Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 47 (1): 171- 174 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

An Efficient Partitioning Clustering Algorithm

S. NIZAMANI[++], K. DAHRI*, Sehrish. NIZAMANI, Sarwat. NIZAMANI**, G, LAGHARI***

Department of Information Technology, Sindh University Campus Mirpurkhas, Pakistan

Received 18th November, 2014 and Revised 17th January 2015

Abstract: Clustering is an unsupervised classification method, used in various disciplines, with a goal of organizing objects into different groups. Various methods and algorithms exist for clustering which attempt to find better clustering results. In clustering, choosing the centroids is a very sensitive concern as it is the essential element to do the clustering. This paper presents a novel partitioning clustering method that generates better clusters which is efficient in terms of Rand Index and Purity compared to traditional K-means algorithms. Experiments are performed on various datasets from UCI machine learning repository to obtain clusters and results show that the proposed method generates better clusters in terms of Purity and Rand Index.

Keywords: clustering algorithm, cluster analysis, unsupervised classification

1. INTRODUCTION

The task of grouping elements having similar characteristics is known as clustering. Similar objects have to be the part of same cluster, while those having dissimilarities could be enlisted in different cluster (Mirkin, 1998). The data elements are organized in groups (clusters) in such a way, that the distance between two elements of the same cluster is minimal whereas the distance is greater between two elements from different clusters. Clustering is particularly useful to organize the data into relevant groups with respect to similarities and dissimilarities of attributes of different data elements where no prior information is available to distinguish the data elements. Clustering falls into the category of unsupervised classification that is, usually, applied by domain experts to categorize the scattered and unlabeled data. The clustering is applicable in a multitude of disciplines including data mining (Fayyad *et al.,* 1996), data compression (Gersho and Gray, 1992), pattern recognition (Duda and Hart, 1973), artificial intelligence and applied statistics.

Clustering algorithms are broadly categorized into two classes, namely; the hierarchical and the partitioning clustering algorithms (Jain and Dubes, 1973). In hierarchical clustering, each object or cluster is gradually embedded in a larger cluster until only one cluster is obtained (Han and Kamber, 2006). Partitioning algorithms partition the dataset into k clusters in a way that the objects in one cluster are similar to each other contrary to the objects in another cluster.

Partitioning clustering algorithms involve two phases. In the first phase, centroids are determined for each cluster and the second phase involves putting data items into appropriate cluster based on some distance function. Various partitioning clustering techniques have been proposed so far to obtain better clustering results. This paper presents a novel partitioning clustering algorithm which generates better clusters in terms of Rand Index and Purity compared to the basic K-means (MacQueen, 1967) algorithm.

2. RELATED WORK

A number of partitioning clustering algorithms exist, which apply distinct, and sometimes overlapping, techniques depending upon user needs and / or the desired output to obtain clustering.

K-means (MacQueen, 1967) is undoubtedly the best-known clustering algorithm for its easiness and effectiveness. It is the most intuitive and popular clustering algorithm, which uses an iterative refinement approach for finding the cluster memberships. It chooses k initial centroids randomly, and decides the association of each data item by estimating the distance of the item from previously selected initial centers. It iteratively calculates the mean and initializes new centroid values. This cycle continues until all data items stop moving from their clusters. It produces local optimal solution for each iteration step by deciding the membership of data elements to group/cluster, assuming that the current model is accurate. It compares the previous and current model until both results are close where its iterative refinements cease.

[++]Corresponding Author: saad.nizamani@usindh.edu.pk
*Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan
**Department of Computer Science, Sindh University Campus Mirpurkhas, Pakistan
***Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan

?Part of this paper is taken from MS thesis entitle "Effective and Efficient Clustering Algorithm" by the first author.

K-means algorithm is extremely sensitive to the initial cluster centroids, that different clustering groups may be produced for the same dataset. Recalculating the mean for every iteration step slightly slows the clustering process. Owing to the iterative procedure, k-means is computationally expensive. Amongst several other disadvantages, k-means does not deal with outliers.

