A Comparative Analysis of Clustering Algorithms

Raj bala
Research Scholar (M.Tech)
Amity University
Haryana, India

Sunil Sikka, PhD
Assistant Professor
Amity University
Haryana, India

Juhi Singh
Assistant Professor
Amity University
Haryana, India

ABSTRACT

Clustering is a process of grouping a set of similar data objects within the same group based on similarity criteria (i.e. based on a set of attributes). There are many clustering algorithms. The objective of this paper is to perform a comparative analysis of four clustering algorithms namely K-means algorithm, Hierarchical algorithm, Expectation and maximization algorithm and Density based algorithm. These algorithms are compared in terms of efficiency and accuracy, using WEKA tool. The data for clustering is used in normalized and as well as unnormalized format. In terms of efficiency and accuracy K-means produces better results as compared to other algorithms.

Keywords

Clustering, K-means algorithm, Hierarchical algorithm, Expectation and maximization algorithm and Density based algorithm and WEKA tool.

1. INTRODUCTION

Data mining is a technique to analyze and retrieve knowledge from large amount of database and transform it into useful information for future use [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 retrieves the data from the database that are related to the analysis task. In Preprocessing, data are cleaned and/or integrated. Data transformation, transforms data into appropriate form for mining, by applying summarization or aggregation functions. Data mining is an essential step where intelligent methods are performed in order to extract useful patterns and knowledge. Interpretation/evaluation identifies patterns that representing knowledge based on some measures.

In data mining, mining of data can be done using two learning approaches- Supervised and Unsupervised learning. Clustering is an unsupervised learning in data 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.

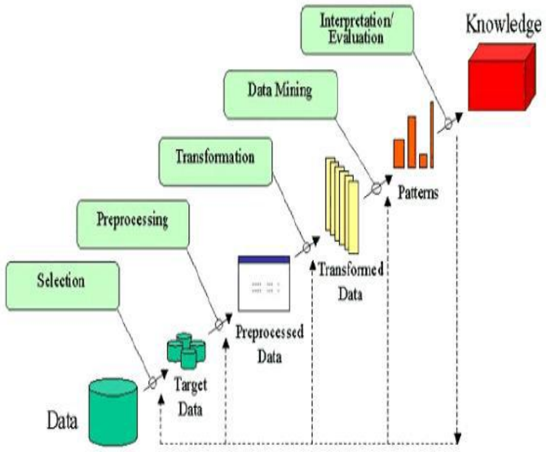

Figure1. Phases of Data Mining [3]

This paper performs a comparative analysis of four clustering algorithms namely K-means algorithm, Hierarchical algorithm, Expectation and maximization algorithm and Density based algorithm. The performance of these clustering algorithms is compared in terms of accuracy and efficiency. The dataset used for clustering has been downloaded from internet [5]. Waikato Environment for Knowledge Analysis (WEKA) tool is used to execute the algorithms.

The rest of the paper is organized into five sections. Section II explains the methodology used in this paper. Section III describes the clustering algorithms compared. Section IV gives description about data set and WEKA tool which are used to compare the algorithms. Section V presents the experimental result in tabular and graphical forms. Finally section VI concludes the paper.

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, four clustering algorithms are chosen namely K-means, Hierarchical, EM and Make Density.

Step2. **Choose the dataset:** The “Bank” data set has been chosen and downloaded from

“http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/preprocess.html” in .CSV file format.

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.

