International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 7– No. 2, April 2014 – www.ijais.org

An Advanced Clustering Algorithm (ACA) for Clustering Large Data Set to Achieve High Dimensionality

Amanpreet Kaur Toor
Research Scholar (M.Tech, C.S.E)
Amritsar College of Engineering & Technology
Manawala, Amritsar, Punjab, India

Amarpreet Singh
Associate Professor
Amritsar College of Engineering & Technology
Manawala, Amritsar, Punjab, India

ABSTRACT

The cluster analysis method is one of the critical methods in data mining; this method of clustering algorithm will manipulate the clustering results directly. This paper proposes an Advanced Clustering Algorithm in order to addresses the concern of high dimensionality and large data set [1]. The Advanced Clustering Algorithm method avoids computing the distance of each data object to the cluster recursively and save the execution time. ACA requires a simple data structure to store information in each iteration, which is to be used in the next iteration. Experimental results show that the Advanced Clustering Algorithm method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the traditional algorithm Kohonen SOM. This paper includes Advanced Clustering Algorithm (ACA) and its simulated experimental results with different data sets.

Keywords

ACA, SOM, Clustering, Large Data Set, High Dimensionality, Cluster Analysis

1. INTRODUCTION

Clustering is the process of organizing data objects into a set of dissimilar classes called Clusters. Clustering is an unsupervised technique of Classification. In unsupervised technique the correct answers are not known in advance. Classification is a technique that assigns data objects to a set of classes. Formally, we have a set of dimensional points and a distance function that gives the distance between two points and we are required to compute cluster centers, such that the points falling in the same cluster are similar and points that are in different cluster are dissimilar. Most of the initial clustering techniques were developed by various communities, where the goal of the communities was to cluster a small number of data instances. However, within the data mining community, the focus has been on clustering large datasets [2]. Developing clustering algorithms to effectively cluster swiftly growing datasets has been discovered as an important challenge.

A number of clustering algorithms have been proposed to solve clustering problems like K-Means, Kohonen SOM and HCA. Experimental results have shown that Kohonen SOM is a superlative clustering algorithm among K-means, HAC [3]. But Kohonen SOM also has some shortcomings that are discussed below.

Kohonen Self Organizing Feature Map or SOM provides a way of representing multidimensional data in much lower dimensional space - usually one or two dimensions. This process of reducing the dimensionality of vectors is essentially a data compression technique known as PCA (Principal Component Analysis). SOM is non-deterministic and can produce different results in different run.

In addition, the Kohonen technique creates a network that stores information in such a way that any topological relationships within the training set are maintained. One of the most interesting aspects of SOM is that they learn to classify data without any external supervision whatsoever. It consists of neurons or map units, each having a location in a continuous multidimensional measurement space as well as in a discrete two dimensional data collection is repeatedly presented to the SOM until a topology preserving mapping from the multi dimensional measurement space into the two dimensional output space is obtained. This dimensionality reduction property of the SOM makes it especially suitable for data visualization. Every SOM is different therefore we must be careful what conclusions we draw from our results. [4]

Various methods have been proposed till date to solve clustering problems but it has been analyzed that the SOM algorithm fails to give optimum result when it comes to clustering high dimensional data set because their complexity tends to make things more difficult when the number of dimensions are added. The Quality of SOM algorithm reduces when used on High Dimensional Data. In data mining this problem is known as "**Curse of Dimensionality"**. This research will deal the problem of high dimensionality and large data set.

Several clustering algorithms had been proposed till date but each of them is used for some specific requirement. There does not have a single algorithm that can effectively handle all sorts of requirement. This makes an enormous challenge for the user to select one among the available algorithm for specific purposes. To deal with this problem, a new algorithm has been proposed in this research that is named as ***"Advanced Clustering Algorithm".***

This paper is organized s follows. Section 2 presents a modified approach. Section 3 describes about the time complexity of the proposed method. Section 4 experimentally demonstrates the performance of the proposed method. And the final Section 5 describes the conclusion.

2. MODIFIED APPROACH

2.1 Advanced Clustering Algorithm (ACA)

To overcome the shortcomings of the SOM algorithm, this paper presents an Advanced Clustering Algorithm method. The main idea of the algorithm is to split the data structures into different subset to keep the labels of the cluster. The distance of all the data objects to the nearest cluster centre is calculated during each iteration that can be used in the next iteration. If the computed distance is smaller than or equal to the distance to the old center, then the data object remain in its own cluster that was assigned to it in the prior iteration.