Keeping in view the limitations of k-means, several authors have presented improved variations of k-means algorithm which produce better results with different datasets(Maitra *et al.,* 2010) . Fuzzy clustering (Nock and Nielsen, 2006; Bezdek, 2012) is one of the variations of k-means that applies fuzzy logic in determining centroids. A variation proposed in article (Maitra *et al.,* 2010) presents a systematic way for initializing centroid values in k-means, K-means++ (Arthur, 2009) is another variation of k-means that uses trees as filtering to expedite each k-means step (Kanungo *et al.,* 2002).

Nazeer and Sebastian (2009) endeavor to find better initial clusters so that overall results for k-means become more stable. The method calculates range and then sorts data based on the range. The dataset is divided into k equal parts and the mean values of each part are selected as initial centroids. The rest of the steps are similar to that of k-means algorithm.

Erisoglu *et al.*. (2011) attempt to find the initial cluster centers in two axes: the main axis and second axis. The main axis is an attribute having maximum value of coefficient variance and the second axis is selected from the column having the smallest coefficient correlation value. Once the values for both axes are obtained, the mean values of both axes are calculated to find the farthest element using some distance metric. The most distant element from the mean values is selected as a first centroid. Similarly the second centroid is obtained as the most distant element from the first centroid. The third centroid is obtained by finding the most distant element from the second centroid and adding the distances 1 and 2 to choose the third centroid. This method of finding centroids continues until desired number of clusters is achieved.

3. MATERIALS AND METHODS

In order to achieve better clustering results and improve the efficiency of clustering algorithm, the novel method of clustering is proposed which determines the initial centroids of the clusters according to data distribution. In this method, the entire dataset is sorted first by quicksort algorithm. The sorted dataset is divided into K equal number of parts (clusters). Then the mean is calculated for each of the clusters. The means are selected as centroids for clusters.

It is significant to note that data elements in the dataset may also have multiple attributes. For these datasets, the attribute with the maximum value of standard deviation is determined. Initially, the standard deviation of each attribute is computed and then the attribute with maximum standard deviation is determined. The entire dataset is then sorted based on the attribute with maximum standard deviation.

Once the centroids are obtained for each cluster, the data elements are put in appropriate cluster having less distance from the cluster centroid. The steps involved in the proposed algorithm are listed below.

Algorithm

Input: Dataset D={d_1, d_2, d_3, …,d_n} [dataset comprising of n number of elements]. K [Number of cluster]

Output: K Cluster Groups.

Step I: Calculate the standard deviation of all the attributes.

Step II: Choose the attribute that has the greatest value of standard deviation.

Step III: Sort the entire dataset according to that attribute which has the highest value of standard deviation.

Step IV: Distribute the dataset into K equal chunks.

Step V: Estimate the arithmetic mean of every part and take those values as centroids.

Step VI: Calculate the distance of every element from all the centroids.

Step VII: Place the object into cluster that has minimum distance from its centroid.

The proposed algorithm determines the initial centroids in more efficient manner based on proper distribution of data. As the data elements are already sorted in each cluster, the algorithm converges quickly and results into organizing data elements into clusters in a relatively less time.

4. RESULTS AND DISCUSSION

The proposed algorithm is tested for its accuracy on the benchmark datasets from UCI Machine Learning Repository (Kanungo *et al.,* 2002). The datasets used include: Iris[1] (Nazeer and Sebastian, 2009), Wine[2] (Zhao and Karypis, 2002), breast-cancer[3] (Rand, 1971) and seeds[4]. Iris dataset is often considered to be the benchmark dataset for testing the clustering results.

