Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

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

Aman Toor

Received: 6 December 2013 Accepted: 4 January 2014 Published: 15 January 2014

Abstract

Cluster analysis method is one of the main analytical methods in data mining; this method of clustering algorithm will influence the clustering results directly. This paper proposes an Advanced Clustering Algorithm in order to solve this question, requiring a simple data structure to store some information [1] in every iteration, which is to be used in the next iteration. The Advanced Clustering Algorithm method avoids computing the distance of each data object to the cluster centers repeat, saving the running time. 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. This paper includes Advanced Clustering Algorithm (ACA) and describes the experimental results and conclusions through experimenting with academic data sets.

Index terms— ACA, SOM, K-MEANS, HAC, clustering, large data set, high dimensionality, cluster analysis

1 Introduction

lustering is the process of organizing data objects into a set of disjoint classes called Clusters. Clustering is an Unsupervised Clustering technique of Classification. Classification refers to a technique that assigns data objects to a set of classes. Unsupervised means that clustering does not depends upon predefined classes while clustering the data objects. Formally, given a set of dimensional points and a function that gives the distance between two points , 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 statistics or pattern recognition communities, where the goal was to cluster a modest number of data instances. However, within the data mining community, the focus has been on clustering large datasets [2]. Developing clustering algorithms to effectively and efficiently cluster rapidly growing datasets has been identified as an important challenge.

A number of clustering algorithms have been proposed to solve clustering problems. One of the most popular clustering methods are K-Means, SOM, HCA. Their shortcomings are discussed below.

The standard k-means algorithm needs to calculate the distance from the each data object to all

Author : e-mail: er.amantoor@gmail.com the centers of k clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases. In K-Means algorithm initial cluster centers are produced arbitrary, it does not promise to produce the peculiar clustering results. Efficiency of original k-means algorithm is heavily rely on the initial centroid. Initial centroid also has an influence on the number of iterations required while running the original K-Means algorithm. Computational Complexity of K-Means algorithm is very high and does not provide high quality clusters when it comes to cluster High dimensional data set.

Kohonon's SOMs are a type of unsupervised learning. The goal is to discover some underlying structure of the data. SOM algorithm is computationally expensive. Large quantity of good quality representative training data required. No generally accepted measure of 'quality' of a SOM e.g. Average quantization error (how well the data is classified). Every SOM is different therefore we must be careful what conclusions we draw from our results. SOM is non-deterministic and can and will produce different results in different run.

Hierarchical clustering algorithms are either topdown or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. [6] Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached. This algorithm is sensitive to outliers and sometimes it is difficult to identify the correct number of clusters from Dendrogram. [7] Various methods have been proposed in literature but it have been analyzed that the K-Means, SOM, HCA fails to give optimum result when it comes to clustering high dimensional data set because their complexity tends to make things more difficult when number of dimensions are added. In data mining this problem is known as "Curse of Dimensionality". This research will deal the problem of high dimensionality and large data set.

A large number of algorithms had been proposed till date, each of them address some specific requirement. There does not exist a single algorithm which can adequately handle all sorts of requirement. This makes a great challenge for the user to do selection among the available algorithm for specific task. To cope with this problem, a new algorithm is going to be proposed in this research that is named as "Advanced Clustering Algorithm". This paper is organized s follows. Section 2 presents an overview of ACA. Section 3 introduces proposed method. Section 4 describes about the time complexity of proposed method. Section 5 experimentally demonstrates the performance of proposed method. And the final Section 6 describes the conclusion.

2 II.

3 Advanced Clustering Algorithm

Experimental results have shown Kohonon's SOM is superlative clustering algorithm among Kmeans, HCA [8]. For the shortcomings of the above SOM algorithm, this paper presents an Advanced Clustering Algorithm method. The main idea of algorithm is to set two simple data structures to retain the labels of cluster and the distance of all the data objects to the nearest cluster during the each iteration that can be used in next iteration. We calculate the distance between the current data object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in it's cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k-1 clustering centers, saving the calculative 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 and assign this point to the nearest cluster center. And then we separately record the label of nearest cluster center and the distance to it's 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, saving a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

4 III.

5 Proposed Algorithm

The process of the Advanced Clustering algorithm is described as follows: Input: The number of desired clusters k, and a database D= {d 1 , d 2 , d n } containing n data objects. Output: A set of k clusters.

