World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 3, No. 5, 97-99, 2013

An Efficient Incremental Clustering Algorithm

Nidhi Gupta
USICT, GGSIPU
Delhi, India

R.L.Ujjwal
USICT, GGSIPU
Delhi, India

Abstract- Clustering is process of grouping data objects into distinct clusters so that data in the same cluster are similar. The most popular clustering algorithm used is the K-means algorithm, which is a partitioning algorithm. Unsupervised techniques like clustering may be used for fault prediction in software modules.

This paper describes the standard k-means algorithm and analyzes the shortcomings of standard k-means algorithm. This paper proposes an incremental clustering algorithm. Experimental results show that the proposed algorithm produces clusters in less computation time.

Keywords - Clustering; Incremental Clustering; K-means; Unsupervised; Partitioning; Data Objects.

I. INTRODUCTION

Clustering is the task of organizing data into groups (known as clusters) such that the data objects that are similar to(or close to) each other are put in the same cluster. Clustering is a form of unsupervised learning in which no class labels are provided.

K-means clustering is a popular clustering algorithm based on the partition of data. However, there are some disadvantages of it, such as the number of clusters needs to be defined beforehand. The proposed algorithm overcomes the shortcoming of k-means algorithm.

The rest of the paper is organized as follows. Section 2 describes the standard k-means algorithm. Section 3 describes the related work. Section 4 present the method proposed in this paper. Experimental results are shown in section 5.Finally conclusions are drawn in section 6.

II. K-MEANS CLUSTERING ALGORITHM

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters

K-means Algorithm

1. The basic step of k-means clustering is to give n data objects and k number of clusters.
2. Input the centroids of each cluster.
3. Determine the distance of each object to the centroids
4. Group the object based on minimum distance.
5. Update the centroids
6. Repeat steps from 3 to 5 until change in Groups.

A. Limitations of K-means Algorithm

1. The number of clusters (K) needs to be determined beforehand.
2. The algorithm is sensitive to an initial seed selection
3. It is sensitive to outliers.
4. Number of iterations are unknown.

III. RELATED WORK

Shi Na, Liu Xumin, Guan Yong[1] proposed an improved K-means Clustering algorithm. The improved method avoids computing the distance of each data object to the cluster centers repeatedly, saving the running time.

K A Abdul Nazeer, D Madhu Kumar ,M P Sebastian[2] proposed a heuristic method based on sorting and partitioning the input data, for finding the initial centroids in accordance with the data distribution, hereby improving the accuracy of the K-means algorithm.

Baolin Yi, Haiquan Qiao, Fan Yang, Chenwei Xu[4] proposed a new method to find the initial center and improve the sensitivity to the initial centers of k-means algorithm. The algorithm first computes the density of the area where the data object belongs to; then it finds k data objects, which are belong to high density area, as the initial start centers. Experimental results shows that the proposed method can produce a high purity clustering results and eliminate the sensitivity to the initial centers to some extent.

Juntao Wang, Xiaolong Su[3] proposed an improved K-means algorithm using noise data filter. The algorithm developed density-based detection methods based on characteristics of noise data.

Fang Yuan, Zeng-HuiMeng, Hong-Xia Zhang, Chun-Ru Dong [11] proposed a systematic method for finding the initial centroids. The centroids obtained by this method are consistent with the distribution of data.

Fahim A M et al [10]. Proposed an efficient method for assigning data-points to clusters. The original K-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids.Fahim's approach makes use of two distance functions for this purpose-one similar to the k-means and other one based on a heuristics.

Abdul Nazeer and Sebastian[9] proposed an algorithm comprising of separate methods for accomplishing the two phases of clustering.

Mushfeq-Us-Saleheen Shameem and Raihana Ferdous[6] proposed a modified algorithm that uses Jaccard distance measure to choose k most different document and use it as k centroid of cluster. Author show in his result that the sum of square in modified k- mean is nearly half of the traditional k-mean.

IV. Proposed Algorithm

In this paper, an incremental clustering approach is used. The basic idea of this algorithm is as follows: Let Tth denotes a threshold of dissimilarity between data objects.

We initially give a value of Tth then choose an object randomly from the given datasets, let it be the center of a cluster, and choose another object from the given datasets again, compute distance between the selected data object and the existing cluster center, If this distance is larger than Tth then form a new cluster and selected object will be the center of the cluster otherwise group the object into existing cluster and update its centroid. Choose an object again from the datasets, repeat the process until all objects are clustered.

Clustering Steps

1. The basic step of proposed clustering is to give n data objects.
2. Assign any random data object to the first cluster.
3. Select next random object.
4. Determine the distance between selected object and centroids of existing clusters.
5. Compare the distance with threshold limit, group the object into existing cluster or form a new cluster with that object.
6. Repeat the steps 3 to 5 until all objects are selected.