[1] UCI Repository of Machine learning Databases, [Data file], Retrieved from: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/

[2] UCI Repository of Machine learning Databases, [Data file], Retrieved from: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/wine/

[3] UCI Repository of Machine learning Databases, [Data file], Retrieved from: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer/

[4] UCI Repository of Machine learning Databases, [Data file], Retrieved from: http://archive.ics.uci.edu/ml/datasets/seeds

The tools used in order to apply the proposed algorithm include R the data mining tool[5], Eclipse with Java code and Microsoft Excel. R is used in this research to calculate the values of Rand Index and Purity by using predefined functions and to compare the results of various algorithms and datasets are taken in Excel format and some basic manual calculation of standard deviation, partitions and distance formulas are done in MS Excel, Eclipse is used as platform to implement the algorithm in Java to automated the process of clustering.

Evaluation Criteria

Two comparison measures are used to evaluate the clustering results and efficiency of clustering method. The comparison measures include: Purity (Zhao and Karypis, 2002) and Rand Index (Rand, 1971).

Purity

Purity is used to estimate the consistency of the clusters, i.e. the quantity of objects a cluster contain in a particular class. It is calculated by assigning each cluster to the class that is most common in that cluster. Then assignment accuracy is calculated by the number of correctly assigned members divided by total number of elements. It is the percentage of most dominant category in the cluster. It usually ranges between 0 and 1. An average purity is calculated by taking means of all clusters. The consistency of a cluster may be estimated using Purity i.e. the quantity of objects a cluster contain in a particular class (Subhashini and Kumar, 2010). It can be considered as a rate of classification with the supposition that all prototypes of clusters are expected to be part of actual dominant class for that cluster. In an accurate cluster that contains elements from a specific category, the value of purity would be 1. Generally, the greater value for purity implies better results.

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |w_k \cap c_j|$$

Where C is a set of classes $\{c_1, c_{2\ldots}, c_j\}$; Ω is a set of clusters $\{w_1, w_2 \ldots w_k\}$; and N is the total number of data points in the dataset[6].

Rand Index

Rand Index is traditionally used as the measure of similarity between two data partitions. The value of Rand Index ranges from 0 to 1, where 0 indicates no agreement on any pair of points by two clusters and 1 indicates that the data clusters are identical. Rand Index is similarity-measuring technique used to find out the resemblance between two clusters and accuracy. It may be computed using equation given below:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Where

a= Elements in S(s1 and s2) are part of X and in the same part Y

b= Elements in S(s1 and s2) are in different part of X and are in different part of Y

c= Elements in S(s1 and s2) are in same part of X and are in different part of Y

d= Elements in S(s1 and s2) are in different part of Y and are in same part of X

The results for proposed clustering method and original k-means algorithm are obtained using different datasets based on two evaluation criteria: Purity and Rand Index. As k-means algorithm creates different clusters at every iteration step, we use only two random iterations. The experimental results presented in tabular form in **(Tables 1 and 2).**

Table 1: Rand Index values obtained from proposed algorithm and k-mean algorithms.

	Rand Index		
Dataset	**Proposed Algorithm**	**K-MEANS1**	**K-MEANS2**
Iris	0.9123	0.7182103	0.8922595
Wine	0.7205	0.9169682	0.6998032
SeedS	0.8844	0.8538619	0.881203

Table 2: Purity values obtained from proposed algorithm and k mean algorithms.

	Purity		
Dataset	**Proposed Algorithm**	**K-MEAN1**	**K-MEAN 2**
Iris	0.9266	0.666	0.906
Wine	0.7134	0.938	0.657
Seeds	0.9047	0.876	0.9

Cluster membership results for various dataset are given in **(Table 3 and 4)**. It can be observed from Tables 3, 4 that the proposed algorithm results in better clusters, which is more effective and defines clusters with appropriate number of memberships in each cluster.

