Journal of
Computer Science & Systems Biology

Toor and Singh, J Comput Sci Syst Biol 2014, 7:4
DOI: 10.4172/jcsb.1000146

Research Article Open Access

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

Amanpreet Kaur Toor and Amarpreet Singh*

Amritsar College of Engineering & Technology, Punjab, India

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 the question of high dimensionality and large data set. The Advanced Clustering Algorithm method avoids computing the distance of each data object to the cluster centers again and again and save the running time. ACA requires a simple data structure to store information in every 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 algorithms (K-Means, SOM and HAC). This paper includes Advanced Clustering Algorithm (ACA) and its experimental results through experimenting with academic data sets.

Keywords: ACA; SOM; K-Means; HAC; Clustering; Large data set; High dimensionality; Cluster analysis

Introduction

Clustering is the process of organizing data objects into a set of disjoint classes called Clusters. Clustering is an unsupervised technique of Classification. In unsupervised technique the correct answers are not known (or just not told to the network). Classification refers to a technique that assigns data objects to a set of classes. 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 [1,2]. 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. Some of the most popular clustering methods are K-Means, SOM, and HCA. Their shortcomings are discussed below.

The standard K-Means algorithm needs to calculate the distance from the each data object to all 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 relying 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 [3].

Kohonen Self Organizing Feature Map or SOM provide a way of representing multidimensional data in much lower dimensional spaces - usually one or two dimensions. This process, of reducing the dimensionality of vectors, is essentially a data compression technique known as vector quantization. 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 SOMs 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 multi- dimensional 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 [4-7].

Hierarchical clustering algorithms are either top-down 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 [8]. 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 [9].

Various methods have been proposed in literature but it has 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 [10,11].

A large number of algorithms had been proposed till date, each

***Corresponding author:** Amarpreet Singh, Associate Professor, Amritsar College of Engineering & Technology Manawala, Amritsar, Punjab, India, Tel: 0183-506-9532; E-mail: er.amantoor@gmail.com

Received March 29, 2014; **Accepted** April 16, 2014; **Published** April 19, 2014

Citation: Toor AK, Singh A (2014) An Advanced Clustering Algorithm (ACA) for Clustering Large Data Set to Achieve High Dimensionality. J Comput Sci Syst Biol 7: 115-118. doi:10.4172/jcsb.1000146

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 select one among the available algorithm for specific task. To cope up with this problem, a new algorithm has been proposed in this research that is named as **"Advanced Clustering Algorithm".**

Proposed Advanced Clustering Algorithm (ACA)

Experimental results have shown Kohonon SOM is superlative clustering algorithm among K-means, HAC [12]. SOM provide a way of representing multidimensional data in much lower dimensional spaces - usually one or two dimensions and SOM is non-deterministic and can produce different results in different run.

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 its 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 accessing time to the k-1 cluster centers [13]. 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 nearest cluster center and then separately record the distance to its 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.

Algorithm 1: The Advanced Method

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

Input: The number of desired clusters K.

Dataset S.

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

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

Output: A set of K clusters.

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

2. Repeat step 3 for m=1 to n.

3. Apply combined approach for sub sample.

4. In each set, take the middle point as the initial centroid.

5. Compute the distance between each data point to all the initial centroids

6. For each data point find the closest centroid and assign to nearest cluster.

7. Choose minimum of minimum distance from cluster center criteria.

8. Now apply new calculation again on dataset S for K clusters.

9. Combine two nearest clusters into one cluster.

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

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 then the required complexity is 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) [14]. Hence the total time complexity is O(nk). While the time complexity of 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 speed of clustering and reduce the computational complexity.

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 (K-Means, SOM and HAC). Two simulated experiments have been carried out to demonstrate the performance of the Advanced in this paper [15,16]. This algorithm has also been applied to the clustering of real datasets on WEKA data mining tool. 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 [17]. 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 (Figures 1-6, Tables 1 and 2).

