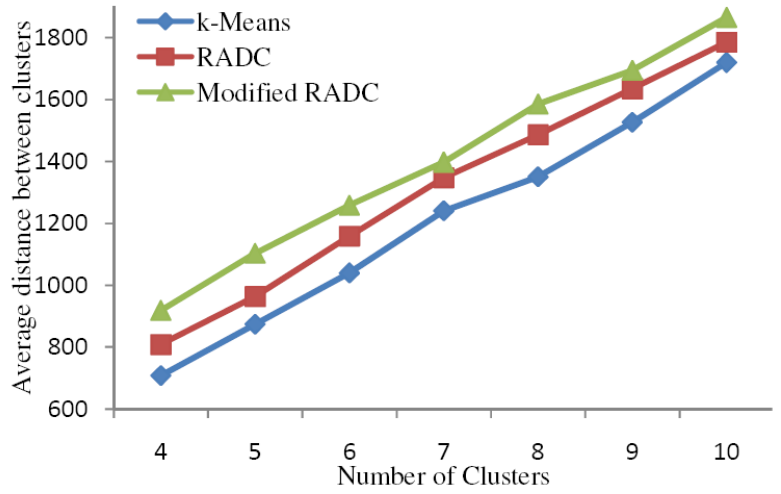
```
//Input:A set D = {d1, d2, d3, ... , dn} of n objects to cluster and a set
      A = {aij | aij = distance between object i and object j for 1 ≤ i, j ≤ n and j > i}.
//Output:A set K = {k1, k2, k3, ... , kk} of k subsets of D as final clusters and a set C = {c1, c2, c3, ... , ck} of
centroids of these clusters.
Proposed clustering algorithm (D,A)
1.  let m = 1;
2.  km = {p|∃p ∈ D}; //Randomly choose any object from D, say p
3.  K = {km};
4.  cm = p;
5.  C = {cm};
6.  Tth = f(A); //defined in eq. (1).
7.  for each object q' not clustered
8.      do if q' is chronologically greater than p
9.          then A = A − {apq'};
10.         else A = A − {aq'p};
11. for each random object q ∈ {D} − p
12. do   for each centroid r ∈ C
13.          do sr = d(q,r); //d(q,r) is the distance between point q & r
14.      sj = min (s1, s2, s3, ... , sm);
15.      if(sj ≤ Tth)
16.      then kj = kj ∪ q;
17.          Update centroid cj for cluster kj.
18.      else m = m + 1;
19.           km = {q};
20.           K = K ∪ {km}
21.          cm = q;
22.      n = n − 1;
23.      Tth = f(A); //defined in eq. (1).
24.      for each object q' ∈ {D} not clustered
25.      do if q' is chronologically greater than q
26.          then A = A − {aqq'};
27.          else A = A − {aq'q};
```

FIGURE 1. Modified single pass clustering algorithm

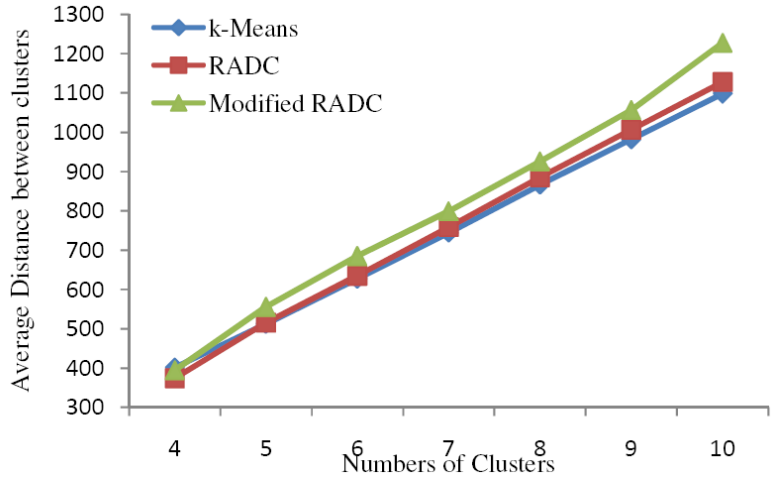
4. **Performance Evaluation.** In this section, five experiments for k-means, SPC and modified SPC have been carried out. In the first experiment, fifteen artificial datasets each containing five hundreds 2-D data objects are taken. These data objects are generated randomly in the range 100 to 499 in both dimensions. The value of k is taken in the range from four to ten for k-means algorithm and value of threshold is taken in the range from 50 to 75 for SPC algorithm. Quality of clustering is assessed by means of separation methods and compactness methods. Figure 2 and Figure 3 show the comparison among



