IJRSET APRIL 2017 Volume 4, Issue 4 Pages: 21-26

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

EFFICIENT CLUSTERING ALGORITHM USING BIRCH CLUSTERS

[1] G. Abdulkalamazad, [2] Dr.J.Jeba Emilyn
[1] PG Scholar, [2] Associate Professor,
[1, 2] Department of Information Technology,
[1, 2] Sona College Of Technology, India.

ABSTRACT: The search for useful patterns in large data sets has recently attracted considerable interest, and one of the most common problems in this area is the identification of clusters or densely populated regions in a multidimensional data set. Premature work does not adequately address the problem of large data sets and minimize 1/0 costs. Clustering is a widely used technique in data mining. At present, there are many clustering algorithms, but most existing clustering algorithms are either limited to handle the single attribute or can handle both types of data, but are not efficient when clustering large data sets. Only a few algorithms can do both well. Clustering is the process of grouping of data, where the grouping is established by finding similarities between data based on their characteristics. Such groups are termed as Clusters. A comparative study of clustering algorithms across two different data items is performed here. The performance of the various clustering algorithms is compared based on the time taken to form the estimated clusters. The experimental results of various clustering algorithms to form clusters are depicted as a graph. Thus it can be concluded as the time taken to form the clusters increases as the number of cluster increases. The BIRCH clustering algorithm takes very few seconds to cluster the data items whereas the simple KMeans takes the longest time to perform clustering. The experimental results suggest that the BIRCH algorithm is effective when compared to k-means algorithm.The results show that the BIRCH algorithm is efficient and produces better quality of clusters.

Keywords: [Clustering, Clustering algorithms, K-means, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).]

1. INTRODUCTION

Data mining is a technique to analyze and retrieve from large amount of database and transform it to useful information for future use [1].data mining is used in classification, clustering, regression, association rule discovery, sequential pattern discovery, outlier detection, etc.[2].Mining can be done in two learning approaches-Supervised and Unsupervised learning in data mining applications. Clustering is the task of groupings a set of objects in such a way that objects in the cluster are more similar to each other than to those in other clusters [4].Clustering techniques have numerous applications in various fields including artificial intelligence, pattern recognition, bio

informatics, segmentation and machine learning.

Clustering is mainly needed to organize the results provided by a search engine. Clustering can also be viewed as a special type of classification. The clusters formed as a result of clustering can be defined as a set of like elements. But the elements from different clusters are not alike. Clustering is similar to database segmentation, where like tuples in a database are grouped together. When clustering is applied to a real world database, many problems occur there such as: handling outlier is difficult; interpreting the semantics of each cluster is difficult, no correct answer for a clustering problem and what data should be used for clustering. The problem of clustering can also be defined as below: Given a collection of data objects, the work of clustering is to divide the data objects into groups such that objects in the same group are similar. Objects in different groups should be dissimilar. Data belonging to one cluster are the most similar; and data belonging to different clusters are the most dissimilar. Clustering algorithms can be viewed as hierarchical and partitional. With hierarchical clustering, a nested set of clusters is created. The hierarchy is divided into various levels. In the lowest level, each item will have its own cluster. In the highest level, all the items will be belonging to a single cluster. With partitional clustering, only one set of cluster is created. Hierarchical clustering is represented using a tree structure called dendrogram. Examples of hierarchical clustering algorithms are agglomerative and divisive clustering algorithms. Examples of partitional clustering algorithms are KMeans, nearest neighbour and PAM. Clustering can be done on large databases also. Most popular clustering algorithms like BIRCH (balanced iterative reducing and clustering using hierarchies) [12], DBSCAN (density based spatial clustering of applications with noise). Clustering can also be performed with categorical attributes. Optimization based partitioning algorithms are represented by its prototype.

Objects of similar prototype are clustered together. An iterative control strategy is used to optimize the clustering. If the clusters are of convex shape, same size, density, and if their number *k* can be reasonably estimated, then the clustering algorithm can be selected correctly. *K*-means, *k*-modes and *k*-medoid algorithms can be differentiated based on their prototypes. The *k*-means method has been shown to be effective in producing good clustering results for many practical applications. However, the *k*-means algorithm requires time proportional to the product of number of patterns and number of clusters per iteration. This computationally may be expensive especially for large datasets. Among clustering formulations that are based on minimizing a formal objective function, perhaps the most widely used and studied is *k*-means clustering. Given a set of *n* data points in real *d*-dimensional space, Rd, and an integer *k*, the problem is to determine a set of *k* points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest centre. A comparative study between various clustering algorithms based on the time taken to form the clusters is considered. The various clustering algorithms taken into consideration are simple KMeans and BIRCH clustering. The rest of the paper is organized in to five sections. Section II explains the methodology used in this paper. Section III explains various clustering algorithms. Section IV gives description about dataset. Section V presents the experimental results in graphical forms. Finally Section VI concludes the paper.