Therefore, there is no need to calculate the distance from the data object to the other k-1 clustering centers. By this methodology ACA saves the computational time to the k-1 cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers and find the nearest cluster center. It assigns this point to the adjacent cluster center and then individually record the distance to its cluster center. Because in each iteration some data points still remain in the original cluster. It means that some parts of the data points will not be calculated and saving total time of calculating the distance. So ACA is enhancing the efficiency of the clustering.
The ACA algorithm takes the dataset and the values of k as the only input needed since the initial centroid are computed automatically. It finds the optimal centroid by algorithm.

Algorithm 1: The Advanced Method

The process of the Advanced Clustering algorithm is described as follows:

Input: The desired number of clusters k.

Dataset S.

D= $\{d_1, d_2, \ldots d_n\}$ contain n data objects.

di=$\{ x_1, x_2, \ldots x_m \}$// Set of attributes.

Output: A set of k clusters.

1. Draw multiple sub-samples {SI, S2, . . . ,Sj } from the original dataset.

2. For each data point calculate the distance from origin.

3. Sort the data points according to the distance.

4. Repeat step1 to 2 until the number of data points reaches Sj.

5. From each subset take the middle point as the initial centroid by calculating the mean.

6. Calculate the distance between each data point di to all the initial centroids cj by using SSE method.

7. Find the closest centroid cj for each data point di by assigning di to cluster j.

8. For each cluster j, recalculate the centroids.

9. Repeat

10. For each data point di,

a. Calculate the distance from the centroid to the nearest cluster.

b. If SSE is less than or equal to the calculated distance, then the data point will present in the same cluster.

c. Else calculate the distance for every centroid cj.

End for;

11. Merge two nearest clusters into one cluster that have relatively low SSE.

12. Recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

3. TIME COMPLEXITY

This paper proposes an Advanced Clustering Algorithm to obtain the initial cluster. Time complexity of the ACA is O (nk). Here some data points remain in the original clusters, while the others move to other clusters. If the data point retains in the original cluster then the required complexity is O (1), else O (k). With the convergence of the clustering algorithm, the number of data points moved from their cluster will reduce.
If half of the data points move from their cluster, the time complexity is O (nk/2). Hence the total time complexity is O (nk). While the time complexity of the SOM clustering algorithm is not known because it produces different results in different run. So the proposed algorithm in this paper can effectively improve the Quality of clustering and reduce the computational complexity.

4. EXPERIMENTAL RESULTS

This paper selects academic data set repository of machine learning databases to test the efficiency of the Advanced Clustering Algorithm (ACA) and the standard algorithms such as SOM. Many simulated experiments have been carried out to demonstrate the performance of the Advanced Clustering Algorithm in this paper. This algorithm has also been applied to the clustering of real datasets on the WEKA data mining tool. The data set is given as input to the SOM algorithm and the Advanced Clustering Algorithm. Experiments compare Advanced Clustering Algorithm with the SOM algorithm. The comparisons of SOM and ACA have been made in terms of the total execution time of clusters and their quality. The Experimental operating system is Windows 8, programming language is Java. A brief description of the datasets used in experimental evaluation is described in Table 1 and it shows some characteristics of the data sets.

Table 1: Data Set Size

S. No	Dataset	Number of attributes	Number of records
1	**Academic Activities**	15	5504
2	**1998 Household**	57	52

	Trend Data		
3	**Employee Data**	10	74
4	**IRIS Data**	5	50

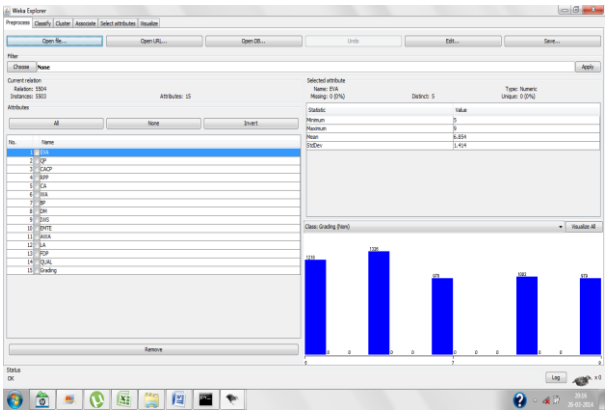

Fig 1: Display data set according to class attributes.

