IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 69-76

Comprehensive Evaluation of Clustering Algorithms In Binary and Multi-Class Attributes

Ms. S. Saranya*, Dr.S.Sasikala**, Ms.P.Deepika***, Mr. Jaikishen J Kamath****, Mr.S.Muthukrishnan*****

*Assistant.Prof., Department of Computer Application, Hindusthan College of Arts and Science,
**Associate Prof., Department of Computer Application, Hindusthan college of Arts and science,
***Assistant.Prof., PG &Research Department of computer science, Hindusthan college of Arts and science
****II BCA, Hindusthan college of Arts and science
*****III BCA, Hindusthan college of Arts and science

Abstract: *Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups or some similar to the other data points in the similar group than those in other groups. In this paper four clustering algorithm are described Expectation–maximization (EM) algorithm, Hierarchical algorithm, Density based algorithm, simple k-means algorithm. These algorithms were implemented in two different dataset. The data set were divide the entire dataset under different labels with similar data points into one cluster. The dataming algorithms are compared in the clustering data using weka tool*

Keywords: *Data mining, clustering algorithm, simple K-means algorithms, Hierarchical clustering, Density based algorithm*

I. Introduction

Data Mining is extracting information from huge sets of data. This data is of no use until it is transformed into valuable information. Data mining will involves in the effective data collection and warehousing as well as computer processing. To segmenting the data and to evaluating the probability of the future events, the data mining uses will sophisticated the mathematical algorithms. Data mining is known as Knowledge Discovery in Data (KDD).

Clustering algorithms are often practical in different fields like data mining, learning theory, pattern recognition to discover clusters in a set of data. Clustering is an unsupervised learning method used for grouping elements or data sets in such a manner that elements in the same group are more similar (in some manner or another) to each other than to those in extra groups. These groups are known as clusters. The Clustering is the subject of active- research in several fields as like statistics, pattern recognition, and machine learning. This paper focused on the clustering techniques in data mining. Data mining adds to the clustering of the complications of very huge datasets with many attributes of different types. This will imposes unique computational requirements on the relevant clustering algorithms. The different clustering algorithms have met these requirements and it have been successfully applied to the real-life data mining problems.

II. Clustering Techniques

Four clustering techniques were considered for evaluation in this paper. It is deployed in different dataset and the data mining tool WEKA is used for the simulation. The description of the clustering techniques was presented in this section.

a. Hierarchical clustering:

Hierarchical clustering will build a tree-based hierarchical taxonomy (a dendrogram) to represent data, each group (or "node") will links to two or more successor groups. The groups are nested and structured as a tree, which ideally ends up as a significant classification system.

Each node in the cluster tree contains a group of similar data and the Nodes group on the graph will point to the next to other, similar nodes. Clusters at one level join with clusters in the next level up, using a degree of similarity; This process will carries until all the nodes are in the tree, which gives a visual snapshot of the data contained in whole set. Total number of the clusters is not predetermined before you start the tree creation.

There are two techniques of hierarchical clustering, Divisive and Agglomerative.

The agglomerative hierarchical clustering algorithm techniques are implemented as follows:

Divisive method
In this method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. to conclude, we continue recursively on each cluster until there is one cluster for each study.

Agglomerative method
In this technique we assign each examination to its own cluster. Then, compute the similarity (e.g., distance) among every of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 awaiting there is only a one cluster left.

Step1.Begin:
Assign number of cluster=number of objects.
Step2. Repeat:
When number of cluster = 1 or specify by user
a) Find the minimum inters cluster distance.
b) Merge the minimum inter cluster.
Step3. End

Divisive hierarchical clustering:

Divisive hierarchical clustering is a top down approach. Divisive hierarchical clustering starts with single cluster that enclose all data objects. Then in every succeeding iteration, it split into the clusters by fulfilling some similarity criteria until every data objects forms clusters its own or satisfies stopping criteria.

Algorithm:
Step1. Begin:
Assign number of cluster=number of objects.
Step2. Repeat: When number of cluster = 1 or specify by User.
a) Find the minimum inters cluster distance.
b) Merge the minimum inter cluster.
Step3. End

b. Expectation–maximization(EM) algorithm:

Expectation-Maximization algorithm is designed to estimate the maximum likelihood parameters of a statistical model in many situations, such as the one where the equations cannot be solved directly. EM is chosen to cluster the data for the following reason among others:

- It has strong statistical basis
- It is a linear in database size
- It is robust to noisy data
- It can recognize the desired number of clusters as input
- It can handle high dimensionality
- It converges fast given a good initialization

c. K-means algorithm:

K-means (MacQueen, 1967) is the simplest unsupervised learning algorithms, it will solve the well known clustering problem. The procedures will follows a simple and easy way to classify the given data set through the certain number of clusters (assume k clusters) fixed a priori. To describe the k centroids, single for each cluster. These centroids should be located in a cunning way because of dissimilar location causes different outcome. So, better choice is to place them as much as possible far away from each other. Next step is to take each point belonging to the given data set and the associate it to the nearest centroid. When there is no point is pending, first step is to completed and an early group age is done. At this point, need to re-calculate k new centroids as bary centers of the clusters resulting from the previous step. After that have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. The result of this loop that the k centroids change their location to step by step until there is no more changes are done. In other terminology centroids do not move any more

Algorithm:
Step 1: Place K points into the space represented by the items that are being clustered. These points correspond to initial group centroids.
Step 2: Assign the each object to the group that has to the closest centroid.
Step 3: When all items have been assigned, compute the place of the K centroids.
Step 4: Repeat Steps 2 and 3 until the centroids no longer progress.

d. Density based algorithm:

A cluster is a dense area of points that is divided by low density regions from the strongly dense regions. This clustering method can be used when the clusters are irregular It happen core substance i.e. objects that have dense neighborhoods. It connects core items and their neighborhoods to structure opaque regions as clusters. Clusters are created as highest sets of density attached points and can identify noise and used when outliers are encountered.

Algorithm :
Step1: Select an arbitrary point r.
Step2: Retrieve the neighborhood of r using 'ε'.
Step3: If the density of the neighborhood reaches to the threshold, clustering process start. Else point is mark as noise.
Step4: Repeat the process until all of the points have been processed.

III. Datasets For Evaluation

This section presents the results of evaluation the four algorithm which has been made using two datasets taken from UCI repository of machine learning.The dataset description are shown below.

A. Iris data

This data set (Fisher, 1936) has often been used as a customary for testing clustering algorithms. This data set has three classes that represents three different varieties of Iris flowers namely Iris setosa(I), Iris versicolor(II) and Iris virginica(III). Fifty samples were obtained from every of three classes, thus a sum of 150 samples is obtainable. Each sample is described by a set of four attributes viz sepal length, sepal width, petal length and petal width. In numerical illustration, two of the classes (virginica, versicolor) have a huge overlap while third is well divided from the other two.

B. Glass Identification Dataset

It has 9 attributes (Refractive index, Sodium, Pottasium, Magnesium, Aluminium, calcium, Silicon, Barium and iron content) and consist of 214 instances of 7 different classes namely Building windows Float processed glass, Vehicle windows float processed glass, Building windows non-float processed glass , vehicle windows non-float processed glass, containers non-window glass, table ware non-window glass and headlamps non – window glass.

IV. Results And Discussion

Among the evaluated algorithms, k-means clustering algorithm is simplest algorithm as compared to other algorithms and its performance is better than Hierarchical Clustering algorithm.

Density based clustering algorithm is not suitable for the data having very huge variations in density and the hierarchical clustering algorithm is more susceptible to noisy data. EM algorithm will takes more time to build cluster when compared to K- Mean, density based clustering algorithms due to this reason the k-mean and density based algorithm are better than EM algorithm. Density based algorithm has takes only less time to build the cluster but it's not better than simple k-mean algorithm since density based algorithm has a high log likelihood value and if the value of log likelihood is high then it will makes bad cluster. Hence we conclude the k-mean is the best algorithm because it takes very less time to build a model when compared to the hierarchal algorithm and cluster instances are also not good in hierarchal algorithm.
TABLE1: Comparison of Various Clustering Algorithms for Iris Dataset.

Algorithm	No. Of clusters	Cluster instance	No of iterations	Sum of squared error	Log likelihood	Time taken to build model(sec)
K-means	3	50(%33) 50(%33) 50(%33)	3	7.8175		0
Hierarchical Algorithm	2	50(%33) 100(%67)				0.08
EM Algorithm	3	48 (32%) 50 (.33%) 52 (.35%)			-2.21077	0.03
Density based Algorithm	2	100(.67%) 50(%33)	7	62.1437	-3.06315	0.02