1. Draw multiple sub-samples {SI, S2, . . . ,Sj } from the original dataset. 2. Repeat step 3 for m=l to i 3. Apply combined approach for sub sample. IV.

6 Time Complexity

This paper proposes an Advanced Clustering Algorithm, to obtain the initial cluster, time complexity of the advanced algorithm 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, this needs O (1), else O (k). With the convergence of 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 standard k-means clustering algorithm require O(nkt). So the proposed algorithm in this paper can effectively improve the speed of clustering and reduce the computational complexity. But the Advanced kmeans algorithm requires the pre estimated the number of clusters, k, which is the same to the standard kmeans algorithm. If you want to get to the optimal solution, you must test the different value of k.

7 V.

8 Experimental Results

This paper selects academic data set repository of machine learning databases to test the efficiency of the advanced algorithm and the standard algorithms. Two simulated experiments have been carried out to demonstrate the performance of the Advanced in this paper. This algorithm has also been applied to the clustering of real datasets. In two experiments, time taken for each experiment is computed. The same data set is given as input

to the standard algorithm and the Advanced Clustering Algorithm. Experiments compare Advanced Clustering Algorithm with the standard algorithm in terms of the total execution time of clusters and their accuracy. Experimental operating system is Window 8, program language is java. This paper uses academic activities as the test datasets and gives a brief description of the datasets used in experiment evaluation. Table 1 shows some characteristics of the datasets.

9 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 standard kmeans, SOM and HAC clustering algorithm. Because the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times during every iteration, which makes the efficiency of standard clustering is not high. This paper presents a simple and efficient way for assigning data points to clusters. The proposed method [1] [2]

Figure 1: C

[1] © 2014 Global Journals Inc. (US)Cluster analysis method is one of the main analytical methods in data mining; this method of clustering algorithm will influence the clustering results directly. This paper proposes an Advanced Clustering Algorithm in order to solve this question, requiring a simple data structure to store some information[1] in every iteration, which is to be used in the next iteration. The Advanced Clustering Algorithm method avoids computing the distance of each data object to the cluster centers repeat, saving the running time. 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. This paper includes Advanced Clustering Algorithm (ACA) and describes the experimental results and conclusions through experimenting with academic data sets.Abstract-

[2] © 2014 Global Journals Inc. (US)An Advanced Clustering Algorithm (ACA) for Clustering Large Data Set to Achieve High Dimensionality