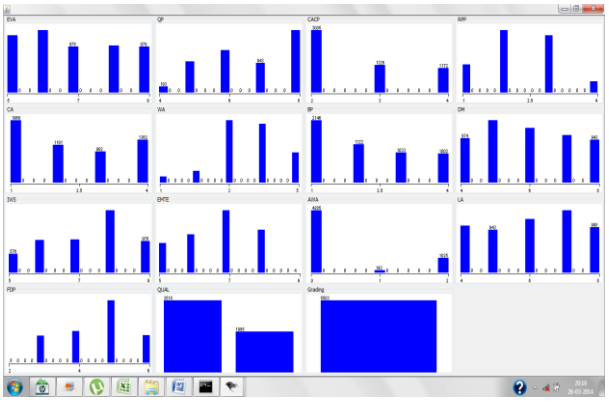

Fig 2: Display All Attributes

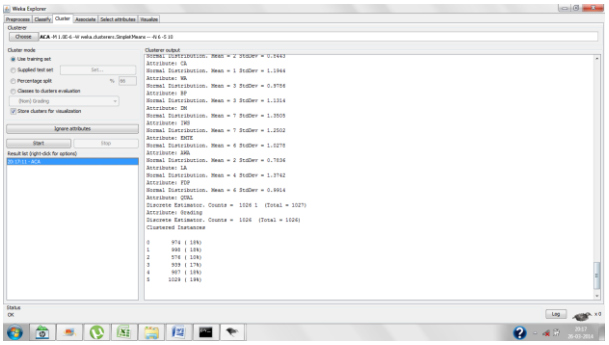

Fig 3: Display clusters and the % of data that each cluster contains.

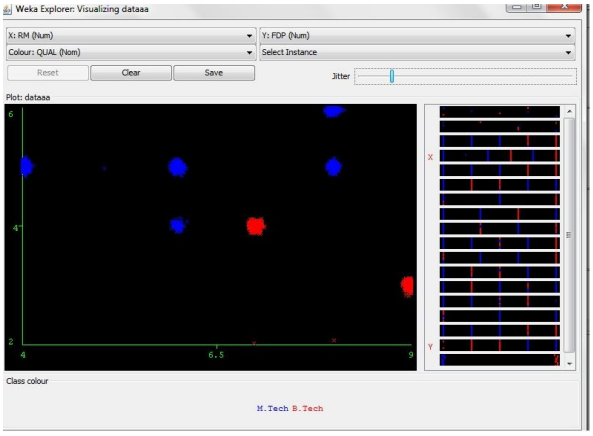

Fig 4: Visualization of scatter plot (Clusters)

4.1 ALGORITHM COMPARISION

The SOM and ACA clustering algorithm are compared according to the following factors:-

a. Number of Clusters

b. Quality (In terms of Error Rate)

c. Size of the data set.

d. Type of the data set.

For each parameter four tests are conducted. Table 2. shows the comparison among SOM and ACA based upon the Error Rate. We have calculated the Error Rate of both algorithms by modifying the value of k. Initially we have defined the value of k=2 and then it vary to 4, 6 and 8. As we have observed from Table 2 that Error Rate of SOM algorithm increases as the value of k becomes greater. At the initial conditions Error Rate of ACA algorithm is high as compared to SOM algorithm, but Error rate decreases as the value of k becomes greater.

Table 2: The association between Number of Clusters and Error rate of SOM and ACA algorithm.

