# A COMPARITIVE PERFORMANCE ANALYSIS OF CLUSTRING ALGORITHIMS IN DATA MINING USING WEKA

**Er. Parminder Singh[1], Vikas[2]**

*[1]Assistant Professor & Head of CSE (RIET,PHAGWARA)*
*[2]M.TECH scholar (RIET, PHAGWARA)*

***ABSTRACT**:-Data mining is the process of discovering patterns in large data sets. It is the analysis step of knowledge discovery in the databases . Clustering play an important role in data mining. 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. The above mentioned objective is achieved by WEKA (Waikato Environment for Knowledge Analysis) machine learning tool as an API (application programming interface). This tool for data pre-processing, clustering, classification and visualization. This research present a comparative analysis for various clustering algorithms. In experiments the effectiveness of algorithms is evaluated by comparing the results on the datasets.*

***KEYWORDS**:-HAC,K-means, DBSCAN, WEKA, API.*

## 1.INRTRODUCTION

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]. Data mining is a multi-stage process [3], data is mined by going through various phases, as shown in Figure 1.

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[1]

In data mining, mining of data can be donusingtwlearningapproaches- Supervised and Unsupervilearning. Clusterinis an unsupervised l idata mining applications. 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.

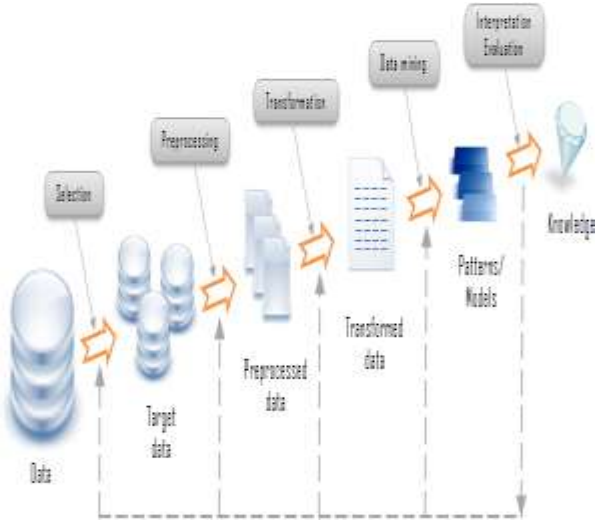

**Figure 1:** Phases of data mining

In this paper, firstly we have discussed the different clustering approaches and techniques used in data mining and then in the later part, we have compared and analyzed few algorithms in terms of accuracy and efficiently. The dataset used for clustering is on banking. Waikato Environment for Knowledge Analysis (WEKA) tool is used to execute the algorithms[12].

## 2. METHODOLOGY

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

**Step1. Choose the clustering algorithms:** To perform the comparative analysis, three clustering algorithms are chosen namely K-means, Hierarchical and Make Density.

**Step2. Choose the dataset:** The "Bank" data set has been chosen from specific location where it is stored. The file format is .CSV.

**Step3. Load data on WEKA:** Load data file for further analysis.

**Step4. 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.

**Step5. Apply clustering algorithms:** Apply the all clustering algorithms on unnormalize as well as normalize dataset.

**Step6. 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.

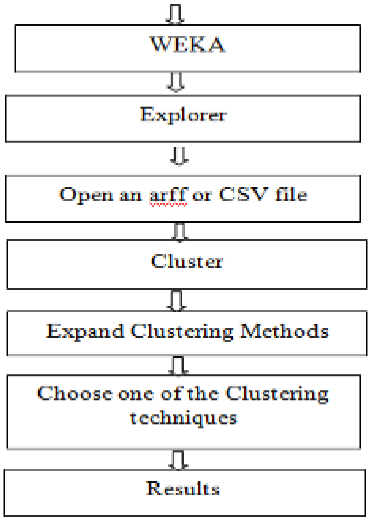
WEKA
Explorer
Open an arff or CSV file
Cluster
Expand Clustering Methods
Choose one of the Clustering techniques
Results

**Figure 2:** Clustering using WEKA

## 3. CLUSTERING ALGORITHMS

### 3.1K-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.

**Algorithm:**

**INPUT**: Number of desired clusters *K*
Data objects D= {d1, d2…dn} OUTPUT:
A set of K clusters

**Steps:**

**Step1**. Begin: Randomly choose k data objects from
Data set D as initial centers.
Number of cluster=K;
**Step2**. Repeat:
a). Assume each cluster ascentriod.
b). Calculate distance of all data points to Centroids.
c). c). Assign data object $d_i$ to the nearest cluster.

**Step3**. Update: For each cluster j (1 <= j<=k),
Recalculate the cluster center.
**Step4**. Until: no change in the center of clusters.
**Step5**. End

### 3.2HIERARCHICAL Clustering

Hierarchical Clustering method merged or splits the similar data objects by constructing hierarchy of clusters also known as dendogram[7]. Hierarchical Clustering method forms clusters progressively. Hierarchical Clustering classified into two forms: Agglomerative and Divisive algorithm.

### 3.2.1Agglomerative clustering

Agglomerative hierarchical clustering is a bottom up method which starts with every single object in a single cluster. Then, in each successive iteration, it combines the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster or specify by the user [7].