(a) Single linkage method

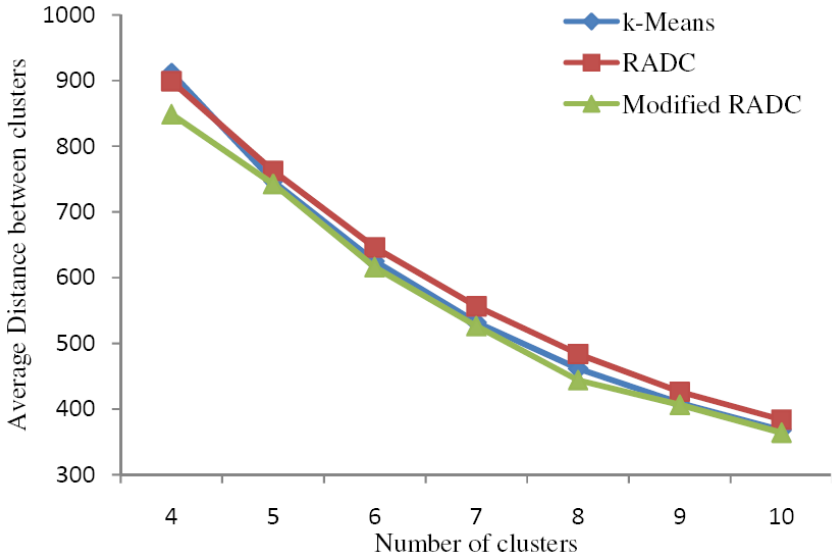


(b) Complete linkage method

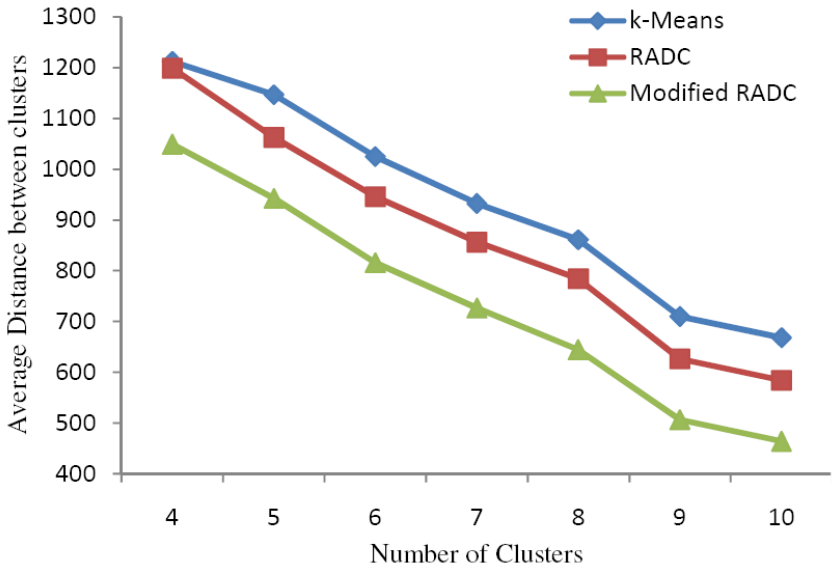


(c) Centroid linkage method

FIGURE 2. Separation comparison among k-means, *SPC* and modified *SPC* algorithm