3. CLUSTERING ALGORITHMS

3.1 K-Means Clustering [6]

K-means clustering is a simple partitioning algorithm. It partitioned 'n' data objects into K sets of clusters for resulting the low inter cluster similarity and high intra cluster similarity. Cluster similarity is measured by the mean value of the objects in a cluster, which can be called as the cluster's centroid. 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 centroids. 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 as centriod.
b). Calculate distance of all data points to Centroids.
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
```

3.2 Hierarchical 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.

Agglomerative 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.
```

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
```

3.3 Expectation–maximization (EM) algorithm

This is an iterative method for finding maximum likelihood or Maximum a Posteriori (MAP) estimates of parameters in statistical systems, where the system depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E steps [8].

[9] EM Algorithm focus on the joint log-likelihood function of the observed variables X and the latent variables Z = $\{z_1. . .z_N\}$,

$$l_\theta(X, Z) = \ln p_\theta(X, Z).$$

Algorithm:

```
Step1. Initialize: Set i = 1 and choose an initial$\theta_1$.
Step2. Repeat :( a) Expectation (E): Compute
Q $(\theta, \theta_i) = E\theta_i[\ln p_\theta(Z, X \mid X)]$
            $= \int \ln p_\theta(Z, X) p\theta_i(Z \mid X) dZ$.
(b) Maximization (M): Compute
    $\theta_i+1 = \arg\max Q(\theta, \theta_i)$.
(c) $i \leftarrow i + 1$
```

3.4 Density based algorithm [10]

A cluster is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering algorithm can be used when the clusters are irregular It finds core objects i.e. objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters. Clusters are formed as maximum sets of density connected points and can detect noise and used when outliers are encountered.

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 Dataset

For performing the comparison analysis 'Bank' dataset has been used. The datasets has been downloaded from web [5]. Banking data are related to customer information and consists of 12 attributes and 600 instances. In the paper "Bank data" is used in .csv file format. The attributes and their description are given in Table 1.

Table 1: Attributes of the Data Set

Id	a unique identification number
Age	age of customer in years (numeric)
Sex	MALE / FEMALE
Region	inner-city/rural/suburban/town
Income	income of customer (numeric)
Married	is the customer married (YES/NO)
Children	number of children (numeric)
Car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
Mortgage	does the customer have a mortgage (YES/NO)
Pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

4.2 Tool

WEKA is a software tool that was developed at the University of Waikato in New Zealand and written on Java [12]. 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.

Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

Figure 2: Main Window of WEKA Tool [12]

5. EXPERIMENTAL 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. The experimental results are presented in Table 2 and Table 3.The simulation result is partitioned into several sub items for easier analysis and evaluation. Table 2 summarizes the result of all clustering algorithms with unnormalize data and Table 3 summarizes the result of all clustering algorithms normalize data.