2. METHODOLOGY

The methodology describes all the steps according to which clustering algorithms is performed.

Step1: Load the data sets

The “Two moons” dataset has been chosen and downloaded from the UCI Machine Learning Repository.

Step2: Normalize data

After loading of the dataset the next step is to normalize the data. Select normalize filter and apply on the same data set. Save the result using save button.

Step3: Similarity Graph

To perform similarity graph, choose the graph type and number of neighbors from the Normal K-Nearest Neighbours [4]. It creates the similarity graph in the connected components.

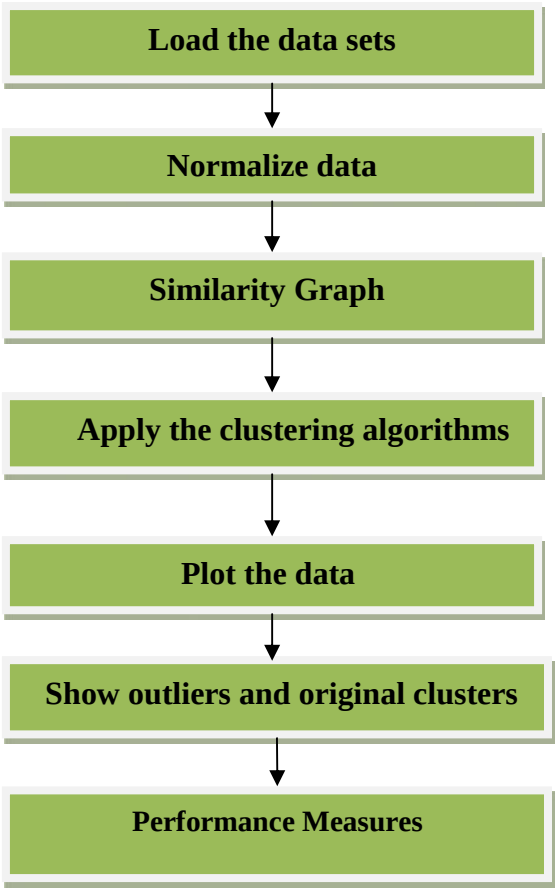

Figure - 1 Methodology of Birch algorithm

Step 4: Apply the clustering algorithms

To perform the clustering algorithms, choose the Hierarchical clustering algorithm namely BIRCH clustering algorithm. It chooses the number of clusters to cluster the original data.

Step5: plot the data

After clustering the data, the data should be plotted in various ways like silhouette data, clustering data and similarity graph.

Step6: Show outliers and original clusters

The BRICH algorithm can define the original cluster points and outliers by means of the correlative threshold by users. These points with higher connection strength than the threshold are defined as core points, vice versa. This method with threshold often has a subordinate efficiency because users must have a deep understanding of the biological sequence data set and also have to make extensive correlation tests for threshold values. There is obviously the possibility of a threshold error, and it is often a biological sequence dataset that reduces clustering precision and maneuverability over the long term.

Step 7: Performance measures -Performance measures show the total output as in graphical format. It is easy to show our data by comparing the existing and proposed analysis using graphical data.

3. CLUSTERING ALGORITHMS

A. K-MEANS CLUSTERING ALGORITHM

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The results of partitioning method are a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.[13,14] **Example:** A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset. K-means is a data mining algorithm which performs clustering of the data samples. As mentioned previously, clustering means the division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to

cluster the database, K-means algorithm uses an iterative approach.

B. BIRCH Algorithm

In the BIRCH [5] tree a node is called a Clustering Feature. It is a small representation of an underlying cluster of one or many points. BIRCH builds on the idea that, points that are close enough should always be considered as a group.

Clustering Features provide this level of abstraction. Clustering Features are stored as a vector of three values: CF = (N; LS; SS). The linear sum (LS), the square sum (SS), and the number of points it encloses (N). A CF tree is a height balanced tree that has two parameters namely, a branching factor, B, and threshold, T. The representation of a non-leaf node can be stated as {CF i, child i}, where, i = 1, 2, …., B, Child i : A pointer to its I th child node. Cf i: CF of the sub cluster represented by the ith child.