(a) Centroid based method



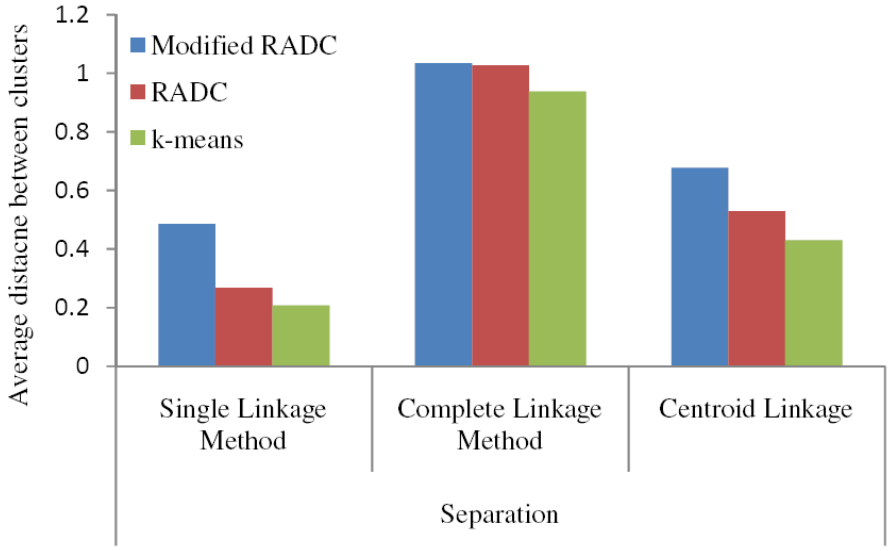
(b) Average paired distance method

FIGURE 3. Compactness comparison among k-means, SPC and modified SPC algorithm

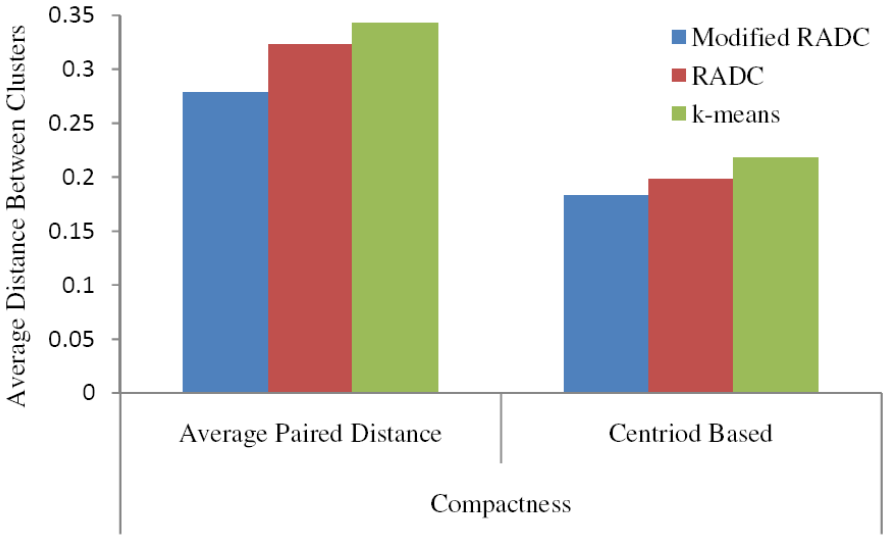
TABLE 1. Characteristics of datasets

Sr.no.	Name of Dataset	# Objects	# Attributes	# Clusters
1	Ecoli	336	8	8
2	Iris	150	4	3
3	Seeds	210	7	3
4	Wine	178	13	3

k-means, SPC and modified SPC algorithms for separation and compactness, respectively. From both figures, it is observed that clusters generated by modified SPC algorithm are well separate and compact than the clusters generated by k-means and SPC algorithms.