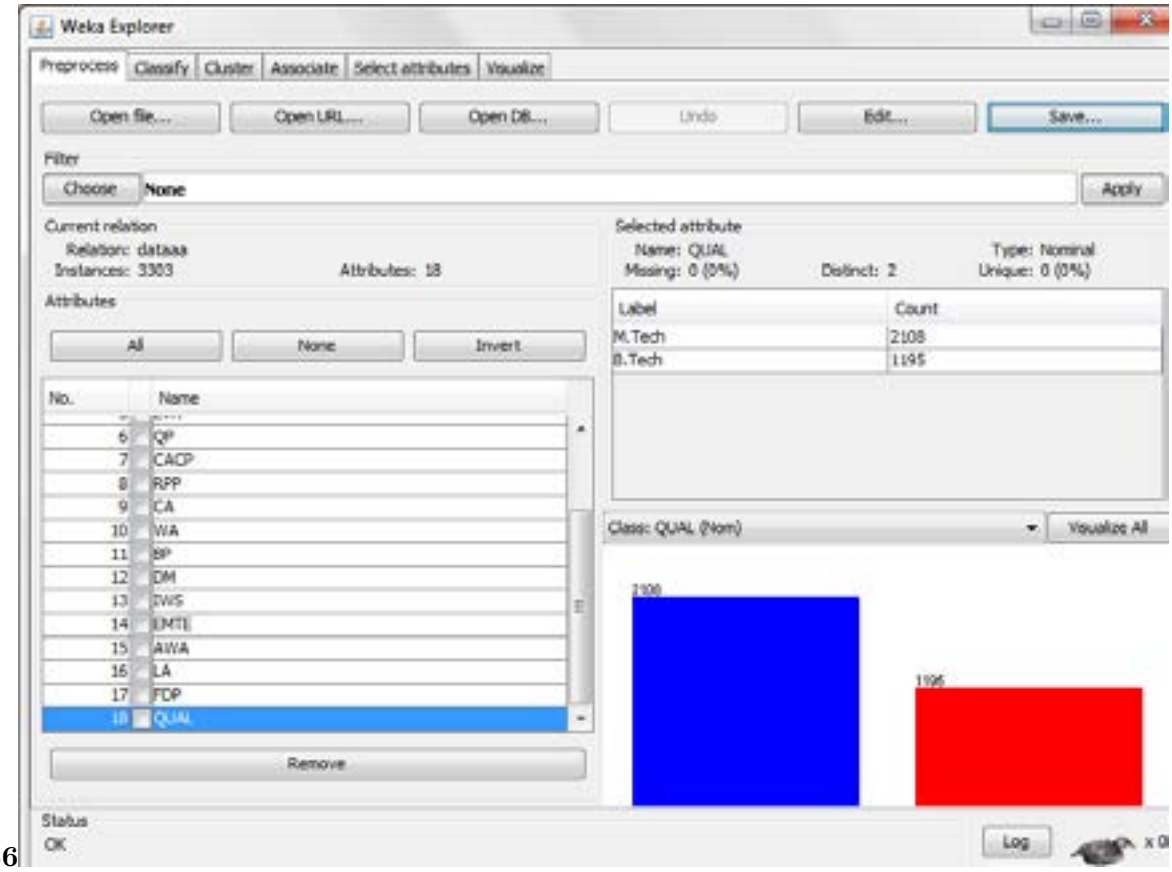


Figure 2: 4. Compute centroid 5 . 6 .

