IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Comparative Evaluation of Clustering Algorithms on Financial Data set using Data Mining Tool"

Prabhjot , Parminder Singh , Naveen Dhillion

M.Tech Scholar in R.I.E.T, Phagwara

Head of Computer Science Department

Principal at R.I.E.T, Phagwara

Abstract: Data mining is the process is to extract information from a data set and transform it into an understandable structure. There are several major data mining techniques have been developing and using in data mining projects recently including classification, clustering, prediction, sequential patterns and decision tree. With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. Data mining refers to the process of retrieving knowledge by discovering novel and relative patterns from large datasets. Analyzing bank databases for analyzing customer behavior is difficult since Bank databases are multi-dimensional, comprised of monthly account records and daily transaction records. Clustering the datasets, assessment and the way of expressing customer's demands and the provinces of requests should be recognized for providing services to the customers, banks, financial and credit institute. It can make a group of abstract objects into classes of similar objects. In the clustering, firstly partition the set of data into groups based on data similarity and then assigns the labels to the groups. The overall goal of this research work is to evaluate the performance of HAC, K-means and density based clustering (DBSCAN) data mining algorithms by considering the different data sets. HAC is a method of cluster analysis which seeks to build a hierarchy of clusters. It has bottom-up and top-down approach. K-means clustering to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean. Density based clusters are the dense areas in the data space separated from each other by sparse areas. This research presents a comparative analysis for various clustering algorithms. In experiments the effectiveness of algorithms is evaluated by comparing the results on the datasets.

Keywords: - HAC, DBSCAN, K-means, WEKA.

1. INTRODUCTION

Data mining is the extraction of intriguing patterns or information from huge stack of data. In other words, it is the exploration of links, associations and overall patterns that prevail in large databases but are hidden or unknown [1]. Data mining is used in classification, clustering, regression, association rule discovery, sequential pattern discovery, outlier detection, etc. [2]. To overcome the challenges of handling such high-dimensional and varied information, several major data mining techniques have been developed and utilized in data mining projects. These include classification, clustering, prediction, sequential patterns, and decision trees. Clustering, in particular, is a fundamental technique that organizes groups of abstract objects into classes of similar objects. In clustering, an algorithm first partitions a dataset into groups based on data similarity, and subsequently assigns distinct labels to these groups. Data mining is a multi-stage process [3] data is mined by going through various phases, as shown in Figure 1.

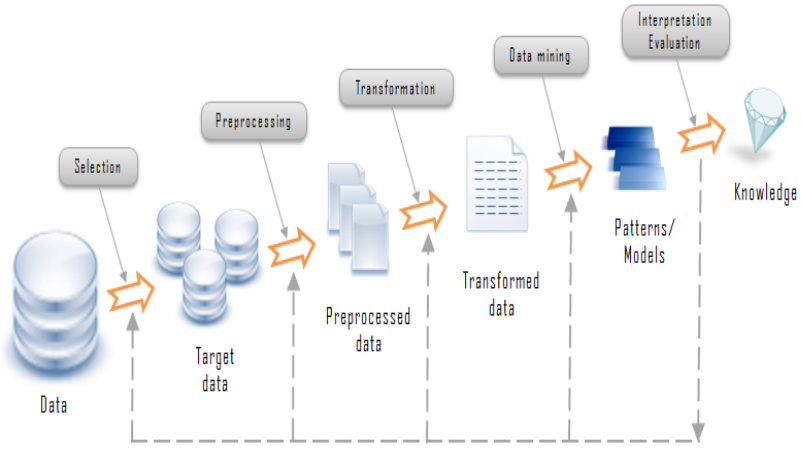

Figure 1: Phases of data mining

Data selection process of extracting valuable information and facts from data has become more an art than science. Even before the data is collected and processed, a preconception of the nature of the knowledge to be extracted from the data exists in the human mind, hence the human intuition remain irreplaceable. Various techniques were developed for the extraction of data, each of them customized for the specific set of information. Clustering is a technique of "natural" grouping of the un-labeled data objects in such a way that objects belonging to one cluster are not similar to the objects belonging to another cluster. It can be considered as the most essential and important unsupervised learning technique in Data Mining. Clustering is the task of grouping 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, bioinformatics, segmentation and machine learning.

