ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

Analysis of clustering technique for the diabetes dataset using the training set parameter

R.Nithya[1], P.Manikandan[2], Dr.D.Ramyachitra[3]

M.Phil Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu[1]

Ph. D Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu[2]

Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu[3]

Abstract: The Clustering technique is used to place data elements into related groups without advance knowledge of the group description. The clustering technique belongs to an unsupervised learning and it is used to discover a new set of categories. The clustering technique groups the data instances in to subsets in such a manner that similar instances are grouped together while different instances belong to different groups. This paper represents the performance of three clustering algorithms such as Hierarchical clustering, Density based clustering and K Means clustering algorithm. The Diabetes dataset is used for the comparison of those clustering algorithms based on the performance of execution time and the number of clustered instances.

Keywords: Hierarchical clustering, Density based clustering, K Means clustering, Diabetes dataset, Training set, Clustering.

1. INTRODUCTION

There are various clustering algorithms such as partition clustering, hierarchical clustering, Density based clustering, and fuzzy clustering, etc. These clustering algorithms are classified according to the creation of clusters of objects [1]. Determining good clusters is one of the main goals in clustering algorithms. The clustering is one of the major problems in variety of domains such as the statistical data analysis, medical image processing, data mining and knowledge discovery, bioinformatics and data classification and compression [2]. Class identification in spatial databases is the main attractive task in clustering algorithms [3]. In supervised learning, the hierarchical clustering is one of the most frequently used methods and it is typically more effective in detecting the true clustering structure of a data set than partitioning algorithms.

In computer vision community, the k-means algorithm is one of the most commonly used clustering algorithms which can be used for its simplicity and effectiveness. It's an iterative algorithm in which, each iteration new cluster centers are computed and each data point is re-assigned to its nearest center. And also the k-means clustering algorithm is widely used in machine learning for clustering and quantization.

Density based clustering algorithms discover the clusters which are dense regions of data points and are separated by sparse regions with respect to given density parameters. And also the density –based clustering can discover the clusters with arbitrary shapes [4].

In this paper the comparison is made to find out which test option is best for the clustering algorithms such as the Hierarchical clustering, Density based clustering and K Means clustering. In the test option there are four kinds of parameter like supplied test set, training set, percentage spilt and class to clusters evaluation. In this paper the training set parameter is used to calculate the data set values. In this paper the Diabetes dataset is used for comparison of those algorithms. The remaining section of this paper is follows. Section 2 describes the literature review, Section 3 describes the methodology for the Diabetes dataset and Section 4 describes the experimental result. Finally Section 5 gives the Conclusion and Future work.

2. LITERATURE REVIEW

Christophe Ambroise, et al., presented a new method for segmenting multispectral satellite images. Their method is unsupervised and consists of two steps. In the first step the pixels of learning set are summarized. And in second step the map are clustered using Agglomerative hierarchical clustering. Each pixel takes the label of its nearest neighbor [5].

Yihong Dong et al., presented an algorithm that applies fuzzy set theory to hierarchical clustering method so as to discover clusters with arbitrary shape. It first partitions the data sets into several sub-clusters using a partitioning method and then constructs a fuzzy graph of sub-clusters by analyzing the fuzzy-connectedness degree among sub-clusters. Their algorithm can be performed in high-dimensional data sets. And also their method generates better quality clusters than traditional algorithms and scales up well for large databases [6].

Clark F. Olson reviewed the important results for sequential algorithms and also describes the parallel algorithms for hierarchical clustering. An Optimal PRAM algorithm using $n/\log n$ processor is given for the average

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

link, complete link, median, centroid, and minimum variance [7].

Younghoon Kim et al., proposed the new density-based clustering algorithm, called DBCURE. It is used to find clusters with varying densities and suitable for parallelizing the algorithm with Map Reduce. Density-based algorithms find each cluster one by one. DBCURE-MR finds several clusters together in parallel. Their experimental results with various data sets confirm that DBCURE-MR finds clusters efficiently without being sensitive to the clusters with varying densities and scales up well with the Map Reduce framework [8].

