An Experimental Study on Clustering Techniques in Data Mining

Hemendra Kumar*[1], Krishna Kant Asopa[2], Shruti Bijawat[3]

*[1]Computer Engineering, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan, India
[2]Computer Engineering, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan, India
[3]Computer Engineering, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan, India

ABSTRACT

Clustering is important in data analysis and data mining applications. Cluster can mean as a conglomerate of data sets which can be seen similar to other data set in the same cluster and also are not similar to the different objects in same clusters.[1]The objective of data mining process is to come out with output of useful and relevant information from a large data set and convert it into an understandable form so that it can be used in future. The Aim of this paper is to identify the high-profit, low error, high efficiency and high-value by one of the data mining technique.

Keywords: Data mining, Simple K means, hierarchical clustering, farthest first

I. INTRODUCTION

Data mining is the phenomenon to analyze the data from different data sources for different perspectives and making the summary of the one into an understandable and meaningful information through various decision producing algorithms. Data mining consists of many functions which have to be performing like it extracts the data and then transform the data, and load transaction data onto the data warehouse system. It store and manage the data in a multi-dimensional database system and present the data in a useful manner and format like a graph or table. It provides the satisfactory data access to business analysts and analyzes the specified data by the application software. Data mining involves the association rule learning, anomaly detection and classification, clustering, summarization and regression. In this paper, we have to do simple clustering analysis by the help of different clustering algorithms [2].Cluster Analysis is a fundamental operation in data and it is an automatic process to find the similar objects from the database. It's important features is that it discovers the patterns in large datasets. To extract the data patterns it used the intelligent methods. Basically in data mining process there are six classes which is anomaly detection, association rule learning, clustering, classification, regression and summarization. There are three stages of clustering in which first raw data is come then clustering algorithm is come after the last stage of clustering is cluster of data.

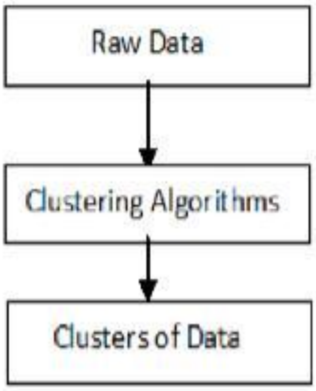

Figure 1. Stages of Clustering

II. LITERATURE SURVEY

Table 1

S. no.	Title of paper	Year of publication	Author's Name	Identified Problem	Methodology
1	Customer Data Clustering	2011	Dr. Sankar Rajagopal	In the real world there are many number of company which having large number of database but it can't manage these dataset.	We apply three algorithm in this paper and purpose of these algorithm is to provide low risk, high value and high profit.
2	Survey paper on Clustering Techniques	2013	Amandeep Kaur Mann and Navneet Kaur	The main problem in the hierarchical clustering algorithm that it don't visit the cluster again after once the visit.	In this survey paper, we have to understand the simply clustering algorithm and analyze the predict results which they produced.
3	Performance Analysis Of Clustering Techniques	2013	Kyle DeFreitas and Margaret Bernard	The main problem in the K Means Algorithm is number of cluster that means we have to define the values of k cluster in starting.	In this paper, we analyze the clustering algorithm and according to that we predict the case based results.
4	Clustering Algorithms in Educational Data Mining	2015	Ashish Dutt, Saeed, and Hamidreza Mahroeian	The main problem is that how the algorithm applied in the education field and produce the result.	The main aim behind in this paper is that to produce the low risk and high profit when we apply the clustering algorithm.
5	Lung Cancer Data Analysis by K-means and Farthest first clustering algorithms	2015	A. Dharmarajan and T. Velmurugan	The main problem of this paper is to identify the how the clustering algorithm apply in the medical field and to identify the yield of the field.	The final outcome of this paper is to analyze the high profit and low risk in the medical field.

III. CLUSTERING ALGORITHM

This Cluster can mean as a conglomerate of data sets which can be seen similar to other data set in the same cluster and also are not similar to the different objects in same clusters. That means the similar data set belong to the same class.To make the clusters for any data set there are so many algorithm which is like hierarchical, K-Means, Farthest First and the Partion Based Clustering Algorithm. These algorithms are mainly used for data mining.