2. CLUSTERING ALGORITHMS

2.1 K-Means clustering

Data clustering refers to an unsupervised learning technique, which offers refined and more abstract views to the inherent structure of a data set by partitioning it into a number of disjoint or overlapping (fuzzy) groups [5].Clustering refers to the natural grouping of the data objects in such a way that the objects in the same group are similar with respect to the objects present in the other groups. There are broadly three types of clustering, namely, Hierarchal clustering, Density based clustering, and Partition based clustering. It follows as: first randomly select K the objects as mean (center) of clusters. After that all objects are assigned to the K clusters which have minimum Euclidean distance between objects and cantroids. Mean is updated until all the objects are assigned as mean. This updation is continuing until the assignment is stable.

The algorithm follows a straightforward iterative process:

1. **Initialization** – Choose the number of clusters *K* and randomly initialize *K* cluster centroids.
2. **Assignment** – Assign each data point to the nearest centroid, typically using Euclidean distance as the similarity measure.
3. **Update** – Recalculate each centroid as the mean of all points assigned to that cluster.
4. **Iteration** – Repeat the assignment and update steps until centroids stabilize or a maximum number of iterations is reached.

2.2 HIERARCHICAL Clustering

Hierarchical Clustering method merged or splits the similar data objects by constructing hierarchy of clusters also known as dendogram[6]. Hierarchical Clustering method forms clusters progressively. Hierarchical Clustering classified into two forms.

1. Agglomerative (Bottom-Up) Approach:
This is the most common method. It starts with each data point as its own singleton cluster. At each step, the two closest clusters are merged into a new cluster. This process repeats until only one cluster remains, containing all points. The sequence of merges forms a dendrogram — a tree diagram that records the history of merges.

2. Divisive (Top-Down) Approach:
This method starts with all data points in one cluster and recursively splits the most heterogeneous cluster into smaller ones. Divisive methods are computationally more expensive and less common, but they can be useful when a small number of high-level clusters is expected[7].

2.3 DENSITY BASED CLUSTERING

In data mining, clustering aims to group similar data points together. While popular methods like K-means excel at finding spherical clusters, they struggle with arbitrary shapes and are highly sensitive to outliers. Density-based clustering offers a powerful alternative by defining clusters as dense regions of data points separated by sparse, low-density regions (noise).
The core philosophy is simple: for each point in a cluster, its neighborhood of a given radius (ε) must contain at least a minimum number of other points (MinPts). Clusters grow wherever the density exceeds this threshold, making the method highly effective for discovering non-linear, complex structures[8].

The DBSCAN Algorithm

The most iconic density-based algorithm is **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**. It classifies every data point into one of three categories:

1. **Core Point:** A point that has at least MinPts points within its ε-radius (including itself). These are the "interior" points of a dense region.
2. **Border Point:** A point that is within ε of a core point but does not itself have MinPts points in its neighborhood. These form the "edges" of a cluster.
3. **Noise Point (Outlier):** Any point that is neither a core nor a border point.

How DBSCAN works: Start from any unvisited point. If it is a core point, create a new cluster and recursively visit all points within its ε-neighborhood. Any reachable core points and their border points are added to the same cluster.If the starting point is not a core point, temporarily label it as noise (though it may later be reclassified as a border point if discovered by another cluster).Repeat until all points are visited[9].

3. METHODOLOGY

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

Step 1. Choose the clustering algorithms: To perform the comparative analysis, three clustering algorithms are chosen namely K-means, Hierarchical and Make Density.

Step 2. **Choose the dataset:** The “Bank” data set has been chosen from specific location where it is stored. The file format is .CSV.

Step 3. Load data on WEKA: Load data file for further analysis.

Step 4. Normalize data: After loading of the dataset the next step is to normalize the dataset using the WEKA tool through filter tab. Select normalize filter and apply on the same data set. Save the result using save button.

Step 5. Apply clustering algorithms: Apply the all clustering algorithms on unnormalize as well as normalize dataset.