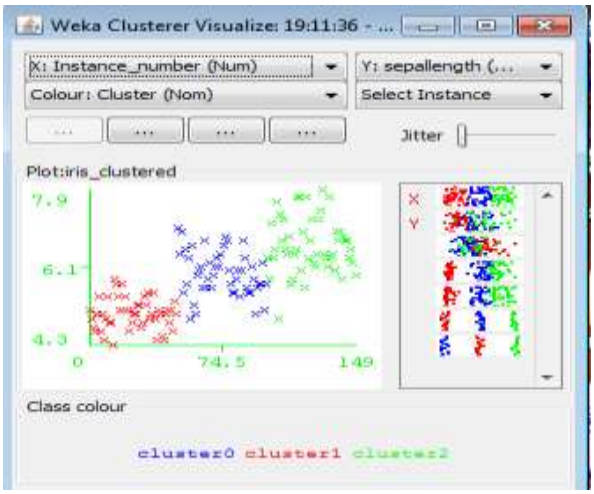

Figure1: Comparison of k-means algorithm for Iris dataset

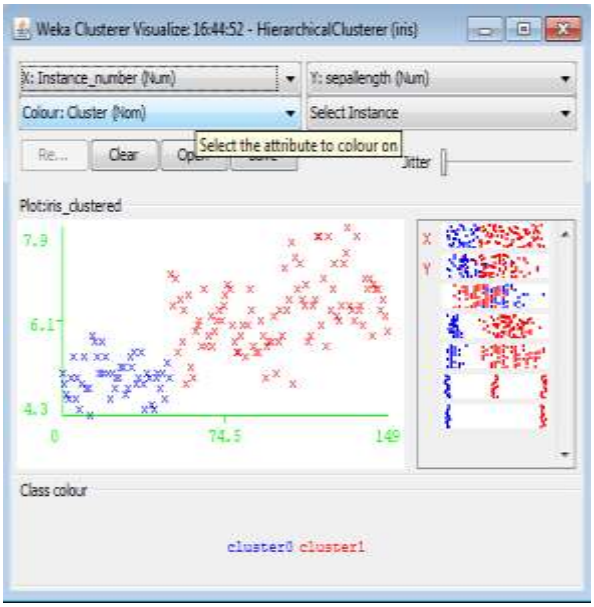

Figure2: Comparison of hierarchical algorithm for Iris dataset

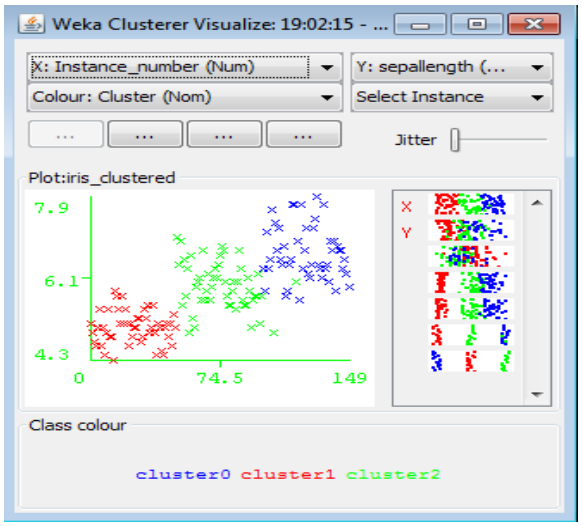

Figure3: Comparison of EM algorithm for Iris dataset

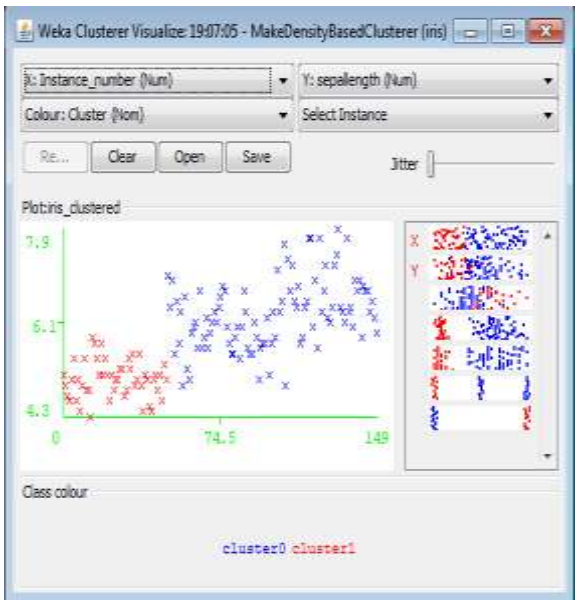

Figure4: Comparison of Density based algorithm for Iris dataset

Table 2: Comparison of Various Clustering Algorithms for Glass Dataset.