P. Viswanath et al., proposed, an efficient density based k-medoids clustering algorithm to overcome the drawbacks of DBSCAN and k-medoids clustering algorithm. Their result will be an improved version of kmedoids clustering algorithm. In their algorithm they perform better than DBSCAN while handling clusters of circularly distributed data points and slightly overlapped clusters [9].

Ashish Ghosh et al., proposed Density based clustering algorithm, which is evaluated on a number of well-known benchmark data sets using different cluster validity measures. Their results are compared with those obtained using two popular standard clustering techniques namely average linkage agglomerative and k-means clustering algorithm and with an ant-based method called adaptive time-dependent transporter ants for clustering. Their experimental results justify the potentiality of the APC algorithm both in terms of the solution (clustering) quality as well as execution time compared to other algorithms for a large number of data sets [10].

AristidisLikas et al., proposed the global k-means algorithm which is an incremental approach that dynamically adds one cluster center at a time through a deterministic global search procedure consisting of N (with N being the size of the data set) executions of the k-means algorithm from suitable initial positions. They proposed the modifications of the method to reduce the computational load without significantly solution quality. Their clustering methods are tested on well-known data sets and they compare favorably to the k-means algorithm with random restarts [11].

Adil M. Bagirov., a new version of the global k-means algorithm is proposed. The starting point for the k-th cluster center in their algorithm is computed by minimizing an auxiliary cluster method. The results of numerical experiments on 14 data sets demonstrate the superiority of their algorithm. It requires more computational time than the global k-means algorithm [12].

Omnia Ossama et al., proposed the algorithm, which uses a key feature of moving objects trajectories that are direction as a heuristic to determine the different number of clusters for the k-means algorithm and also used the silhouette coefficient as a measure for the quality of their approach. The experimental results on both real and synthetic data show their performance and accuracy of their proposed technique [13].

3. METHODOLOGY

In this paper the clustering techniques are used to find the best algorithm for the Diabetes dataset based on the training set parameter. Weka tool is used for the comparative analysis for those clustering algorithms. The flow diagram for the comparative analysis is shown in Fig 1.

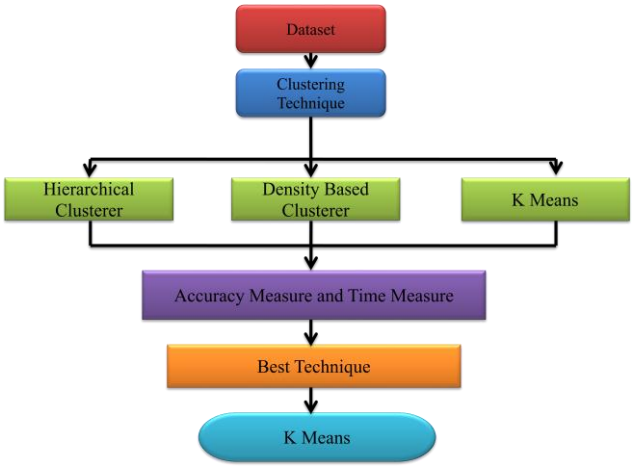

Fig 1: Flow diagram for comparative analysis of clustering technique.

3.1 Dataset

The diabetes dataset is collected from the UCI repository and it contains 769 instances and 9 attributes. Diabetes mellitus is a group of metabolic diseases characterized by high blood sugar (glucose) levels that result from defects in insulin secretion. Patients with high blood sugar will typically experience polyuria, they will become increasingly thirsty and hungry [14]. Glucose is vital to human health, because it's an important source of energy for the cells that make up the muscles and tissues and also the brain's main source of fuel [15].