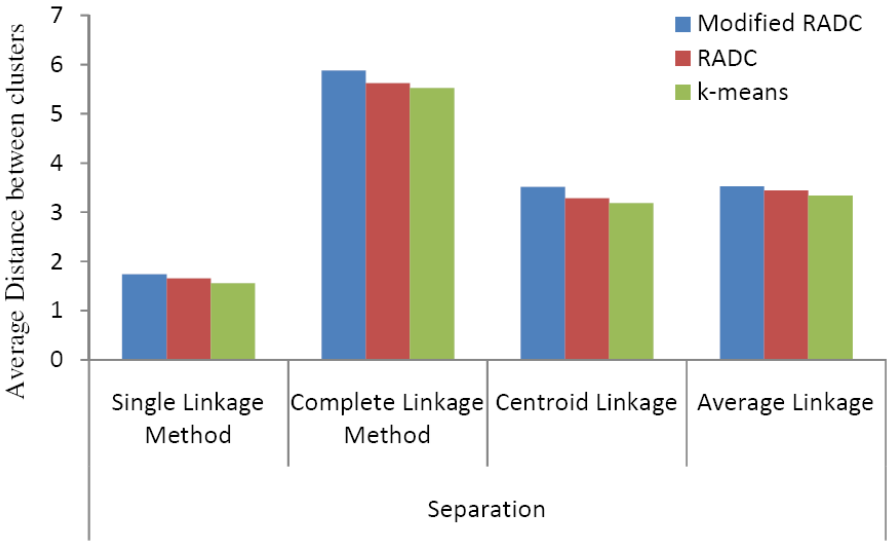
(a) Separation based comparison



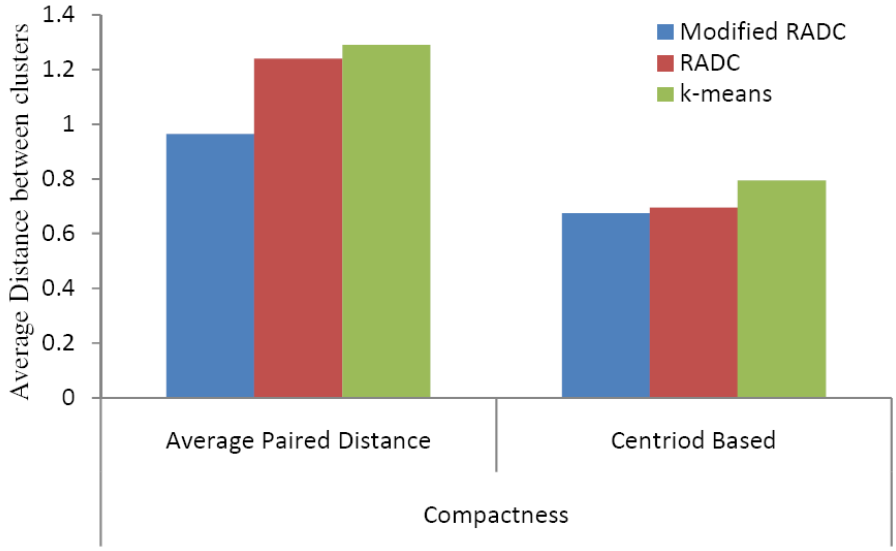
(b) Compactness based comparison

FIGURE 4. Comparison on Ecoli dataset

Rests of four experiments are performed on four real datasets – Ecoli, Iris, Seeds and Wine available in UCI repository. Table 1 shows the characteristics of all datasets used. As mentioned earlier, in k-means algorithm, the number of clusters (k) acts as an input parameter while in the proposed algorithm, threshold similarity value (T_{th}) acts as a criterion to partition the dataset into unknown k numbers of clusters. In k-means algorithm, initially, randomly k data objects are taken as the centroids of k clusters; on the other hand, in the proposed algorithm, mean of the distance of all pairs of objects is taken as an initial threshold similarity value which varies during the execution of algorithm depending upon the number of objects left to be clustered as describe in Equation (1). It is observed that proposed methodology generates the actual number of clusters present in the database.



