A COMPARISON OF VARIOUS CLUSTERING ALGORITHMS FOR SAMPLE DATASET

K. Kiruba[1] *B. Rosiline Jeetha*[2]

ABSTRACT

The data mining process is to extract information from the large database, and it is non-trivial process of identifying valid, novel, potentially useful and understandable pattern in data. It is contain the many machine learning algorithms Data mining involves the outlier detection, classification, clustering, regression and summarization. The clustering is the most important technique in data mining, which divides data into groups of similar object. Each group are called cluster. Clustering can be done by using different types of algorithms such as hierarchical algorithm, partitioning algorithm, density based clustering algorithm, Expectation maximization algorithm. We are using zoo dataset from uci repository. In this paper, we have done a comparison of three clustering algorithms that are using zoo datasets are taken for Experimental results on each clustering algorithms.

Keywords

Cluster analysis, hierarchical clustering, partition clustering, density based clustering, Expectation maximization, WEKA tools, sample dataset.

[1]Research Scholar, Department Computer Science, RVS College of Arts & Science, Coimbatore, Tamil Nadu, India. Email : kiruba2092@gmail.com

[2]Associate Professor, Department Computer Applications (MCA), RVS College of Arts & Science, Coimbatore, Tamil Nadu, India. Email : jeethasekar@gmail.com

I. INTRODUCTION

The data mining process is to extract information from the large database, and it is non-trivial process of identifying valid, novel, potentially useful and understandable pattern in data and data mining is sometimes called as data or knowledge discovery is the process of analysing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analysing data. Data mining involves the outlier detection, classification, clustering, regression and summarization [1]. The Clustering is a most important data mining technique in data mining that to group the similar data into a cluster and dissimilar data into different clusters. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. Clustering is the process of organizing the objects into groups whose members are similar in some way. A cluster is a collection of data objects which are similar between themselves and are dissimilar the objects belonging to other clusters. Clustering is the unsupervised classification of patterns i.e. observations, data items, or feature vectors into groups of clusters. Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups.. Clustering can be done by using

different types of algorithms such as hierarchical clustering algorithm such as hierarchical clustering (connectivity based), partitioning clustering (centroid based), density based clustering (density of data points), Expectation maximization(EM) (Iterative based) here we are comparing four different clustering algorithm for using Zoo datasets given experimental results.

II. Clustering

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous sub groups or clusters. Clustering and clusters are not synonymous. A clustering is an entire collection of clusters; a cluster on the other hand is just one part of the entire picture. Clustering is a division of data into group of similar objects. Each group, called cluster consist of objects that are similar amongst themselves and dissimilar compared to objects of other groups. [2]

III. Cluster Analysis

Cluster analysis is the process of grouping data objects based on the information present, the grouping criteria is that objects within a cluster or group be similar to one another and different from objects in another group. Clustering is different from the previously discussed concept of classification in that clustering does not assign labels that have been previously determined to the groups; it only separates the data objects into groups. Any labeling that may occurs does not depend on previously classified data objects that we have access to.

IV. Weka Tool

The WEKA is a data mining tool. The WEKA or Woodhen (Gallirallus Astralis) is an endemic bird of New Zealand. WEKA is (Waikato Environment for Knowledge Analysis)

The WEKA is suite contains a collection of visualization tools and algorithm for data analysis and predictive modelling, together with Graphical User Interface (GUI) for easy access to this functionality. It provides the many different algorithms for data mining and machine learning. WEKA is a open source and freely available it is also an platform independent [5] [6].

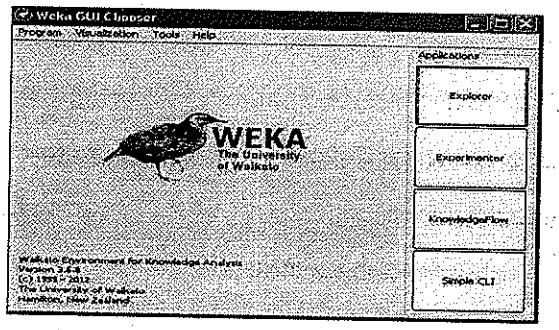

Figure 1 : Front view of WEKA tool

The GUI Chooser consists of four buttons:

- **Explorer:** An environment for exploring data with WEKA.
- **Experimenter:** An environment for performing experiments and conducting statistical tests between learning schemes.