The non-leaf node provides a representation for a cluster and the contents of the node represents all of the sub clusters. In the same manner a leaf-node's contents represents all of its sub clusters and has to confirm to a threshold value for T.

Phase 2 (optional): Condense into desirable range by building a smaller CF tree

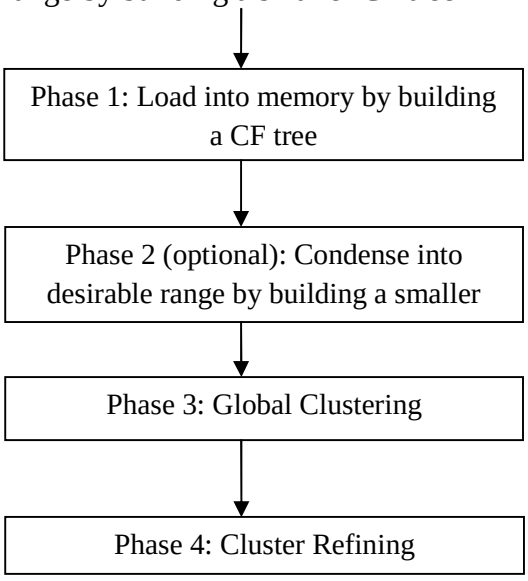

Figure - 2 BIRCH algorithm

The BIRCH clustering algorithm [5] is implemented in four phases. In phase-1, the initial CF is built from the database based on the branching factor B and the threshold value T. Phase-2 is an optional phase in which the initial CF tree would be reduced in size to obtain a smaller CF tree. Global clustering of the data points is performed in phase3 from either the initial CF tree or the smaller tree of phase2.

Good clusters can be obtained from phase-3 of the algorithm. If it is required to improve the quality of the clusters, phase-4 of the algorithm would be needed in the clustering process. The execution of Phase1 of BIRCH begins with a threshold value T.

ADVANTAGES

It is local in that each clustering decision is made without scanning all data points and currently existing clusters.

It exploits the observation that data space is not usually uniformly occupied and not every data point is equally important.

It makes full use of available memory to derive the finest possible sub-clusters while minimizing I/O costs.

It is also an incremental method that does not require the whole dataset in advance.

STEPS

Phase 1: Scan all data and build an initial in-memory CF tree, using the given amount of memory and recycling space on disk.

Phase 2: Condense into desirable length by building a smaller CF tree.

Phase 3: Global clustering.

Phase 4: Cluster refining – this is optional, and requires more passes over the data to refine the results.

4. DATASET

For performing clustering algorithms Abalone dataset has been used. The datasets has downloaded from the UCI Machine Learning Repository. Any type of dataset can be dynamically included. The categorical,

integer, real attributes [11] are used in the datasets.

A.1. Abalone dataset: Predict the age of abalone from physical measurements. The total number of instances is 4177.The number of attributes is 8.some of the attributes used here are sex, length, diameter, height, whole weight etc.

A.2.Parkinsons Dataset: This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient; the name of the patient is identified in the first column.

5. EXPERIMENTAL RESULT

A. Experimental setup

This section presents the experimental results of various clustering algorithms are measured based on the time to form the clusters. Here, different datasets are used namely, Abalone dataset, Atom dataset, Parkinsons data set, Rainbow dash dataset, Two Moons dataset.

Dataset	No. of Instances	No. of Attributes
Abalone	4177	9
Atom	800	3
Parkinsons	195	23
Rainbow Dash	6400	3
Two Moons	14977	2

TABLE - 1 DATASETS USED

B. Results for datasets on different clustering algorithms

The different datasets like Abalone dataset, Atom dataset, Parkinsons data set, Rainbow dash dataset, Two Moons dataset are processed on various clustering algorithms such as K-Means, Birch clustering.

Datasets	Number of clusters	K-Means (Time in secs)	Birch (Time in secs)
Abalone	2	0.297149	0.27
Atom	2	0.404290	0.07
Parkinsons	2	0.031714	0.05
Rainbow Dash	2	5.192541	0.30
Two Moons	2	27.217073	1.57

TABLE - 2 TIME TAKEN TO FORM THE RESPECTIVE NUMBER OF CLUSTERS

From Table II, it is shown that with a number of clusters based on the time taken to form estimated clusters under different datasets.