(a) Separation based comparison



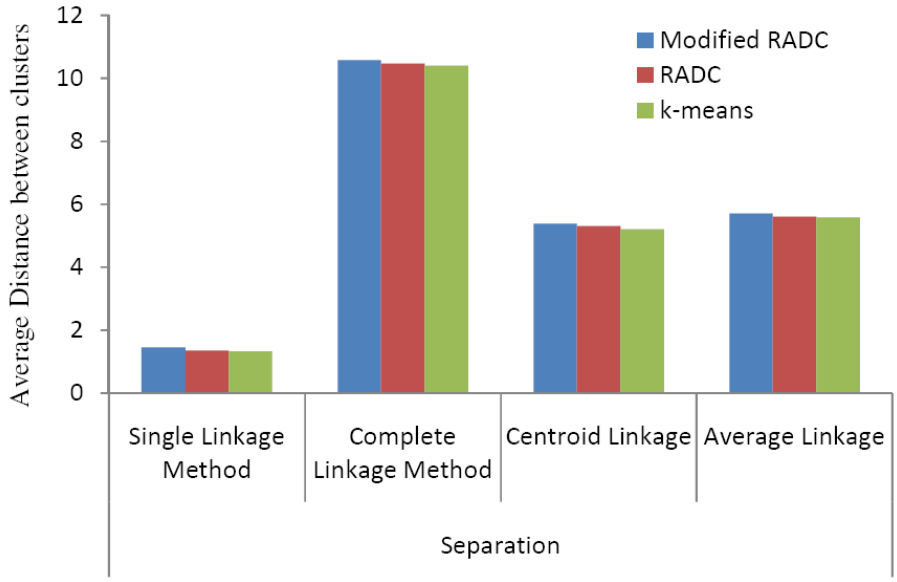
(b) Compactness based comparison

FIGURE 5. Comparison on Iris dataset

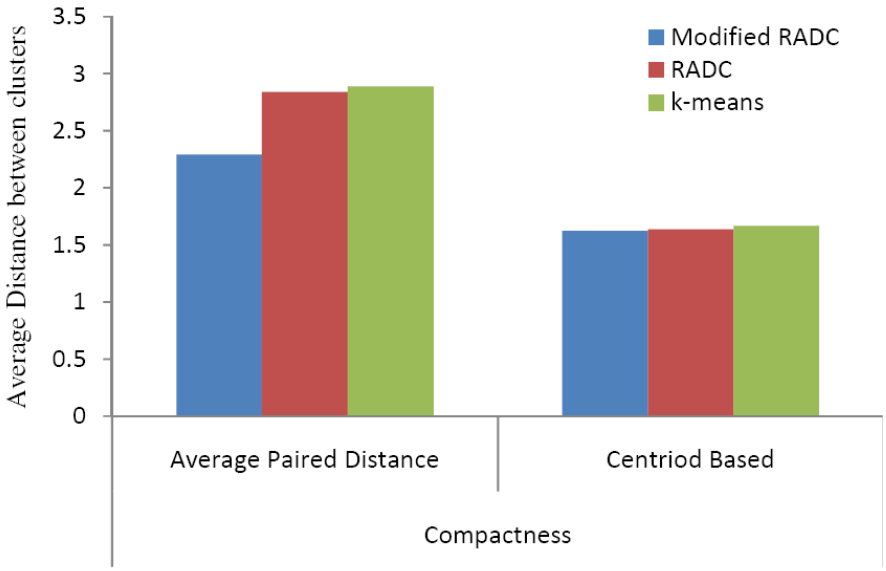
For the real datasets, Ecoli, Iris, Seeds and Wine, a comparison for all validity measures among k-means, *SPC* and modified *SPC* is presented in Figure 4, Figure 5, Figure 6, and Figure 7, respectively.

From Figure 4, Figure 5, Figure 6, and Figure 7, it is observed that the stimulated experiments confirm good performance of the modified *SPC* in terms of well separate and compact clusters.