3.2 Clustering

Clustering technique is the group of a collection of patterns into clusters based on similarity. The patterns with in valid clusters are more parallel to each other than they are to a pattern belonging to a different cluster. The difference between clustering and discriminate analysis is the important concept to be learnt. Normally, the labeled patterns are used to label a new pattern. In clustering, the crisis is to group a given collection of unlabeled patterns into meaningful clusters [16]. In this paper, the clustering algorithms are evaluated to predict which of the algorithm is most suitable for the Diabetes dataset. In the clustering technique three algorithms are compared such as hierarchical clustering, density based clustering, k means clustering to find out which one fits effectively for the diabetes dataset.

3.3 Clustering Algorithms

Clustering technique is used for finding hidden patterns in datamining. In this paper three clustering algorithms are used for finding the best algorithm for the Diabetes dataset and they are as follows.

- ❖ Hierarchical clustering
- ❖ Density Based clustering
- ❖ K Means clustering

DOI 10.17148/IJARCCE.2015.4936

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

3.3.1 Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering is generally of two types:

Agglomerative:

This is a "bottom up" approach. Each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.

Divisive:

This is a "top down" approach. All observations start in one cluster and splits are performed recursively as one moves down the hierarchy [17].

3.3.2 Density Based Clustering

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. The Density-based clustering is defined as areas of higher density of the data set. Density-Based Spatial Clustering of Applications with Noise is most widely used density based algorithm [18].

3.3.3 K Means Clustering

The popular clustering algorithm that minimizes the clustering error is the K-means algorithm. The global k means clustering algorithm which constitutes a deterministic effective global clustering algorithm for the minimization of the clustering error that employs the k means algorithm as a local search procedure. And the k means clustering algorithm progress in an incremental way to solve the clustering problem [19].

4. EXPERIMENTAL MEASURES

In this paper the experimental measures are calculated by using the performance factors such as the clustering accuracy and execution time. And also the comparative analysis for the Diabetes datasets is performed to predict the finest algorithm. The accuracy measure and the execution time for the clustering algorithms are depicted in Table 1.

Algorithms	No. of clu st- ers	Cl ust er (0)	Clu ster (1)	Cl ust er 0 (%)	Clu ster 1 (%)	No. of Iter atio ns	Tim e (sec ond s)	Un cluste red insta nces
Hierarchica l clustering	2	1	768	0	100	0	4.89	0
Density Based Clustering	2	54 1	228	70	30	0	0.3	0
Simple K Means Clustering	2	53 9	230	70	30	4	0.1	0

Table 1: Comparison of performance factors for clustering algorithms

From the experimental results it is inferred that by using the training set parameter for the hierarchical clustering algorithm number of clustered instances in the group and the execution time is higher than the density based clustering and k means clustering algorithms. For the density based clustering algorithm it is inferred that the number of clustered instances is higher than the k means algorithm and lower than the hierarchical clustering, and the execution time is higher than the k means and lower than the hierarchical clustering. For k means algorithm it is inferred that for the training set parameter the number of clusters and execution time is lower than the hierarchical and density based clustering algorithms.