Step 6. Store the result: After running all algorithms, results are stored into the tabular forms and based on number of iteration, sum of squared error, time taken to build clusters, correctly clustered data, and comparative analysis is performed.

Step7. Plot the graph: Represent results in graphical format.

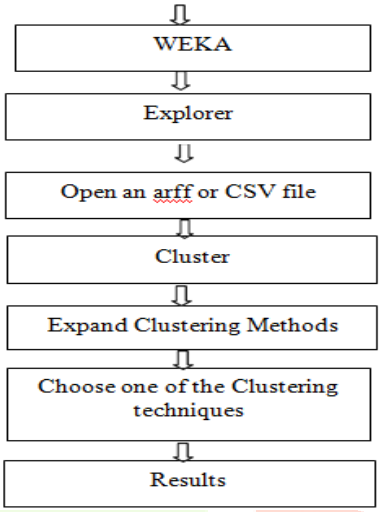

Figure 2: Clustering using WEKA

4. DATASET AND TOOL

4.1 DATA SET

For performing the comparison analysis 'Bank' dataset has been used.

Qualified for rebate
Rate of interest
Interest compound for period
With drawl restrictions
Interest on tax
Loan/advance against deposit
Payment of return
Nomination facility
Premature closer
Payment rule
Transferability
Minimal deposit
Banking services

Table1: Attributes of the Data Set

It is real world data. The dataset is described by the types of attributes, the number of instances stored within the dataset[10]. Banking data are related to customer information and consists of 13 attributes and 5264 instances. In the paper "Bank data" is used in .csv file format. The attributes and their description are given in Table 1.

4.2 Tool

WEKA is a software tool that was developed at the University of Waikato in New Zealand and written on Java [11]. WEKA is platform-independent, open source and user friendly with a graphical interface that allows for quick set up and operation, WEKA is a collection of machine learning algorithms for data mining tasks and its main window is shown in Figure 2. The algorithms can either be applied directly to the dataset or called from your own Java code. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization.

WEKA tool contains Attribute-relationship file format (.arff) and .csv file of the data set. Data set consists of attribute names, types, values and the data. In WEKA, the data objects are called as instances and features of data are considered as attributes [12].

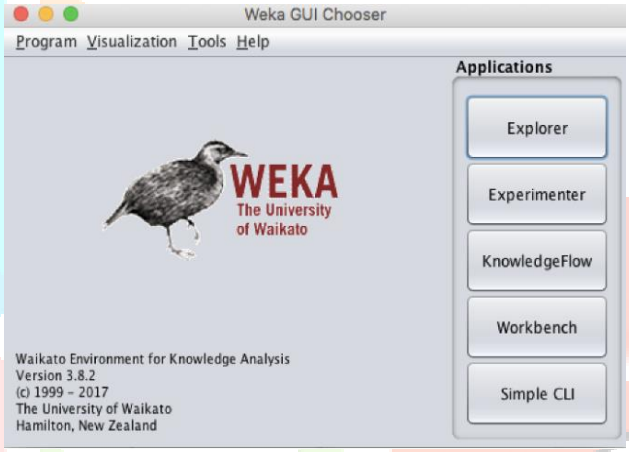

Figure 4.1:Weka Tool

5. EXPERIMENT RESULT

Having introduced the clustering algorithms, now turn to the discussion of these algorithms on the basis of a practical study. This section presents the experimental result of each of the four clustering algorithms using bank data.

