International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June 2018 www.irjet.net p-ISSN: 2395-0072

CLUSTERING TECHNIQUES FOR MUSHROOM DATASET

A.Ameer Rashed Khan [1], Dr.S.Shajun Nisha[2], Dr.M.Mohamed Sathik [3]

[1]M.Phil.,Research Scholar, PG & Research Dept. of Comp Science Sadakathullah Appa College,
[2] Asst.Prof & Head,PG & Research Department of Computer Science, Sadakathullah Appa College
[3] Principal,.Sadakathullah Appa College, Tirunelveli, India

---

Abstract - Data mining plays a vital role in our daily life era. All the data has been digitalized so we need to analyze it to make useful information for our knowledge. Classification and Clustering are the two important major techniques used for extracting the data from the database. Clustering is known as the unsupervised learning which is partition a dataset in to a group by their similarities. The objective of this paper is to evaluate the performance of different clustering algorithm such as Expectation Maximization (EM), Farthest Fast and K-means by correctly clustered instances and time taken to build the model for mushroom dataset using data mining tool WEKA (Waikato environment for Knowledge Analysis). The mushroom dataset consists of 8124 instances and 22 attributes with two classes whether it is eatable or poisonous. The dataset is collected from the UCI machine learning repository.

Keywords: Data mining, Clustering, Expectation–Maximization (EM), Farthest Fast, K-means, WEKA.

I. INTRODUCTION

Data mining is the process of extracting knowledge from the database to make it useful information. It is used to discover the patterns and relationship for given data sets. The data sets consists large number of different types of data like categorical or numerical. The application of data mining is becoming popular in every field to maintain the records and in other forms mostly in Business, Education, Medical, Banking and Agriculture etc. It is classified into two types of learning supervised and un- supervised and there are different type's techniques and algorithms are used in data mining. In this paper we concentrate only with clustering techniques and its algorithm. Clustering is one of the major techniques used in data mining to group the similar data in to a cluster and dissimilar data in to other clusters. There are different algorithms used in clustering techniques the major algorithm was Expectation Maximum (EM), Farthest Fast and K-Means under partitioned method. 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 [6]. Farthest Fast performs the process faster because it is modification of K-means that places each cluster centre in turn at the point furthest from the existing cluster centers [6]. K- Mean is fast and efficient algorithm which is easier to understand. It based on distance between the objects and the cluster mean and it is very sensitive to the outliers in data [6]. Mushroom cultivation is becoming popular outcome throughout the world; it is not only the nutritional food, medicinal purposes and additionally adds to the income, particularly those who are having poor land. It becomes hobby for the aged persons as well as homemakers those who can grow mushrooms in small areas and it faces only less difficulties for cultivations. Cultivation of mushroom has been in trend for almost 200 years. Although India has started mushroom farming recently. It is an excellent source of vitamins, minerals, protein and is a good source of iron for anemic patients. There are many characters of mushrooms to analyze suitably with data mining tools. The dataset is collected from the UCI machine learning repository which consists of various numbers of data sets in it. We have chosen only the mushroom data set for analysis.

1.1 RELATED WORK

In our day to day life there will be a large amount of data which is increased as terabytes to petabytes. Companies have spent money for construct data warehouse which contain millions of records and attributes but they are not taking the ROI [11]. So data mining technique is used. Data mining is the process of extracting and discovering of information from data base and it converts in to useful information to make knowledge. It also enhances the knowledge and helping to develop the models that can uncover connections with in millions of billions of records [1]. Data mining involves the analyses of data stored in a warehouse. Three major data mining techniques are classification, clustering and regression [6].Clustering is the process of grouping objects with similar properties [2]. It use various notations to create the groups and these notations can be like as clusters include group with low distance among the cluster members, dense area of the