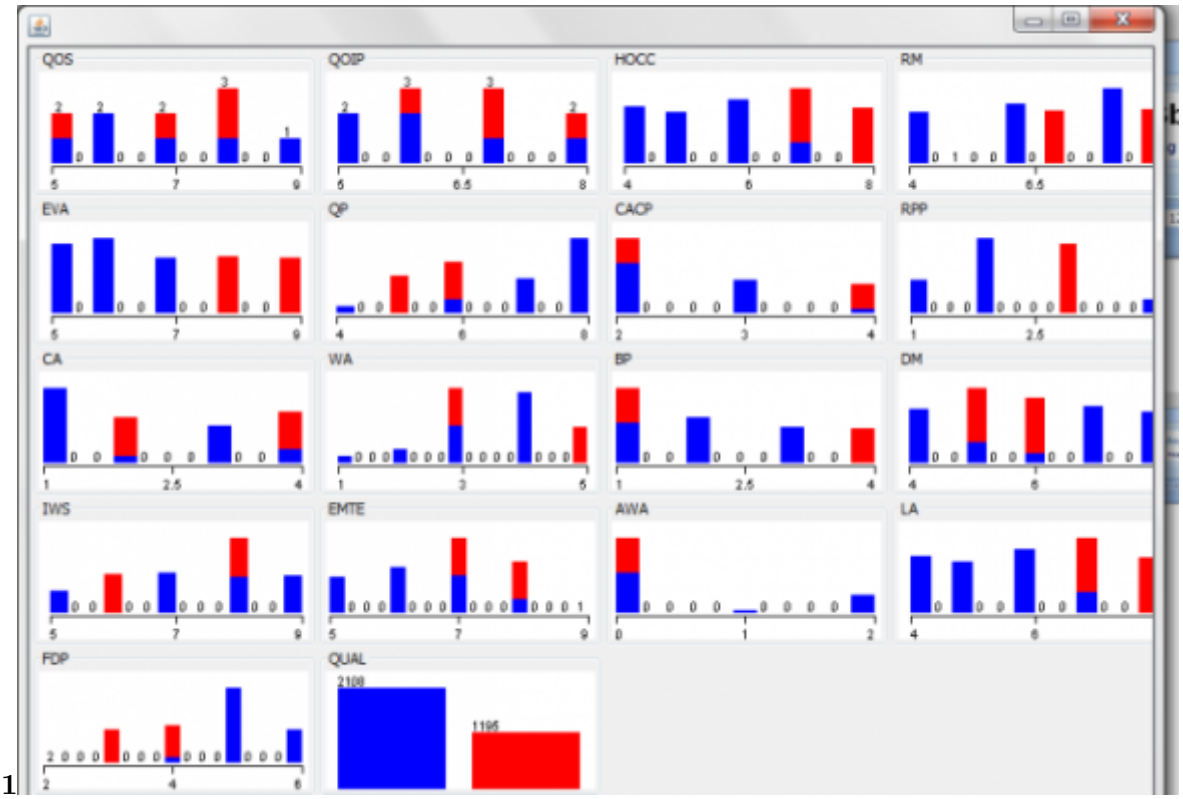
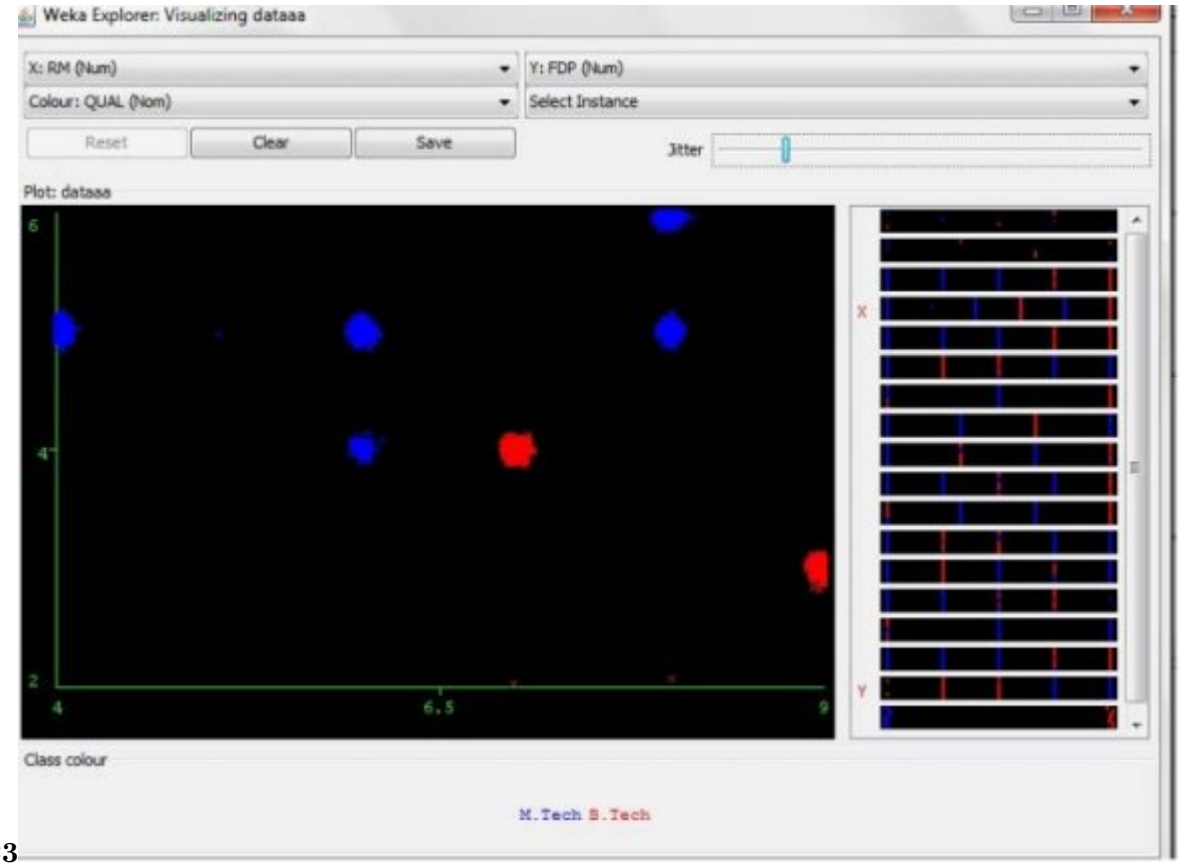


Figure 3: Figure 1 :



23

Figure 4: Figure 2 :Figure 3 :

1

Dataset	Number of attributes	Number of records
Academic Activities		

Figure 5: Table 1 :

2

	Clustering Algorithm			
Parameter	SOM	K-Means	HAC	ECA
Error Rate 0.8189 0.8456 0.8379 0.3672				
Execution Time	297 ms	1281 ms	1341 ms	1000 ms
Accessing Time	Fast	Slow	Slow	Very fast
Number of Clusters	6	6	6	4

Figure 6: Table 2 :

.1 Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org

[Huang ()] 'A fast clustering algorithm to cluster very large categorical data sets in data mining'. Z Huang . *Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, (of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge DiscoveryTucson) 1997. p. .