Table 2.Clustering algorithms result for Bank's Dataset without normalize

Algorithm	Number of Clusters	Cluster instance	Number of Iteration	Time	Accuracy

K-means	2	256(43%)
		344(57%)
			4	0.03s	56.66%

Hierarchical Algorithm	2	599(100%)
		1(%)
			4	1.51s	54.16%

EM Algorithm	2	311(52%)
		289(48%)
			4	0.44s	57.83%

Density based Algorithm	2	256(43%)
		344(57%)
			4	0.12s	56.66%

Table 3.Clustering algorithms result for Bank's Dataset with normalize

Algorithm	Number of Clusters	Cluster instance	Number of Iteration	Time	Accuracy

K-means	2	210(35%)
		390(65%)
			6	0.02s	55.20%

Hierarchical Algorithm	2	599(100%)
		1(%)
			4	0.59s	54.16%

EM Algorithm	2	365(61%)
		235(39%)
			4	0.14s	53.83%

Density based Algorithm	2	213(35%)
		387(65%)
			6	0.05s	50.83%

Figure 3 and figure 4 presents the graphical representation of the clustering algorithms result. Figure 3 represents the accuracy result with unnormalize data whereas figure 4 represents the accuracy result with normalize data. The Figure 5 shows the time taken by the clustering algorithms to make clusters when datasets are deployed in the WEKA Tool.

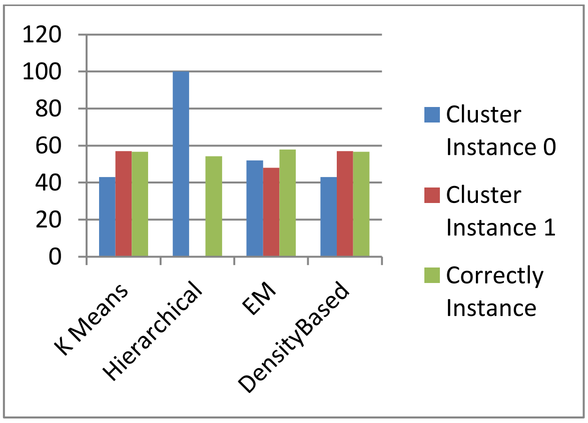

Figure3: Comparison of correctly and incorrectly instances without Normalization filter

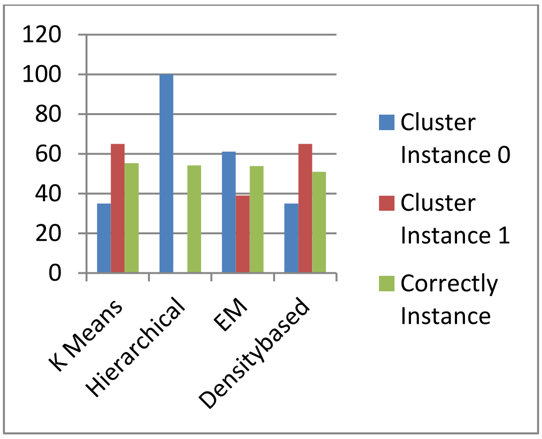

Figure4: Comparison of correctly and incorrectly instances with Normalization filter

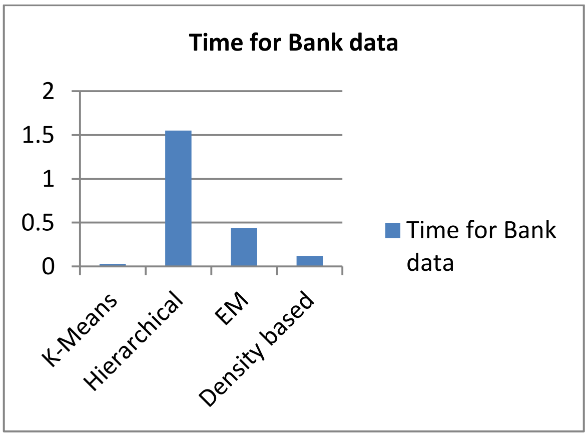

Figure5: Time take by the K-means, Hierarchical, EM and Density Based clustering for datasets

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. 0.03 seconds and more accuracy i.e. 56.66% when data is unnormalized. Distribution of cluster instance is more properly done in Density based algorithm but it takes more time i.e. 0.12 seconds as compare to K-Means. When data is normalized, K-Means takes lowest time 0.02seconds and more accuracy i.e. 55.20% as compared to all other algorithms.

So in terms of efficiency and accuracy K-Means clustering algorithm produce better result as compared to other algorithms with normalized and unnormalized data.

6. CONCLUSION

In this paper, comparative study has been performed on the K-means, Hierarchical, EM 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 and graph. The comparative study is performed on the basis of accuracy and efficiency parameters. Hierarchical clustering takes more time to form clusters and less accuracy with both normalized and unnormalized data. Density based clustering form clusters with equal accuracy as K-means clustering but it takes more time to make clusters with unnormalized data. After apply normalization only 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.

7. ACKNOWLEDGMENT

Many thanks to past and present members of the Waikato machine learning group and the external contributors for all the work they have put into WEKA.

8. REFERENCES

[1] Usama Fayyad, Gregory Piatetsky Shapiro and padhraic Symyh, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communication of the ACM, Vol. 39, No. 11, pp. 27-34,1996.

[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] Data Preprocessing in WEKA, Available at: http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/preprocess.html.

[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: 2277 128X.

[11] Data Clustering Algorithms, Available at: https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithmen.

[12] Introduction to Weka, Available at: http://transact.dl.sourceforge.net/sourceforge/weka/WekaManual-3.6.0.pdf