Input: D= { d_1,d_2,d_3,……..d_n} //Set of *n* objects to cluster
Output: K={k_1,k_2,k_3….k_k },C= {c_1,c_2,c_3,… c_k } //K is set of subsets of D as final clusters and C is set of centroids of those clusters
Algorithm:

Proposed Algorithm (D)

1. let k=1
2. D_l=RAND()
3. k_k={ d_k}
4. K= { k_k }
5. C_k= d_i
6. Assign some constant value to T_{th}
7. for i= 2 to *n* do
8. Determine distance (m) between d_i and each centroid Cj of any kj in K such that m is minimum. (1<=j<=k)
9. if (m<= T_{th}) then //T_{th}–threshold limit for max. distance allowed
10. k_j= k_j U d_i
11. Calculate new mean (centroid c_j) for cluster k_j;
12. else k= k+1
13. k_k= d_i
14. K= K U k_k
15. C_k= d_i

V. Experimental Results

A synthetic data set is taken which contains 600 data points and each data point contains 4 attributes. The same data set is given as input to the standard K-means algorithm and the proposed algorithm. First we run the proposed algorithm and note down the clusters formed for taking different value of the threshold. For the same number of clusters we check the k-means algorithm by specifying the value of K equals to the number of cluster formed using proposed algorithm.

Experiments compare k-means algorithm with the proposed algorithm in terms of total execution time of clusters.

The results of the experiments are tabulated in Table 1.

TABLE 1: COMPARISON OF THE K-MEANS AND PROPOSED ALGORITHM USING A SYNTHETIC DATA SET.

	K-means Algorithm	**Proposed Algorithm**	
Number of Clusters	**Time Taken(s)**	**Threshold value**	**Time Taken(s)**
9	0.769231	15	0.219780
8	1.043956	17	0.274725
7	0.769231	18	0.549451
6	0.879121	19	0.467033
5	0.659341	20	0.549451
4	0.549451	22	0.219780
3	0.384615	25	0.164835
2	0.274725	35	0.219780

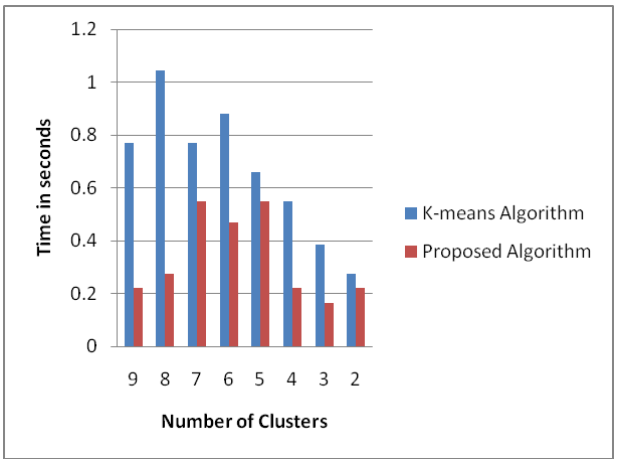

Figure1. Comparison of time taken by the algorithms.

VI. CONCLUSION

In this paper, we propose a new clustering algorithm that can remove the disadvantages of K-means algorithm. In Proposed algorithm we do not need to specify the value of K i.e. the number of cluster required. An experimental result shows that the proposed algorithm takes less time than K-means algorithm. From our result we conclude that the proposed algorithm is better than the K-means algorithm.

VII. REFERENCES

[1] Shi Na, Liu Xumin,, Guan yong ,"Research on k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics.

[2] K A Abdul Nazeer, S D Madhu Kumar, M P Sebastian," Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroid", 2011 Second International Conference on Emerging Applications of Information Technology, IEEE,978-0-7695-4329-1

[3] Juntao Wang, Xiaolong Su," An improved K-Means clustering algorithm", 2011 IEEE, 978-1-61284-486-2

[4]Baolin Yi, Haiquan Qiao, Fan Yang, Chenwei Xu, "An Improved Initialization Center Algorithm for K-means Clustering," IEEE 2010 .

[5] Abdul Nazeer K A, Sebastian M P, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," Proceedings of the International Conference on Data Mining and Knowledge Engineering, London, UK, 2009.

[6]Mushfeq-Us-Saleheen Shameen, Raihana Ferdous,"An Efficient K-Means Algorithm integrated with Jaccard Distance ",2009 IEEE,978-1-4244-4570-7.

[7] Jirong Gu ,Jieming Zhou, Xianwei Chen," An Enhancement of K-means Clustering Algorithm",2009 International Conference on Business Intelligence and Financial Engineering, 978-0-7695-3705-4.

[8]Xiaoping Qing, Shijue Zheng,"A new method for initializing the K-means Clustering algorithm", 2009 Second International Symposium on Knowledge Acquisition and Modeling, IEEE, 978-0-7695-3888-4.

[9] K A Abdul Nazeer and M P Sebastian, "A O(n logn) clustering algorithm using heuristic partitioning," Technical Report, Department of Computer Science and Engineering, NIT Calicut, March 2008.

[10] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):1626–1633, 2006.

[11] Fang yuan,Zeng-Hui Meng, H. X Zhang and C. R Dong, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29,August 2004.