Table 3: Cluster membership result of proposed method on Iris, Wine and Seeds datasets

Dataset	**Cluster 1**	**Cluster 2**	**Cluster 3**
Iris	50	53	47
Wine	59	69	50

Table 4: Cluster membership result of proposed method on Breast Cancer dataset

Breast ue	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
Tissue	17	19	17	45	0	8

[5]The R project for statistical computing, R data mining tool [Computer Software], Available at: http://www.r-project.org/

[6] http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

It can be observed from Tables 1 and 2, the proposed clustering method has improved results for Purity and Rand Index compared to the k-means algorithm. For Iris and seeds datasets, the values of Purity and Rand Index are comparatively better than two iterations of k-means algorithm. For wine dataset the proposed method produces better results for Purity and Rand Index compared to 2^{nd} iteration of k-means, where as k-means outperforms the proposed method in 1^{st} iteration. However, the proposed algorithm gives stable results for every execution contrary to k-means algorithm with varying results. Also the proposed method generates improved and consistent clusters compared to k-means algorithm.

5. CONCLUSION

In this paper, an effective and efficient clustering algorithm is presented and experimentally tested. The proposed algorithm finds the initial clusters more effectively than the traditional k-means algorithm. The centroids are chosen using sorting based approach. The cluster memberships are assigned by calculating the distances, which makes this process simpler in comparison with the traditional k-means algorithm.

The proposed algorithm results in better clusters, which is more effective and defines clusters with appropriate number of memberships in each cluster. This algorithm is simple in comparison with k-means and easier to implement with less computational complexity.

There are fewer limitations associated with this research. The proposed algorithm chooses the centroids using sorting based approach, which works well with small datasets but may cause overhead while sorting huge amount of data. A new mechanism may be discovered which may reveal the more efficient way of choosing centroids.

Another limitation is based on testing this algorithm on limited datasets. However, the results of those specified datasets give efficient and better consequences in comparison with traditional k-means algorithm.

REFERENCES:

Arthur, D. (2009). *Analyzing and improving local search: k-means and icp*, Ph.D. Dissertation. Stanford University, Stanford, CA, USA. Advisor(s) Rajeev Motwani. AAI3363935.

Bezdek, J. C. (2012). Pattern Recognition with Fuzzy Objective Function Algorithms. Springer US, 20-May-2012, US.

Duda, R. O., P. E. Hart, (1973). *Pattern classification and scene analysis, (Illustrated)*, New York: Wiley, ISBN: 0471223611, 9780471223610.

Erisoglu, M., N. Calis, and S. Sakallioglu, (2011). A new algorithm for initial cluster centers in k-means algorithm, *Pattern Recognition Letters*, vol. *32, no.*14, 1701-1705,

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-40.

Han, J., and M. Kamber, (2006). *Data Mining, Southeast Asia Edition: Concepts and Techniques* (2nd ed.).Southeast Asia: Morgan Kaufmann.

Kanungo, T., D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, (2002). An efficient k-means clustering algorithm: Analysis and implementation", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. *24*, 7, 881-892.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* vol. 1, No. 14, 281-297.

Maitra, R., and A. D. Peterson,.(2010). A systematic evaluation of different methods for initializing the K-means clustering algorithm", *IEEE Transactions on Knowledge and Data Engineering*, 522-537.

Mirkin, B. (1998). *Mathematical classification and clustering: From how to what and why* 172-181. Springer Berlin Heidelberg.

Nazeer, K. A., and M. P Sebastian, (2009). Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, *Proceedings of the World Congress on Engineering*, vol. 1, July 1-3, 2009, London, U.K.

Nock, R. and F. Nielsen, (2006). On Weighting Clustering, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8. 1–13.

Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods", *J of the American Statistical association*, vol. *66*, no.336, 846-850.

Subhashini, R., and Kumar, V.J.S (2010). Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval, First International Conference on *Integrated Intelligent Computing (ICIIC), pp. 27*,31., 5-7 August, Banglore, India.

Zhao, Y. and Karypis, G (2002). Evaluation of hierarchical clustering algorithms for document datasets. *Proceedings of the eleventh international conference on Information and knowledge management*, 515-524, 4-9 November, ACM New York, NY, USA.