- **Knowledge Flow**: This environment supports essentially the same functions as the Explorer but with a drag and-drop interface. One advantage is that it supports incremental learning.
- **Simple CLI**: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

V. Types Of Clustering Algoritiims

There are different types of clustering algorithms are used by data mining. Such as

a) Partitioning algorithm (centroid based clustering)

b) Hierarchical algorithm (connectivity based clustering)

c) Density based algorithm (density point based clustering)

Expectation Maximization(Iterative Based clustering) [2]

a) Partitioning algorithm : (K-means, K-medoids)

A partitioning clustering is also known as the centroid based clustering. The partition clustering technique partitioning database into a predefined number of clusters. Given a database of n objects and k, the number of cluster to form, a partitioning algorithm organizes the objects into k partition (kd" n), where partition is represents a cluster. The clusters are formed to optimize an objective partitioning criterion, often called a similarity function, such as distance, so that the objects within a cluster are "similar", whereas the objects of different clusters are "dissimilar", in term of the database attributes. There are many methods of partitioning clustering. They are k-means method, k-medoids method, PAM (partitioning around medoids) and probabilistic clustering.

a.1) K-means clustering:

K-means is a widely used partition clustering method in the industries. The k means algorithm is the most commonly used partition clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

K-Means Algorithm : The algorithm for partitioning, where each cluster's centre is represented by mean value of objects in the cluster.

Input : k: the number of clusters. D: a data set containing n objects.

Output : A set of k clusters.

Method :

1. Arbitrarily choose k objects from D as the initial cluster centres.
2. Repeat.
3. (re)assign each object to the cluster to which the object is most similar using Eq. 1, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
5. Until no change.

This figure show that the result of k-means clustering Methods using WEKA tool. After that we saved the result, the result will be saved in the ARFF file format. We also open these files in the ms excel.

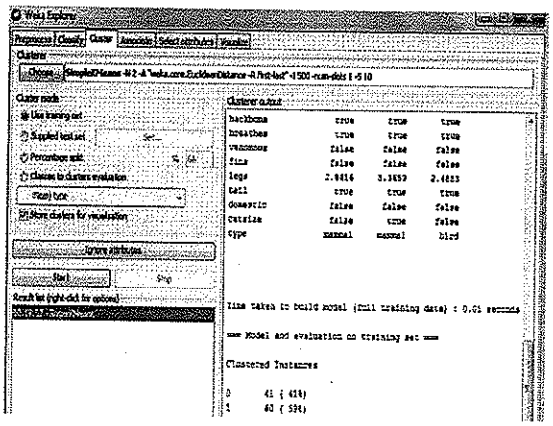

Figure. 2 : Result of k-means clustering

b) Hierarchical clustering :

A hierarchical clustering is also known as the connectivity based clustering. The hierarchical clustering method works by grouping data object into a tree cluster. In hierarchical clustering, which is implemented in the popular numerical software is MATLAB. Then the Euclidean distance are usually used for individual points. Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as dendrogram. [3]

Hierarchical clustering method can be further classified into two types that is given by,

- Agglomerative hierarchical clustering
- Divisive hierarchical clustering

• Agglomerative (bottom up)

1. Start with 1 point (singleton).
2. Recursively adds two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

• Divisive (top down)

1. Start with a big cluster.
2. Recursively divides into smaller clusters.
3. Stop when k number of clusters is achieve

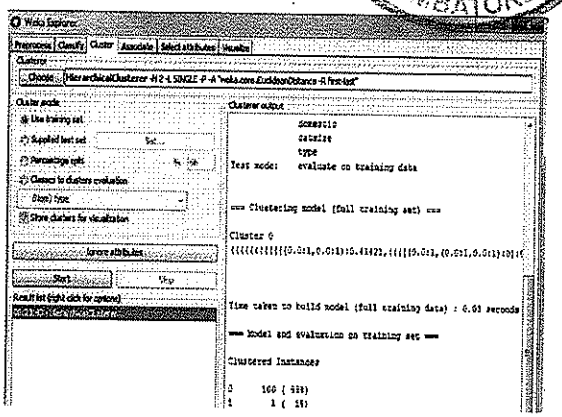

Figure 3 : Result of hierarchical clustering

This figure shows the result of hierarchical clustering method with single linkage between data points using WEKA tool.

c) Density based clustering :

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (Min Pts). One of the most well known density-based clustering algorithms is the DBSCAN. [2]

DBSCAN separates data points into three classes:

- Core points: These are points that are at the interior of a cluster.
- Border points: A border point is a point that is not a core point, but it falls within the neighborhood of a core point.
- Noise points: A noise point is any point that is not a core point or a border point.