Algorithm	No. Of clusters	Cluster instance	No of iterations	Sum of squared error	Log likelihood	Time taken to build model(Sec)
K-means	3	87 (41%) 84 (39%) 43 (20%)	7	77.1543		0.02
Hierarchical Algorithm	2	212 (99%) 2 (1%)				0.14
EM Algorithm	3	54 (25%) 33 (15%) 127 (59%)			-0.10077	0.03
Density based Algorithm	2	144 (67%) 70 (33%)	9	118.2077	0.5539	0

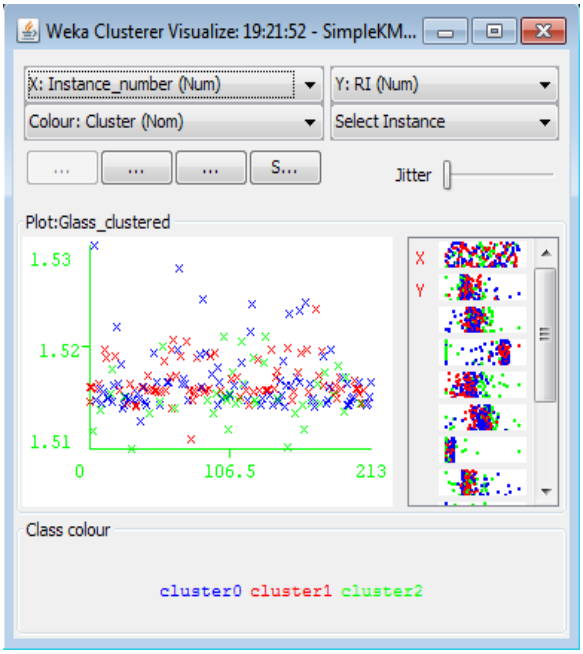

Figure5: comparison of simple k-means algorithm for glass dataset

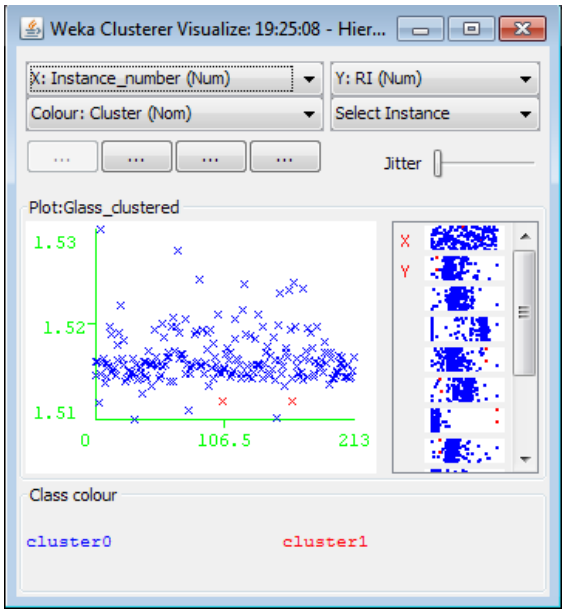

Figure6: comparison of Hierarchical algorithm for glass dataset

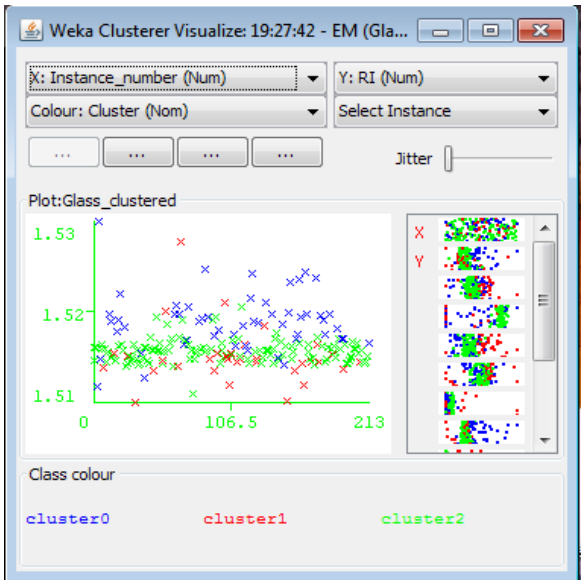

Figure7: comparison of EM algorithm for glass dataset

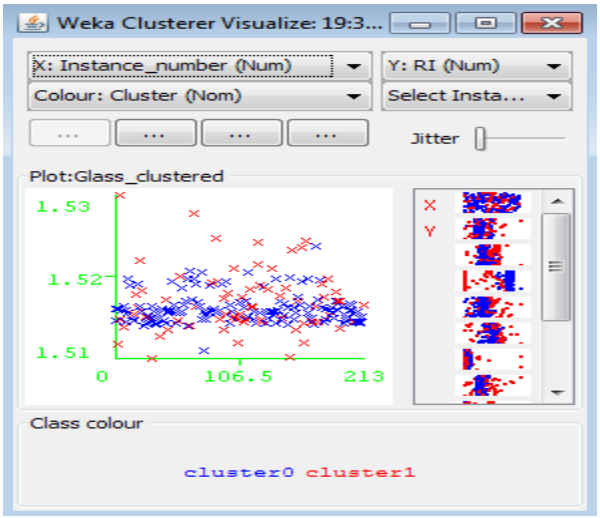

Figure8: comparison of Density based algorithm for glass dataset

V. Conclusion

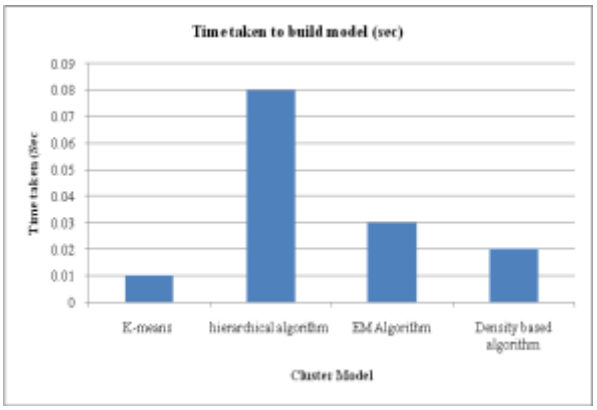

FIGURE 9 (a): Time taken for the IRIS Dataset

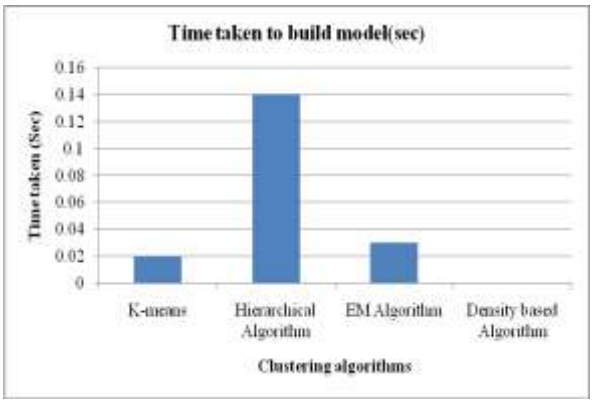

Figure 9 (b): Time taken for the GLASS Dataset

Comparative analysis of various clustering algorithms has been made. The results have been validated using six datasets taken from UCI machine repository. Time taken for the clustering algorithms is depicted in graph at Figure 9. From this paper, we concluded that k-means clustering algorithm is simplest algorithm as compared to other algorithms and its performance is better than Hierarchical Clustering algorithm. Density based clustering algorithm is not suitable for the data having very huge variations in density and the hierarchical clustering algorithm is more susceptible to noisy data.

References:

[1]. R. C. Tryon, Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality. Edwards brother, Incorporated, lithoprinters and publishers, 1939.