5. **Conclusions and Future Direction.** Performance evaluation of clustering algorithms is one of the most important issues in cluster analysis in order to justify the selection of right technique for clustering. In this paper, a modified *SPC* based clustering algorithm has been proposed which uses a threshold similarity value as a function of data objects left to be clustered. Moreover, it relinquishes the requirement of user defined threshold similarity value and the requirement of user defined number of clusters.



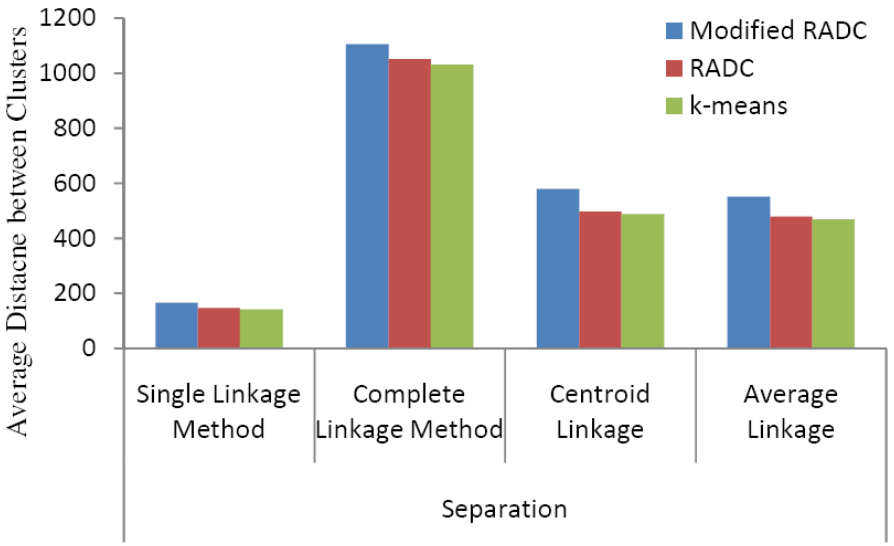
(a) Separation based comparison



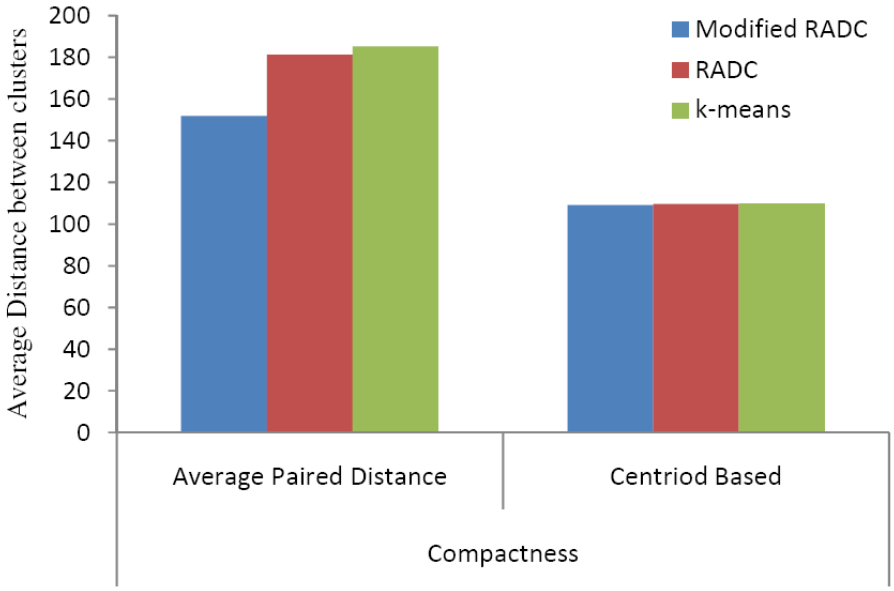
(b) Compactness based comparison

FIGURE 6. Comparison on Seeds dataset

Further, this method has been compared with existing k-means and *SPC* methods on real datasets, and validated for major validity measures. The experiments carried out in this work reveal that modified *SPC* algorithm generates actual number of clusters present in the dataset and performs better than k-means and *SPC* algorithms. In future, new methods to evaluate the threshold similarity value can be identified.