To find a cluster,

DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. The algorithm makes use of a spatial data structure(R*tree) to locate points within Eps distance from the core points of the clusters. Another density based algorithm OPTICS, which is an interactive clustering algorithm, works by creating an ordering of the data set representing its density-based clustering structure.

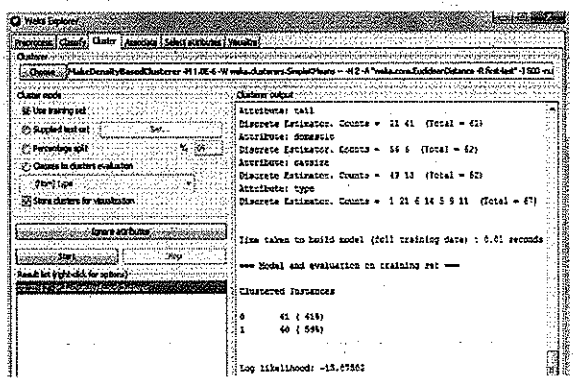

Figure 4 : Result of density based clustering

Above figure shows the result of density-based clustering methods using WEKA tool.

d) Expectation Maximization (EM) :

EM algorithm is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. an expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which computes 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 step.

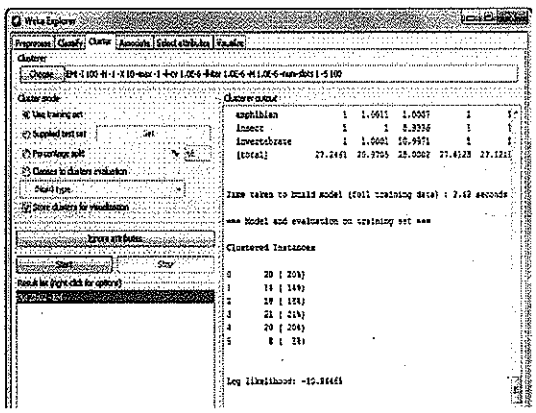

Figure 5: Result of Expectation maximization

VI. Comparison Of Various Clustering Algorithm

Above section involves the study of each of the three techniques introduced previously using WEKA clustering tool on a yellow page data set consist of 5 attribute 19 instances. Clustering of data set is done with each of the clustering algorithm using WEKA tool and the results are:

Table 1: Comparison results of algorithm using WEKA tool

Name	No of cluster	Cluster instances	No of Iteration	With cluster sum of squared error	Time taken to build model	Log likelihood	Unclustered instances
K- means	2	0:41(41%) 1: 60 (59%)	2	424.16	0.01 sec		0
Hierarchical clustering	2	0:100 (99%) 1:1(1%)	-		0.03sec		0
Density based clustering	2	0:41(41%) 1:60(59%)	2	424.16	0.2 sec	-13.87502	0
Expectation Maximization clustering	6	0:20(20%) 1:14(14%) 2:18(18%) 3:21(21%) 4:20(20%) 5:8 (8%)	3		2.42sec	-10.94465	0

VII. Advantages

(i) Increasing number of processors does not affect the computational time for running the algorithm for small values of n.(ROCK)

(ii) Consistently and capture the informative sequence patterns better [1]

(iii) During clustering, we use this goodness measure in order to maximize the criterion function. This goodness measure helps to identify the best pair of clusters to be merged during each step of ROCK [3]

(iv) Results are effective [4]

(v) The incremental clustering algorithms for categorical data should be designed for satisfying practical demand.[4]

VIII. Requirements

The main requirements that a clustering algorithm should satisfy are:

(i) Scalability

(ii) Dealing with different types of attributes

(iii) Discovering clusters with arbitrary shape

(iv) Minimal requirements for domain knowledge to determine input parameters

(v) Ability to deal with noise and outliers

(vi) Insensitivity to order of input records

(vii) High dimensionality

(viii) Interpretability and usability.