Conclusion

SOM algorithm is a typical clustering algorithm and it is widely used for clustering large sets of data. This paper elaborates

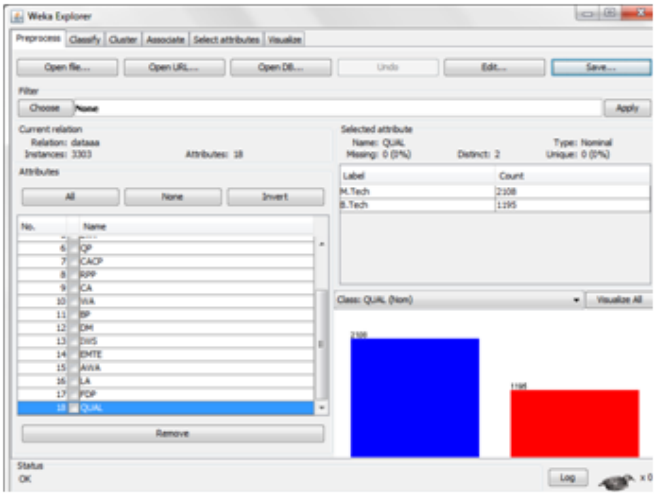

Figure 1: Display data set according to class attributes.

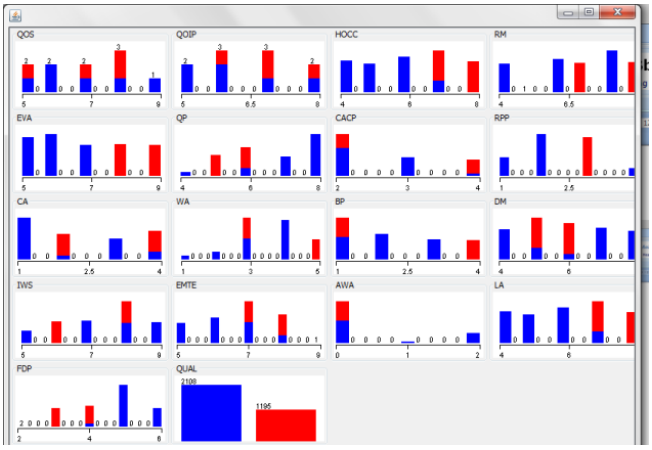

Figure 2: Display All Attributes.

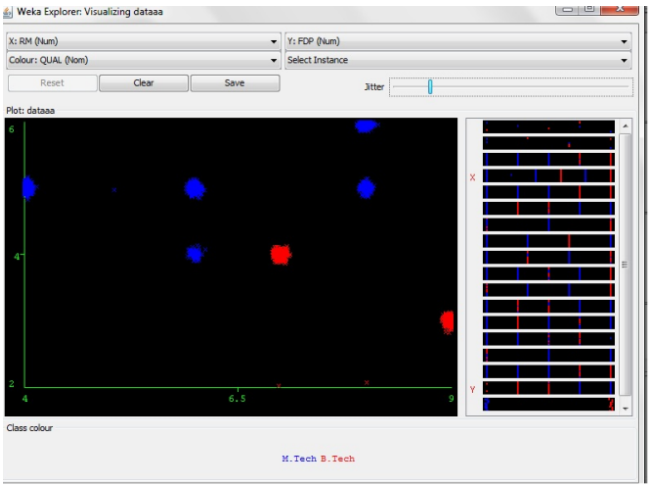

Figure 3: Visualization of scatter plot.

Figure 4: Performance Comparison Based on Error Rate (Quality).

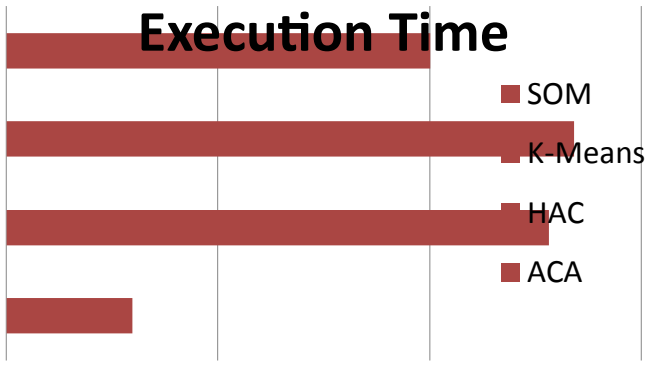

Figure 5: Performance Comparison Based on Execution Time.