A) Hierarchical Clustering Algorithm-

Hierarchical clustering is one of the method of clustering Algorithm which is used to build a hierarchy of clusters of a particular dataset. The hierarchical algorithms is the connectivity based algorithms of clustering and it mainly build the clusters gradually of a dataset. Hierarchical clustering Algorithm having two types: First is agglomerative methods, which is the bottom up approach that means all the work has to be done the bottom to top fashion and the second is the divisive methods which is the top down approach that means all the work has to be done in the top to down fashion. The Agglomerative hierarchical clustering algorithm is a bottom up approach and the pairs of clusters are club together and become as one and then moves up the hierarchy and here each observation starts in its own cluster. The Divisive hierarchical clustering splits in the recursive manner and the move down the hierarchy.

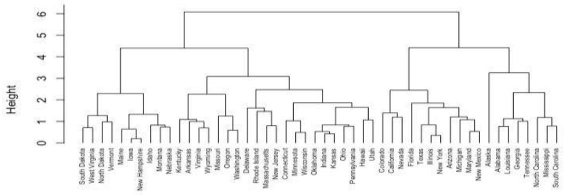

Figure 2. Hierarchical Clustering Algorithm

Advantages of hierarchical clustering

1. It is easy to implement and gives the best result in some cases.
2. Regarding the level of granularity it gives the embedded flexibility.
3. There is no need to require the predefined number of clusters.
4. It accepts any distance which is valid measure.

Disadvantages of hierarchical clustering

1. It is high in the time complexity.
2. It gives error rate high that is efficiency of correctness is low.
3. The main problem in the hierarchical clustering algorithm that it don't visit the cluster again after once the visit.

B) K Means Clustering Algorithm-

K-Means clustering Algorithm divides the n objects of a dataset into the k clusters and the main work in this k mean clustering algorithm is that here each object which having its mean value belongs to the cluster according to the nearest property of value. This Clustering Algorithm will produces the k different clusters which having the good quality. In this algorithm we simply make the two clusters and determine the mean value of each cluster. And analyze the mean value of these cluster set and now we take this mean value and find the new cluster set making this current value of mean value as a reference. and this process has to be continue until the we get two cluster set similar of last two process. That means when we find cluster set of continuously two process is same then our k means algorithm process is completed.

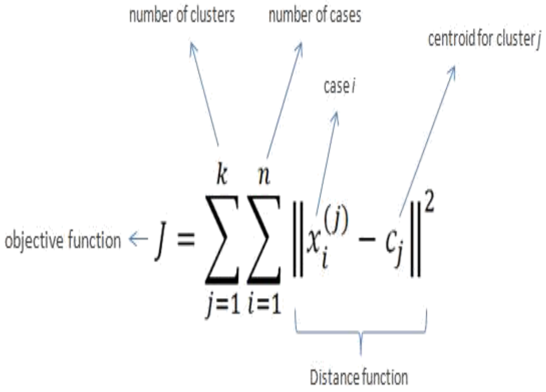
number of clusters
number of cases
case i
centroid for cluster j
objective function ← $J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2$
Distance function

Figure 3

Procedure -

1. Firstly we have to make the k groups of the dataset and here k is the predefined. That mean we make the clusters initially.
2. After that we have to take the k value in the random fashion for making the center of the cluster.
3. And then we assigned the objects according to it's distance that mean which having minimum value from centre having one cluster and which having large value having other cluster. In this we use the Euclidean distance function.
4. Now we calculate the mean value of each cluster for next step.