IX. Applications

Cluster analysis has a vital role in numerous fields ranging from biology to machine learning. Its application depends on whether clustering is used as a stepping stool and a basis for future analysis or as a tool for understanding. Clustering used in the domains of,

(i) web mining,

(ii) text mining,

(iii) prediction and forecasting,

(iv) computational and systems biology,

(v) biometry etc.

For example, in web mining, clustering the different web documents into coherent groups important for efficient information retrieval.

X. Conclusion

In the recent few years data mining techniques covers every area in our life. We are using data mining techniques mainly in the medical, banking, insurances, education etc.

Before start working with the data mining models, it is very necessary to knowledge of available algorithms. The main aim of this paper is to provide the detailed introduction of WEKA clustering algorithms. WEKA is the simplest data mining tool for classify the data various types. It is the first model for provide the graphical user interface of the user. For perform the clustering we used the sample data repository. It's showing the working of various algorithms in WEKA. The WEKA is more suitable tool for data mining applications. This paper shows only the clustering operations in the WEKA, In future enhancements i will use the WEKA tool for my research work to compare and find various clustering algorithms with real dataset.

References

[1] Michel J.A. Berry Gordon S. Linoff – DM Techniques – Second Edition.

[2] Han J.and Kamber M., Datamining: Concepts and Techniques, Morgan Kaufmann publishers, 2001.

[3] Hierarchical Clustering Algorithm – A comparative study by Dr. N. Rajalingam and K.Ranjini in International Journal of Computer Applications April 2011.

[4] Comparisons between data clustering algorithms in The International Arab Journal of Information Technology, Vol.5, No. 3, July 2008 by Osama Abu Abbas, Computer Science department, Yarmouk University, Jordan.

[5] Proceedings of the 4th National Conference; INDIACom-2010 Computing For National Development, February 25 – 26, 2010 *" K-MEANS CLUSTERING USING WEKA INTERFACE"*

[6] International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013 *" A Study on WEKA Tool for Data Pre-processing, Classificatioz and Clustering"*.

[7] M Roubens, *"Fuzzy clustering algorithms and their cluster validity"*, European Journal of Operational Research Volume 10, Issue 3, July 1982, Pages 294–301

[8] J. C. Dunn, *"A fuzzy relative of the iso data process and its use in detecting compact well-separated clusters,"* J. Cybern. Syst., vol 3, no. 3, pp. 32–57, 1973.

[9] L. Hubert and P. Arabie, *"Comparing partitions,"* J. Classification, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[10] R Tibshirani, G Walther, T Hastie, *"Estimating the number of clusters in a data set via the gap statistic"*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) Volume 63, Issue 2, pages 411–423, 2001

[11] GW Milligan, MC Cooper , *"An examination of procedures for determining the number of clusters in a data set"*, Psychometrika, 1985 – Springer, Volume – 50, Issue – 2, pp 159-179.

Authors Biography

Kiruba Kumarasamy is a M.Phil Scholar in Rathinavel Subramaniam College of Arts & Science, Sulur, Coimbatore which is affiliated to Bharathiar University. She has received her under graduate degree B.Sc (CS) from Rathinavel Subramaniam College of Arts & Science, Sulur, Coimbatore in the year 2009. Received her B.Ed (CS) degree from Dr. G.R. Damodharan College of Education in the year 2010, Coimbatore, and M.Sc (CS) degree from Bharathiar University in the year 2012.

Dr. B. Rosiline Jeetha is a Associate professor in Rathinavel Subramaniam College of Arts & Science, Sulur, Coimbatore. She has more than 10 years of Collegiate Teaching experience and specialized in Data mining and Data warehousing. She has published research papers in International Conferences and Journals. She plays the role of journal review committee member in various National and International journals. She has served as the Moderator for the National conferences and delivered the guest lectures at various colleges. She has produced 6 M. Phil Scholars and guiding 3 Scholars in full and part time stream under Bharathiar University. She has received the Faculty Excellence award for the academic years 2010 –2011 and 2011 – 2012 and citation for the academic years 2012–2013and2013–2014.