**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.
```

**3.2.2 Divisive hierarchical clustering** [7]: Divisive hierarchical clustering is a top down approach. Divisive hierarchical clustering starts with one cluster that contain all data objects. Then in each successive iteration, it divide into the clusters by satisfying some similarity criteria until each 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
```

### 3.3DENSITY BASED CLUSTERING

It is based on the concept of local cluster criterion. Clusters in the data space are considered as the regions with higher density as compared to the regions having low object density (noise). The major feature of this type of clustering is that it can discover cluster with arbitrary shapes and is good at handling noise. It requires two parameters for clustering, namely,

a. ⍰- Maximum Neighborhood radius
b. Min points- Min number of points in the ⍰ neighborhood of that point.

The density based approach uses the concepts of density reach ability and density connectivity [6].

**Algorithm [11]:**

```
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.
```

## 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. 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].

## 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.

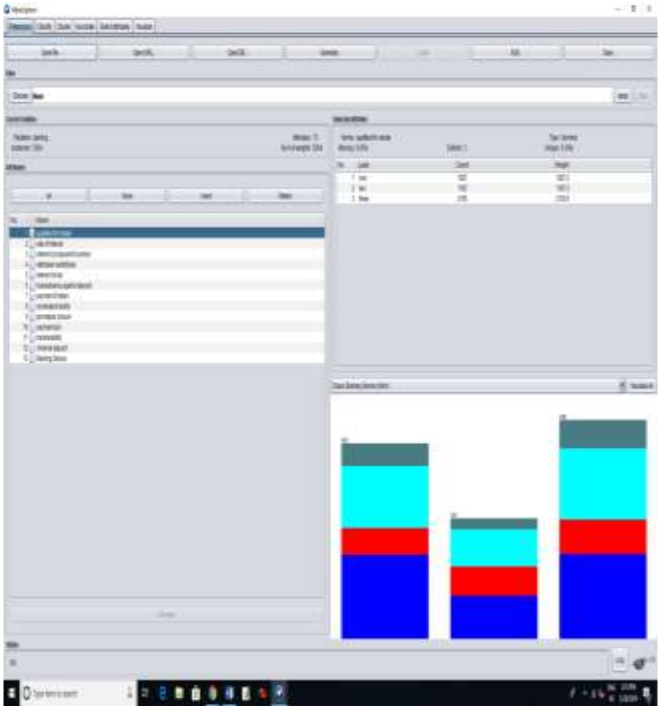

**Figure 3** : Banking instances

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

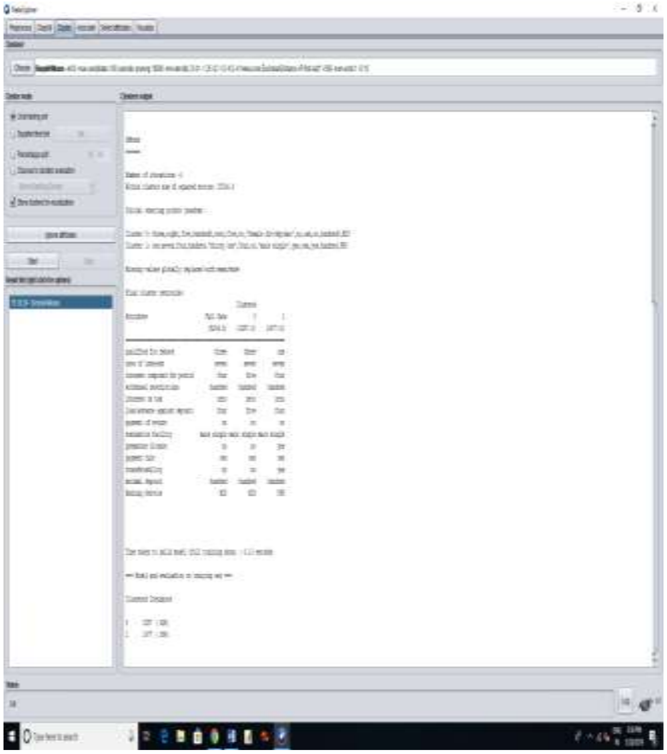

**Figure 4** : implementation of K-means

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

### RESULT

After implementation of these algorithms on data set ,the following results obtained.

| Parameters | Density | HAC | K-means |
|---|---|---|---|
| Number of Cluster | 2 | 2 | 2 |
| Number of Iteration | 4 | 4 | 4 |
| Computational time (seconds) | 0.27 | 0.29 | 0.13 |
| Accuracy (%) | 50.83 | 54.16 | 55.20 |

**TABLE 2:**Comparative result of three Algorithms

For performing comparative analysis, this paper principally focus on the time taken to form clusters, accuracy and number of iterations. Result shows that K-Means algorithm takes lowest time i.e. 0.13 seconds and more accuracy i.e. 55.20%. Distribution of cluster instance is more properly done in Density based algorithm but it takes more time i.e. 0.27 seconds as compare to K-Means. So in terms of efficiency and accuracy K-Means clustering algorithm produce better result as compared to other algorithms.

## 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/sourceforge/weka/WekaManual-3.6.0.pdf

12. Data Processing is WEKA is available at:

    http://facwed.cs.depa.edu/mobasher/classes/etc584/weka