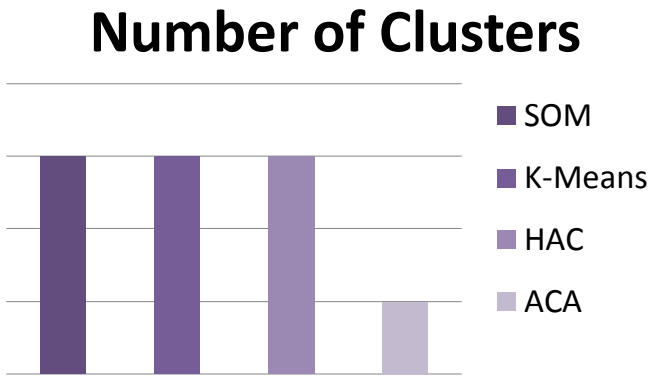

Figure 6: Performance Comparison Based on Number of Clusters.

Dataset	Number of attributes	Number of records
Academic Activities	15	5504

Table 1: Data Set Size.

Parameter	SOM	K-Means	HAC	ACA
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

Table 2: Analysis between traditional (K-Means, SOM, HAC) and Advanced Clustering Algorithm.

Advanced Clustering Algorithm and analyses the shortcomings of the standard k-means, 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 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.

References

1. Yuan F, Meng ZH, Zhang HX, Dong CR (2004) A New Algorithm to Get the

Initial Centroids. Proceedings of 2004 International Conference on Machine Learning and Cybernetics 2: 1191-1193.

2. Sun J, Liu J, Zhao L (2008) Clustering algorithms research. Journal of Software 19: 48-61.
3. Sun S, Qin K (2007) Research on Modified k-means Data Cluster Algorithm. Fine particles, thin films and exchange anisotropy. Computer Engineering, Jacobs IS, Bean CP editors, 33: 200-201.
4. ftp://ftp.ics.uci.edu/pub/machine-learning-databases
5. Fahim AM, Salem AM, Torkey FA (2006) An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University Science A 10: 1626-1633.
6. Zhao YC, Song J (2001) GDILC: A grid-based density isoline clustering algorithm. 2001 International Conferences on Info-tech and Info-net, 2001. Proceedings. ICII 2001-Beijing. 3: 140-145.
7. Toor AK, Amarpreet S (2013) A Survey paper on recent clustering approaches in data mining. International Journal of Advanced Research in Computer Science and Software Engineering 3.
8. Toor AK, Amarpreet S (2013) 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 2: 94-98.
9. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2: 283-304.
10. AbdulNazeer KA, Sebastian MP (2009) Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. Proceeding of the World Congress on Engineering 1: 1-5.
11. Fred ALN, Leitão JMN (2000) Partitional vs Hierarchical clustering using a minimum grammar complexity approach. Lecture Notes in Computer Science 1876: 193-202.
12. Gelbard R, Spiegler I (2000) Hempel's Raven paradox: a positive approach to cluster analysis. Computers & Operations Research 27(4): 305-320.
13. Huang Z (1997) A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 146-151.
14. Ding C, He X (2004) K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. Proceedings of the 2004 ACM symposium on applied computing, 584-589.
15. Hinneburg A, Keim D (1998) An efficient approach to clustering in large multimedia databases with noise. American Association for Artificial Intelligence 58-65.
16. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An efficient data clustering method for very large databases. SIGMOD International Conference on Management of Data. Montreal, ACM Press, Canada, 103-114.
17. Birant D, Kut A (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering 60: 208-221.