Number of Clusters	**Quality (Error Rate)**	
	SOM	**ACA**
2	0.2941	0.5563
4	0.6681	0.2444
6	0.7468	0.1666
8	0.9968	0.1297

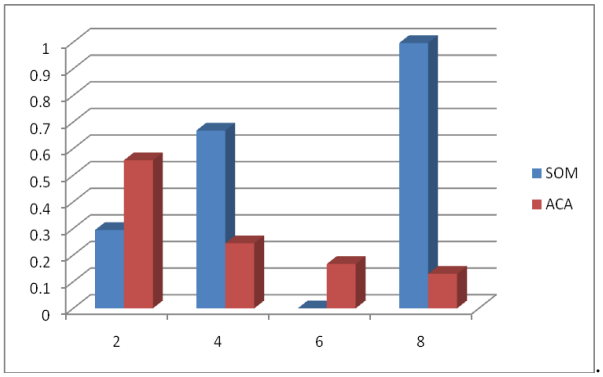

Fig 5: Graphical analysis of Quality based on Error Rate

Table 3 shows an association between Number of Clusters and Execution Time of SOM and ACA algorithm. Experimental results have proven that the Execution time of the ACA algorithm as compared with SOM algorithm is quite high, but it remains constant even if the value of K becomes greater. But this is not true in case of the SOM algorithm. In SOM algorithm as the value of k becomes greater the Execution Time also increases and at a limit of k and it crosses the ACA results. As a result of this Execution Time of ACA algorithm becomes better.

Table 3: The association between Number of Clusters and Execution Time of SOM and ACA algorithm.

Number of Clusters	Execution Time	
	SOM	ACA
2	46 ms	1000 ms
4	63 ms	1000 ms
6	78 ms	1000 ms
8	125 ms	1000 ms

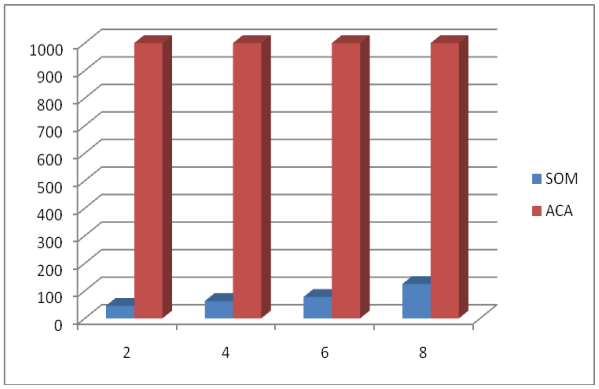

Fig 6: Graphical analysis of Execution Time of SOM and ACA algorithm.

According to the dataset size (Table 4) a large data set of academic activities contains 15 columns and 5504 instances and small data set contains 7 columns and 1000 instances. The small data set has been extracted from the large data set. Experimental results show that the Quality of ACA algorithm becomes excellent when using on a huge data set. The SOM algorithm produces low quality results when used on large data set as compared to ACA algorithm. As a conclusion, ACA algorithm is well suited for large data set.

Table 4: The effect of the data set size and Number of Clusters on Quality of SOM and ACA algorithm

K= 8	Quality	
Data Set Size	SOM	ACA
82560	0.9968	0.1296
7000	0.9151	0.2328

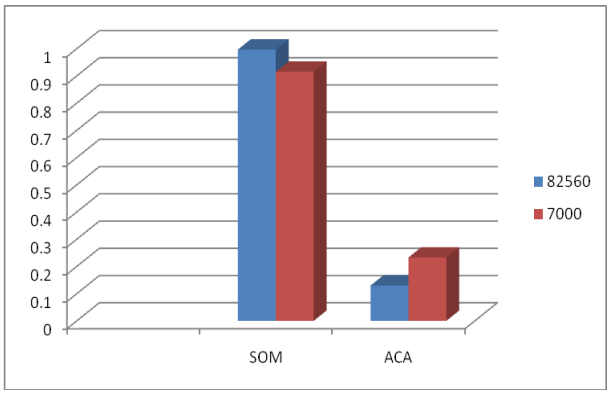

Fig 7: Graphical analysis of Data Set Size among SOM and ACA algorithm.

According to the type of the data set (Table 5), we have used 3 different data type of data set. These data sets are extracted from the Internet. The 1998 Household Trend Data contains all numeric instances. Employee Data and IRIS Data are both multivariate. Experimental results show that Quality of ACA is extremely high even we have implemented it on different type of data set. So ACA algorithm is quite feasible.