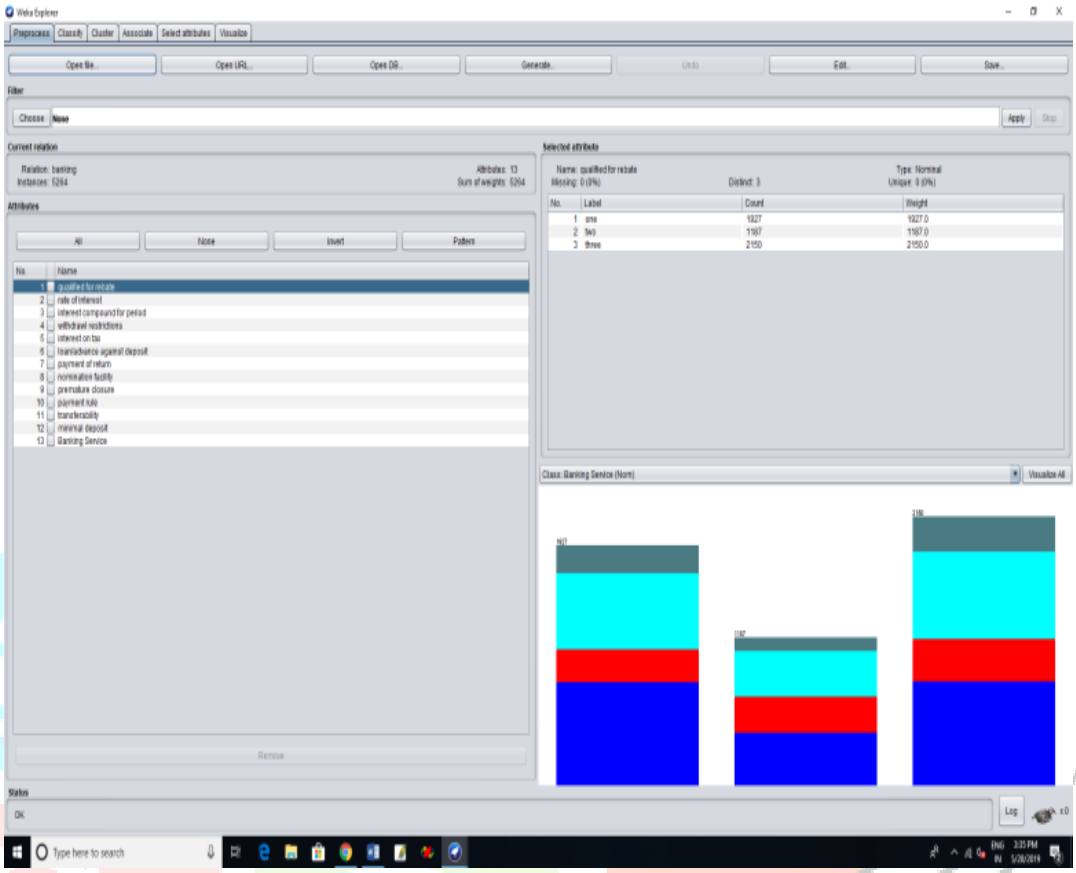

Figure 3: Banking instances

Figure shows the number of banking instances under each type in the dataset is shown in numerically.