(a) Separation based comparison



(b) Compactness based comparison

FIGURE 7. Comparison on Wine dataset

REFERENCES

[1] B. Maria-Florina, B. Avrim and G. Anupam, Clustering under approximation stability, *Journal of the ACM*, vol.60, no.2, 2013.

[2] P. S. Bradley and U. M. Fayyad, Refining initial objects for k-means clustering, *Proc. of the 15th International Conference on Machine Learnining*, pp.91-99, 1998.

[3] D. Pelleg and A. Moore, X-means: Extending k-means with efficient estimation of the number of clusters, *Proc. of the 17th International Conf. on Machine Learning*, pp.727-734, 2000.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley Sons, New York, 1973.

[5] A. M. Fahim, A. M. Salem, F. A. Tokey and M. Ramadan, An efficient enhanced k-means clustering algorithm, *Journal of Zhejiang University Science A*, vol.7, no.10, pp.1626-1633, 2006.

[6] M. Halkidi, M. Vazirgiannis and I. Batistakis, Quality scheme assessment in the clustering process, *Proc. of PKDD*, Lyon, France, 2000.

[7] M. Halkidi, M. Vazirgiannis and Y. Batistakis, On clustering validation techniques, *Journal of Intelligent Information Systems*, vol.17, nos.2-3, pp.107-145, 2001.
[8] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.
[9] V. Hautamaeki, S. Cherednichenko, I. Kaerkkaeinen, T. Kinnunen and P. Fraenti, Improving k-means by outlier removal, *SCIA, LNCS*, vol.3540, pp.978-987, 2005.
[10] A. K. Jain, Data clustering: 50 years beyond k-means, *Pattern Recognition Letters*, vol.31, no.8, pp.651-666, 2010.
[11] A. K. Jain, M. N. Murty and P. J. Flyn, Data clustering: A review, *Computing Surveys*, vol.31, no.3, 1999.
[12] B. Jiang, J. Pei, Y. Tao and X. Lin, Clustering uncertain data based on probability distribution similarity, *IEEE Trans. on Knowl. Data Eng.*, vol.25, no.4, pp.751-763, 2013.
[13] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
[14] S. S. Khan and A. Ahmad, Cluster center initialization algorithm for k-means algorithm, *Pattern Recognition Lett.*, vol.25, pp.1293-1302, 2004.
[15] D. T. Nguyen, C. Lihui and K. C. Chee, Clustering with multiviewpoint – Based similarity measure, *IEEE Trans. on Knowl. Data Eng.*, vol.24, 2012.
[16] S. Lloyd, Least squares quantization in PCM, *IEEE Transactions on Information Theory*, vol.28, pp.129-137, 1982.
[17] I. Melnykova and V. Melnykova, K-means algorithm with the use of Mahalanobis distances, *Statistics and Probability Letters*, vol.84, pp.88-95, 2014.
[18] M. Mittal, V. P. Singh and R. K. Sharma, Random automatic detection of clusters, *IEEE International Conference Image Information Processing*, pp.1-6, 2011.
[19] E. Murat, C. Nazif and S. Sadullah, A new algorithm for initial cluster centers in k-means algorithm, *Pattern Recognition Letters*, vol.32, pp.1701-1705, 2011.
[20] D. Reddya and K. J. Prasanta, Initialization for k-means clustering using Voronoi diagram, *Procedia Technology*, vol.4, pp.395-400, 2012.
[21] G. Salton, *The SMART Retrieval System*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
[22] R. Scitovski and K. Sabo, Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters, *Knowledge-Based Systems*, vol.57, pp.1-7, 2014.
[23] S. Theodoridis and K. Koutroubas, *Pattern Recognition*, Academic Press, 1999.
[24] Z. K. Rizman, An efficient k'-means clustering algorithm, *Pattern Recognition Letters*, vol.29, pp.1385-1391, 2008.