[Zhao and Gdilc ()] 'A grid-based density isoline clustering algorithm'. Y C Zhao , Song J Gdilc . http://iee-explore.ieee.org/iel5/7719/21161/00982709.pdf *Proc. of theInternet Conf. on Info-Net*, Y X Zhong, S Cui, Y Yang (ed.) (of theInternet Conf. on Info-NetBeijing) 2001. IEEE Press. 140.

[Yuan et al. (2004)] 'A New Algorithm to Get the Initial Centroids'. F Yuan , Z H Meng , H Zhang , DongC . *Proc. of the 3rd International Conference on Machine Learning and Cybernetics*, (of the 3rd International Conference on Machine Learning and Cybernetics) August 2004. p. .

[Amanpreet Kaur Toor and Singh (2013)] *A Survey paper on recent clustering approaches in data mining*, Amarpreet Amanpreet Kaur Toor , Singh . November 2013. 3.

[Birant et al. ()] 'An algorithm for clustering spatial-temporal data'. D Birant , A Kut , St-Dbscan . *ACM SIGMOD Int'l Conf. on Management of Data*, (Montreal) 1996. 2007. ACM Press. 103 p. .

[Zhang et al. ()] 'An efficient data clustering method for very large databases'. T Zhang , R Ram Akrishnan , . Birch . *Proc. of the*, H V Jagadish, I S Mumick (ed.) (of the) 1996.

[Fahim et al. (2006)] 'An efficient enhanced k-means clustering algorithm'. A M Fahim , A M Salem , F Torkey . *Journal of Zhejiang University Science A* July 2006. 10 p. .

[Amanpreet Kaur Toor and Singh (2013)] *Analysis of Clustering Algorithm based on Number of Clusters, error rate, Computation Time and Map Topology on large Data Set*, Amarpreet Amanpreet Kaur Toor , Singh . November-December 2013. 2.

[Jigui et al. (2008)] 'Clustering algorithms Research'. Sun Jigui , Liu Jie , Zhao Lianyu . *Journal of Software* January 2008. 19 (1) p. .

[Huang ()] 'Extensions to the k-means algorithm for clustering large data sets with categorical values'. Z Huang . *Data Mining and Knowledge Discovery* 1998. 2 p. .

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

[Nazeer and Sebastian (2009)] 'Improving the Accuracy and Efficiency of the k-means Clustering Algorithm'. K A Nazeer , M P Sebastian . *Proceeding of the World Congress on Engineering*, (eeding of the World Congress on Engineering) July 2009. 1.

[Ding et al. ()] 'Nearest-Neighbor in data clustering: Incorporating local information into global optimization'. C Ding , X He , K- . http://www.acm.org/conferences/sac/sac *Proc. of the ACM Symp. on Applied Computing*, (of the ACM Symp. on Applied ComputingNicosia) 2004. 2004. ACM Press. 584.

[Fred and Leitão ()] 'Partitionalvs hierarchical clustering using a minimum grammar compexity approach'. Aln Fred , Jmn Leitão . Ohttp://www.sigmod.or-g/dblp/db/conf/sspr/sspr2000.htm *Proc. of the SSPR & SPR 2000*, (of the SSPR & SPR 2000) 2000.193?202. 1876.

[Shibao et al. (2007)] 'Research on Modified kmeans Data Cluster Algorithm'. Sun Shibao , ; I S Qin Keyun , C P Jacobs , Bean . *Computer Engineering* July 2007. 33 (13) p. . (Fine particles, thin films and exchange anisotropy)

[Merz and Murphy] *UCI Repository of Machine Learning Databases*, C Merz , P Murphy . ftp://ftp.ics.uc-i.edu/pub/machine-learningdatabases