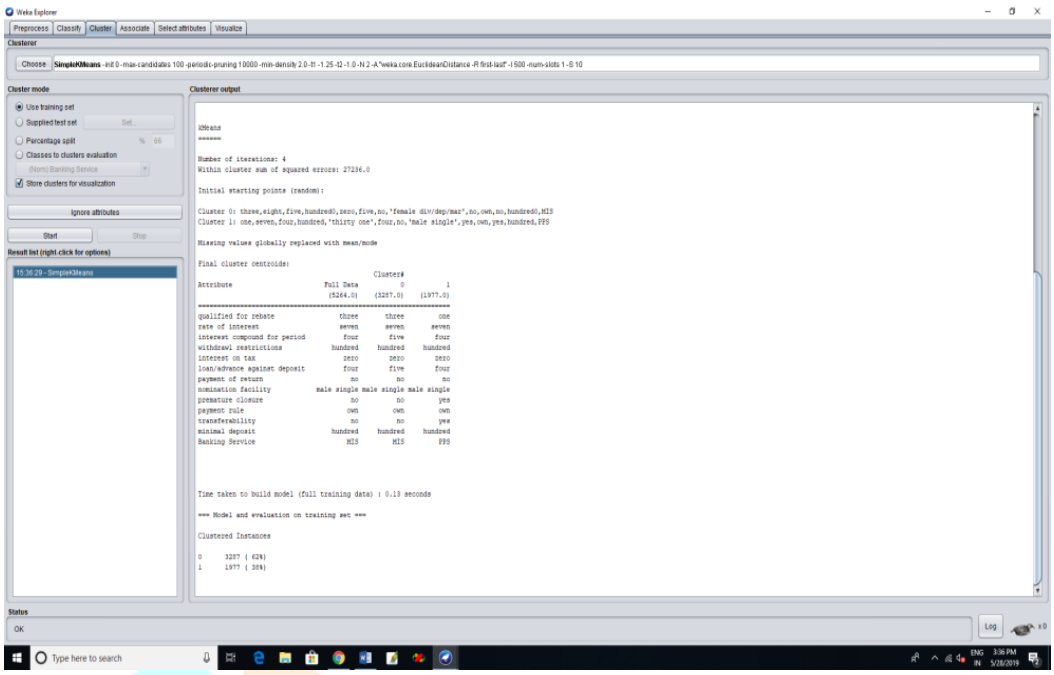

Figure 4: implementation of K-means

Figure shows the implementation of K-means: number of iteration, error rate and number of cluster**.**

RESULT

We compare the clustering algorithm of K-Means, DBSCAN and Hierarchical cluster terms of sum of squared error rate and cluster quality. We can analysis the error rate different clustering algorithms on Banking dataset.

Data set	**Banking**
K-Means	27236
DBSCAN	13663
Hierarchical	12695

Table 5.1: Error rate of clustering algorithm

We recorded sum of squared error rate for K-Means is higher than DBSCAN and Hierarchical clustering algorithm in dataset of Banking.

We compare the clustering algorithm of K-Means, DBSCAN and Hierarchical cluster terms of execution time and cluster quality. We can analysis the execution time different clustering algorithms on Banking data set.

Data set	**Banking**
K-Means	0.13
DBSCAN	0.02
Hierarchical Cluster	0.03

Table 5.2: Execution time of clustering algorithm in Seconds.

We recorded execution time for K-Means is higher than DBSCAN and Hierarchical clustering algorithm in both dataset of Banking.

6. CONCLUSION

In this paper, comparative study has been performed on the K- means, Hierarchical, and Density based clustering algorithms. Comparison is performed on Bank dataset using WEKA tool and the comparative results are presented in the form of table. The comparative study is performed on the basis of accuracy and efficiency parameters. Hierarchical clustering takes more time to form clusters and less accuracy. Density based clustering form clusters with less accuracy as K-means clustering. Simple K-means clustering algorithms forms clusters with less time and more accuracy than other algorithms. In terms of time and accuracy K-means produces better results as compared to other algorithms.

REFERENCES

1. Prakash and Aarohi "Performance analysis of clustering algorithms in data mining in WEKA " IJAET vol. 7 issue 6 ,pp. 1866-1873.

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. AmandeepKaurMann ,NavneetKaur ,"Survey Paper on Clustering Techniques "Volume 2, Issue 4, April 2013 ISSN: 2278 – 7798.

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

5. Thangaraju, Umarani and Poongodi "comparative study of clustering algorithms" IJIRCCE, vol. 5 issue.9 ,September 2017

6. Jiawei Han, MichelineKamber," Data Mining: Concepts and Techniques "second edition.

7. Dr.N.RajalingamK.Ranjini,"Hierarchical Clustering Algorithm - A Comparative Study" Volume 19– No.3, April 2011, ISSN: 0975 –8887.

8. Sharmila, R.C Mishra "Performance Evaluation of Clustering Algorithms" International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013, ISSN:2231-5381.

9. Thomas Schön, "Machine Learning, Lecture 6 Expectation Maximization (EM) and clustering", Available at: http://www.control.isy.liu.se/student/graduate/MachineLearning/Lectures/le6.pdf.

10. S.Revathi,Dr.T.NalinI,"Performance Comparison of Various Clustering Algorithm"Volume 3, Issue 2, February 2013, ISSN: 2277128X.

11. Introduction to Weka, Available at: http://transact.dl.sourceforge.net/sourcefor ge/weka/WekaManual-3.6.0.pdf

12. Data Processing is WEKA is available at: http://facwed.cs.depa.edu/mobasher/classes/etc584/weka.