data space, intervals particular statistical distribution [12].Grouping of objects is done on the principles of maximizing the intra-class similarity such a way the objects in same cluster share some similar properties [8].Clustering is an unsupervised learning observations which coalesce the data in to segments [9].Clustering is the preprocessing step for the other algorithms such as characterization, attribute subset selection and classification [7].Cluster can also be viewed as a special type of classification it can be performed with categorical attributes [10].It is divided in to two methods hierarchical and partition. Hierarchical clustering algorithm group's data objects to form a tree shaped structure. Partition clustering algorithm splits data points into K partition where each partition represent a cluster is done based on certain objective function [2]. Three major algorithms are used under partition method they are Farthest Fast, Expectation–Maximization (EM) and k- mean. Expectation–Maximization (EM) algorithm is an iterative method for findings maximize likelihood or maximum a posteriori (MAP) estimates of parameter in statistical models and the model depends on un-observed latent variables [4]. Farthest Fast algorithms is a modified of k-means that places each cluster center in turn at the point farther most from the existing cluster center [4]. K-mean algorithm is the most commonly used partition algorithm because it can be easily implemented and it is the most efficient one in terms of execution time [3]. WEKA (Waikato Environment Knowledge Analysis) is the widest tools for the data mining. It is an open source interface developed in java [5].

1.2 MOTIVATION AND JUSTIFIFCATION

Mushroom cultivation business has great scope in India. It requires very little land and can be a good source of income for educated youth, home makers and aged person. It has excellent medicinal properties. It is rich in protein, fiber, and amino acids. It is a 100% vegetarian food and is good for diabetes, blood related problems and joint pains. It has no cholesterol and helps in purifying blood. There are different varieties in edible mushrooms. In India button mushrooms are good demand while comparing to other like milky, oyster etc. Clustering is the process of grouping which have similar properties. It can be viewed as a special type of classification techniques. Different algorithms are used for clustering but the major algorithms are farthest fast, Expectation–Maximization (EM) and K-means. The Expectation–Maximization (EM) algorithm gives extremely useful result for the agricultural data set. The farthest first partition based method provides results very fast and suitable for large scale data sets comparing with other algorithm. The K-Means clustering algorithm is efficient for processing larger data set and it is most suitable for numerical data sets. This motivates to work on this algorithm.

II METHODOLOGY

2.1 Outline of the Work

Our objective of data mining techniques is to analyze the performance of partition clustering algorithm such as Expectation Maximum (EM), Farthest Fast and K- Means by correctly clustered instances and time taken to build the model for the given mushroom data set.

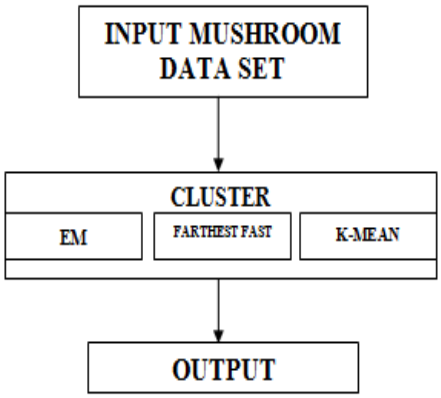

Fig 1: outline of the proposed work

Dataset

The Mushroom dataset has been collected from the UCI Repository. It consists of 8124 instances and 22 attributes in .CSV file format.

Clustering

In data mining one of the important techniques is clustering, which divides the data in to groups by its similarities and dissimilarities data in other groups. Homogeneous data are contains in one cluster and heterogeneous data on other cluster. It is not predefined class and the shapes of the cluster are in dynamic. It is like a database segmentation which groups same instances in a database. It can also be done on large datasets performed with categorical data too. The applications of clusters are used in various business fields like marketing, production, industries, education, agriculture etc. It is not a particular algorithm but common tasks have been done. Different clustering algorithms have been developed which results to a high-quality performance on datasets for cluster structure. Some of the clustering algorithms which are used rapidly k-means, Expectation–Maximization (EM),

Farthest First, Hierarchical clustering etc. In clustering each and every algorithm has its own measures of similarity so the cluster formed by the different clustering algorithms need not be same. One algorithm may make 4 clusters the other may define only 3 cluster groups. The number of clusters formed by the data depends upon the uniqueness and dissimilarity present in the data and the variables considered as metrics for clustering [6].