We have to repeat the steps 2, 3 and 4 until we get same cluster set in consecutive rounds

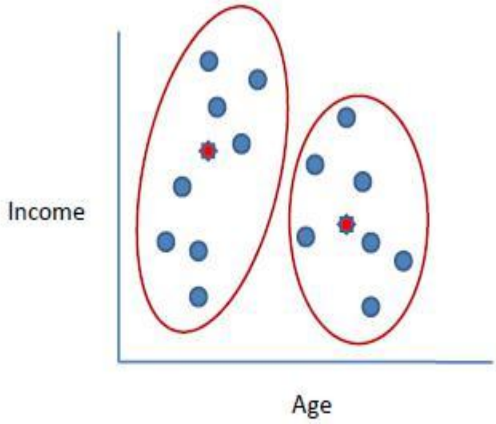

Figure 4

IV. METHODOLOGY

In this paper we have to do a review of clustering and its different Algorithms in data mining. here we used three clustering algorithm in data mining and according to that we analyze the result of these algorithms. The Algorithm which we used in this paper are HIERARCHICAL, K-Means and Farthest First Clustering Algorithm. For determining The Characteristics of these algorithm WEKA tool is used. Weka is a tool which is used for finding the properties and functionality of the algorithms. According to the WEKA tool we analyze the algorithm in terms of time, cluster instance and the efficiency factor. We have to predict the result by taking these factors of the algorithm which we get from the weka tool. Here we take the three different dataset and apply the clustering algorithm and then analyze the result and predict own results of the algorithm according to that factors which we have to take in this paper

IV. RESULT ANALYSIS

Table 2

dataset	Hierarchical Algo			K-Means Algo			Farthest-First Algo		
	Time (sec)	Cluster instance	efficiency	Time(sec)	Cluster instance	efficiency	Time(sec)	Cluster instance	efficiency
Dataset1	0.03	50% & 50%	good	0.02	50% & 50%	best	0.01	68% & 32%	better
Dataset2	6.49	100% & 0%	Not good	0.06	61% & 39%	good	0.05	89% & 11%	better

From Above result analysis table we can see that from all datasets the efficiency of Hierarchical clustering algorithm is not so good and it takes more time to predict the result.[3]And the K-Means clustering algorithm efficiency is good and it gives result in less time as compared to the hierarchical clustering algorithm but the farthest first clustering algorithm gives the result in very less time as compared to all the algorithm and it's efficiency is better. All the algorithm will produce the two cluster instance with different cluster percentages according to its dataset. And the error percentage rate is high in the Hierarchical approach but the correct result with in the short period of time is produced in the Farthest First Clustering algorithm

V. FUTURE SCOPE OF WORK

The main aim of this data mining procedure is to extract information from a large datasets and convert it into an understandable form so that it can be used in future. Clustering algorithm is useful not only for data analysis but for major data mining applications. It is one of the most prominent process of grouping a set of data objects so that the objects, which are similar to each other are usually come in one group and the dissimilar objects are present in other group. Clustering can be performed and executed not only by a particular and specific methodology but also by the different number of algorithms likewise hierarchical, K-Means and Farthest First Clustering Algorithm. Hierarchical clustering is one of the connectivity based clustering approach and Algorithm hence it takes too long time to predict the result of any datasets. And the K Means clustering algorithm is good and it take less time to produce the result But The Farthest First Clustering algorithm takes lesser time as compared to all the other algorithm to produce the results and the efficiency of this algorithm is quite better in terms of the output generation and the correctness of the result is good.

VI. ACKNOWLEDGEMENT

Our thanks to the head of department, Computer Science and Engineering Poornima Institute Of Engineering And Technology and Mr. Abhishek Dadhich for his encouragement in the course of this work. We say thank you all.

VII. REFERENCES

[1] Dr. Sankar Rajagopal "Customer Data Clustering Using Data Mining Techniques", vol.3, No.4, Nov. 2011.

[2] Navneet Kaur,Amandeep Kaur Mann, "Survey Paper on Clustering ",april, 2013.

[3] Ashish Dutt, Saeed Aghabozrgi, Maizatulm, Akmal Binti Ismail and Hamidreza Mahroeian, "Clustering algorithms applied in educational data mining", March 2015.

[4] A. Dharmarajan and T. Velmurugan," Lung Cancer Data Analysis by K means and Farthest First Clustering Algorithms",2015

[5] Amjad Abu Saa, "Education Data Mining & Student'PerformancePrediction", 2016.