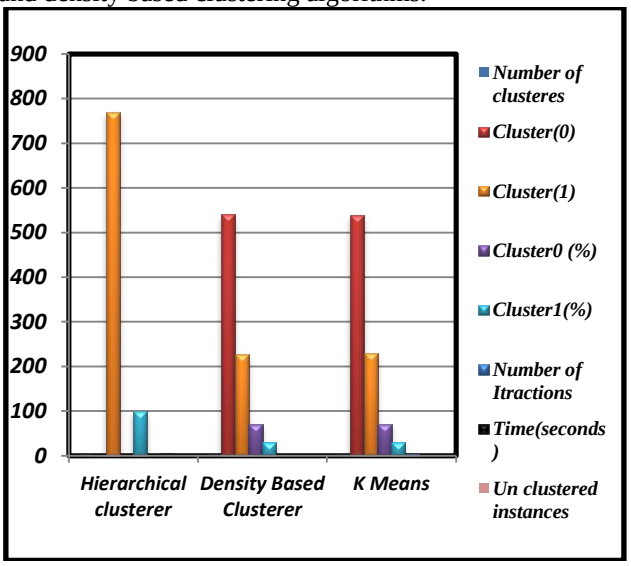

Fig 2: Comparison of performance factors for clustering algorithms

The performance measures for the clustering algorithms such as Hierarchical Clustering, Density Based Clustering and Simple K means algorithm is graphically represented in the Figure 2.The experiment was carried out to the diabetes datasets by using the training set parameter. From the results it is inferred that k means clustering algorithm performs better as compared to the hierarchical clustering and density based clustering. The k means algorithm gives more correctly clustered instances compared to others.

5.CONCLUSION AND FUTURE WORK

In this paper the performance metrics is evaluated for the clustering algorithms such as Hierarchical Clustering, Density Based Clustering and Simple K means clustering algorithms. The algorithms are analyzed by using the trained set parameter based on its class attribute. The performance measures are analyzed based on the number of clustered instances and the execution time taken for clustering the instances. From the experimental results it is inferred that the K means algorithm gives better performance when comparing with the other two algorithms by using the Diabetes dataset. In future the clustering algorithms can be experimented on other datasets also. And in future the k means clustering algorithm will modify to obtain more effective results. And also the k means clustering algorithm can be analyzed using various parameters such as the cross validation, percentage split, and supplied test set.

REFERENCES

1. Peter Scherer et al., "Using SVM and Clustering Algorithms in IDS Systems" 2011, pp. 108–119, ISBN 978-80-248-2391-1.
2. Ming-chuan hung et al., "An Efficient k-Means Clustering Algorithm using simple partitioning" journal of information science and engineering 21, 1157-1177 (2005).
3. Martin Ester et al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

4. Younghoon Kim, et al., "DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce", Information Systems 42 (2014) 15–35.
5. Christophe ambroise, et al., "Christophe ambroise, et al., "Hierarchical clustering of self-organizing maps for cloud classification, Neurocomputing, Volume 30, Issues 1–4, January 2000, Pages 47–52.
6. Yihong Dong et al., "A hierarchical clustering algorithm based on fuzzy graph connectedness", Fuzzy Sets and Systems, Volume 157, 1 July 2006, Pages 1760–1774.
7. Clark F. Olson., "Parallel algorithms for hierarchical clustering" Parallel Computing, Volume 21, Issue 8, August 1995, Pages 1313–1325
8. Younghoon Kim et al., "DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce", Information Systems, Volume 42, June 2014, Pages 15–35.
9. P. Viswanath et al., "Rough-DBSCAN: A fast hybrid based density based clustering method for large datasets", Pattern Recognition Letters 30 (2009) 1477–1488.
10. Ashish Ghosh et al., "Aggregation pheromone density based data clustering" Information Sciences, volume 178, Issue 13, July 2008, Pages 2816-2831.
11. AristidisLikas et al., "The global k-means clustering algorithm", Pattern Recognition 36 (2003) 451–461.
12. Adil M. Bagirow., "Modied global k-means algorithm for minimum sum-of-squares clustering problems", Pattern Recognition 41 (2008) 3192—3199.
13. Omnia Ossama et al., "An extended k-means technique for clustering moving object", Egyptian Informatics Journal (2011)12 45-52.
14. http://www.medicinenet.com/diabetes_mellitus/page2.htm#what_is_diabetes
15. http://www.mayoclinic.org/diseases-conditions/diabetes/basics/definition/con-20033091
16. Osama Abu Abbas., "Comparisons between Data Clustering Algorithms" The International Arab Journal of Information Technology, vol.5, No.3, July 2008.
17. http://members.tripod.com/asim_saeed/paper.html
18. http://en.wikipedia.org/wiki/Hierarchical_clustering
19. Aristidis Likas et al., "The global k means clustering algorithm", Pattern Recognition, Volume 36, Issue 2, February 2003, Pages 451-461.