C. Graph Representation For Performance Evolution

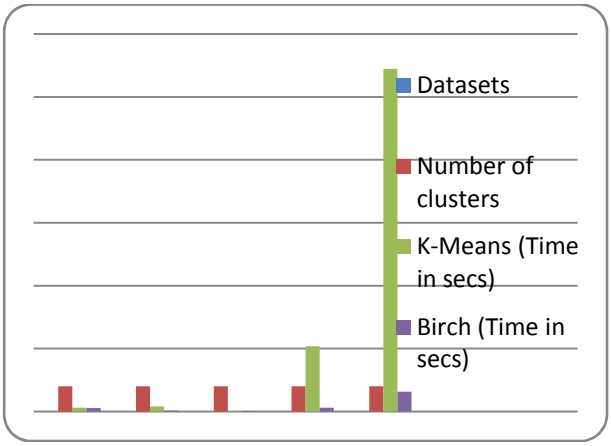

Figure - 3 Execution Time Of different Data sets.

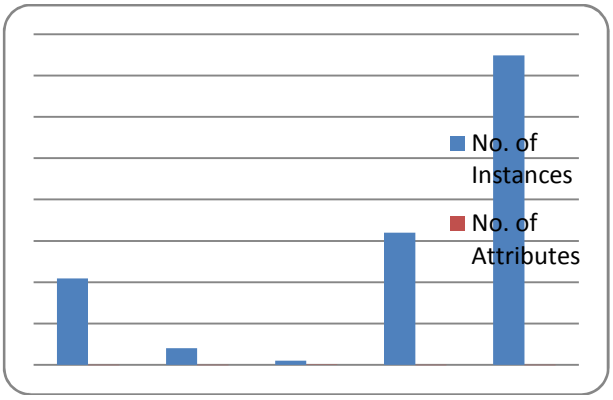

Figure - 4 Number of clusters based on its size.

CONCLUSION

A comparative study of clustering algorithms across different data items is performed here. The performance of the various clustering algorithms is compared based on the time taken to form the estimated clusters. The experimental results of various clustering algorithms to form clusters are depicted as a graph. As the number of clusters increases gradually, the time to form the clusters also increases. The BIRCH clustering algorithm takes very few seconds to cluster the data items whereas the simple KMeans takes the longest time to perform clustering.

REFERENCES

[1].Usama Fayyad, Gregory Piatetsky Shapiro and padhraic Symyh, “The KDD Process for Extracting Useful Knowledge from Volumes of Data”, Communication of the ACM, Vol. 39, No. 11, pp. 27-34,1996.

[2].Chauhan R, Kaur H, Alam M A, “Data Clustering Method for Discovering Clusters in Spatial Cancer Databases”, International Journal of Computer Applications , (0975 – 8887) Vol.10– No.6, November 2010.

[3].Jain A.K., Murty M.N., and Flynn P.J., “Data Clustering: A Review”, ACM Computing Surveys, 31 (3). pp. 264-323, 1999

[4].Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur” Efficent K-means Clustering Algorithm Using Ranking Method In Data Mining”ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012

[5].Tian Zhaung, Raghu Ramakrishna, aud Miron Livny,BIRCH: An Efficient Data Clustering Method for Very Large Databases, Technical Report, Computer Sciences Dept., univ. of Wisconsin-Madison, 1995.

[6].J.MacQueen. some methods for classification and analysis of multivariate observations. Pro.5th Berkeley Symp.Math.Statist, Pro., 1:128-297, 1967.

[7].Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovering, 2:283-304, 1998.

[8].R.Ng and J.Han. Efficient and effective clustering method for spatial data mining. In Pro.1994 Int.Conf .Very Large Data Bases, pages144-155, 1994.

[9].M. Ester, H.-P.Kriegel, J.Sander, and X.Xu. A density-based algorithm for discovering clustering in large spatial database with noise. In Proc. 1996 Int. Conf. Knowledge Discovering and Data Mining.1996 : 266-231.

[10]. T.Chiu, D.P.Fang, J.Chen and Y.W. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In Proc. ACM-SIGKDD int.conf. Knowledge discovery and data mining (KDD’2001)page 263-268,2001.

[11]. Peijun Chen, Yu Wang. An Efficient clustering algorithm for categorical and mixed typed attributes. Computer Engineering and Application. page 190-191, 2004(1)

[12].K. A. Abdul Nazeer, M. P. Sebastian “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm” Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K

[13].Kehar Singh, Dimple Malik and Naveen Sharma “Evolving limitations in K-means algorithm in data mining and” IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011.

[14]. Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur” Efficent K-means Clustering Algorithm Using Ranking Method In Data Mining”ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.