Table 5: The effect of the data set type on Quality of SOM and ACA algorithm

K= 6	Quality	
Data Set Type	SOM	ACA
1998 Household Trend Data	0.1483	0.1219
Employee Data	0.5784	0.1666
IRIS Data	0.7742	0.1666

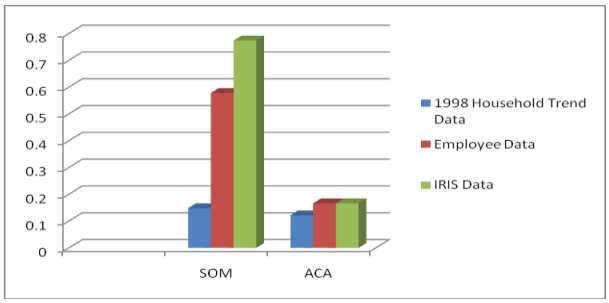

Fig 8: Graphical analysis of Data Set Type among SOM and ACA algorithm

5. CONCLUSION

SOM algorithm is a typical clustering algorithm and it is widely used for clustering large sets of data. This paper elaborates Advanced Clustering Algorithm and analyses the shortcomings of the SOM algorithm. Because the computational complexity of the SOM algorithm is offensively high owing to the need to reassign the data points a number of times during each iteration, which effects its efficiency. This paper presents a simple and efficient way of assigning data points to clusters. The proposed method ACA in this paper ensures the entire process of clustering in O (nk) time without sacrificing the accuracy of clusters. Experimental results show the Advanced Clustering Algorithm can improve the execution time, quality of SOM algorithm and works well on High Dimensional data set. So the proposed method is feasible.

6. FUTURE WORK

In this paper we have compared the one predefined algorithm and one new clustering algorithm. We have given some conclusions above. But still we are not able to cover up all the factors for comparing these algorithms. As a future work, comparisons can be made by using much more parameters and the execution time of ACA for less number of K can be reduced.

7. ACKNOWLEDGMENTS

I would like to thank the Department of Computer Science & Engineering of Amritsar College of Engineering & Technology, Manawala, GT Road Amritsar, Punjab, India. I would also like to thank my parents for their moral and intellectual support for this work.

8. REFERENCES

[1] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.

[2] Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[3] Amanpreet Kaur Toor, Amarpreet Singh, " Analysis of Clustering Algorithm based on Number of Clusters, error rate, Computation Time and Map Topology on large Data Set", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 2, Issue 6, November- December 2013.

[4] Amanpreet Kaur Toor, Amarpreet Singh, " A Survey paper on recent clustering approaches in data mining", International Journal of Advanced Research in Computer Science and Software Engineering Vol 3, Issue 11, November 2013.

[5] Sun Shibao, Qin Keyun," Research on Modified K-means Data Cluster Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200– 201,July 2007.

[6] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[7] Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.

[8] Zhao YC, Song J. GDILC: A grid-based density isoline clustering algorithm. In: Zhong YX, Cui S, Yang Y, eds. Proc. of theInternet Conf. on Info-Net. Beijing: IEEE Press,2001.140−145.http://ieeexplore.ieee.org/iel5/7719/21161/00982709.pdf

[9] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998.

[10] K.A.AbdulNazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm",Proceeding of the World Congress on Engineering, vol 1,london, July 2009.

[11] Fred ALN, Leitão JMN. Partitionalvs hierarchical clustering using a minimum grammar complexity approach. In: Proc. of the SSPR & SPR 2000. LNCS 1876, 2000.193−202.http://www.sigmod.org/dblp/db/conf/sspr/sspr2000.htm

[12] Gelbard R, Spiegler I. Hempel's raven paradox: A positive approach to cluster analysis. Computers and Operations Research, 2000,27(4):305−320.

[13] Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997.146−151.

[14] Ding C, He X. K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. In: Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press, 2004. 584−589.http://www.acm.org/conferences/sac/sac2004/

[15] HinneburgA,KeimD.An efficient approach to clustering in large multimedia databases with noise.In:AgrawalR,StolorzPE,Piatetsky- Shapiro G,eds.Proc.of the 4th Int'l Conf.on Knowledge Discovery and Data Mining(KDD'98).New York:AAAIPress,1998.58~65.

[16]ZhangT,RamakrishnanR,LivnyM.BIRCH:An efficient data clustering method for very large databases.In:JagadishHV,MumickIS,eds.Proc.of the 1996 ACM SIGMOD Int'l Conf.on Management of Data.Montreal:ACM Press,1996.103~114. [15] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial- temporal data. Data & Knowledge Engineering, 2007,60(1): 208-221.