[2]. Y. Yang, X. Guan, and J. You, "CLOPE: a fast and effective clustering algorithm for transactional data," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 682–687.
[3]. Aastha Joshi and Rajneet Kaur .: "A Review: Comparative Study of Various Clustering Techniques in Data Mining" IJARCSSE, 2013
[4]. Swasti Singal and Monika Jena: "A Study on WEKA Tool for Data Preprocessing, Classification and lustering", IJITEE, 2013.
[5]. Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides" http://www2.cs.cmu.edu/~awm/tutorials/kmeans.html.
[6]. I. Davidson and S. S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in Knowledge Discovery in Databases: PKDD 2005, Springer, 2005, pp. 59–70.
[7]. Z. Huang."Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery,2:283–304, 1998.
[8]. Chen G JaradatS.,Banerjee N., Tanaka T., Kom and Zhang M: "Evaluation and comparison of clustering Algorithms in analyzing ES cell Gene expression Data ", Statistica Sinica, vol 12.
[9]. Ms S Saranya, Ms P Deepika, Ms S Sasikala, S Jansi, Ms A Kiruthika: "Accelerating Unique Strategy for Centroid Priming in K-Means Clustering" IJIRST (International Journal for Innovative Research in Science & Technology),volume3,pg:40-47
[10]. S.Saranya and Dr. Punithavalli:" An Efficient Centroid Selection Algorithm for K-means Clustering", International Journal of Multidisciplinary Research Academy,vol1,issue3,pg:124-140
[11]. S. Saranya," Evaluation and Efficient Initial Centroid Selection of New Algorithm for High Dimensional Data", International Journal of Computer Science and Mobile Computing,vol:3,issue:6,pg:722-729