Expectation–Maximization (EM)

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. The result of the cluster analysis is written to a band named class indices. The values in this band indicate the class indices, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc. The class indices are sorted according to the prior probability associated with cluster, i.e. a class index of '0' refers to the cluster with the highest probability [6].

Farthest First

Farthest first is a modification of K means that places each cluster centre in turn at the point furthest from the existing cluster centers. This point must lie within the data area. The farther points are clustered together first. It performs the process faster because of this modification from K-means in many situations like less reassignment and adjustment is needed. The points are selected randomly. It is suitable for large-scale data mining application which is faster [6].

K-means

K-mean algorithm was first used by James Macqueen in 1967. It is an unsupervised learning, partition clustering algorithm and it is a centroid based algorithm. It is based on the distance between the objects and the cluster mean. It is very sensitive to the outliers in data. It is fast, robust, relatively efficient which is easier to understand. Time complexity of the algorithm is O (tknd), t is the numbers of iterations until optimal clusters are not obtained, k is number of predefined clusters, where n is number of objects in the data set and d is number of attributes/ dimension of each objects. It aims to partition n observations into k clusters which each observation belongs to cluster with the nearest mean value. First choose the k centroid. Then take instances or points belonging to the data set and associate them to the nearest centers. After finding k new centroid, a new binding has to be done between the same data set points and the new nearest center. Process is repeated until there is no change. Finally, this algorithm minimizing intra cluster distance and automatically inter cluster distance is maximized [6].

III EXPERIMENTAL RESULTS:

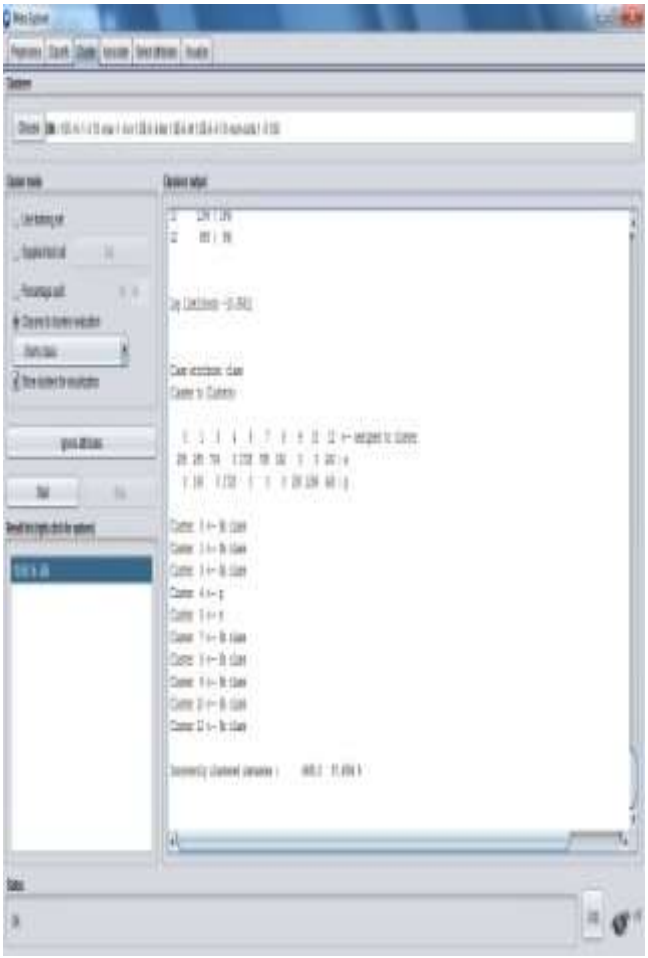

Fig.3.1. Result for Expectation–Maximization (EM) Algorithm

Fig.3.2. Result for Farthest Fast Algorithm

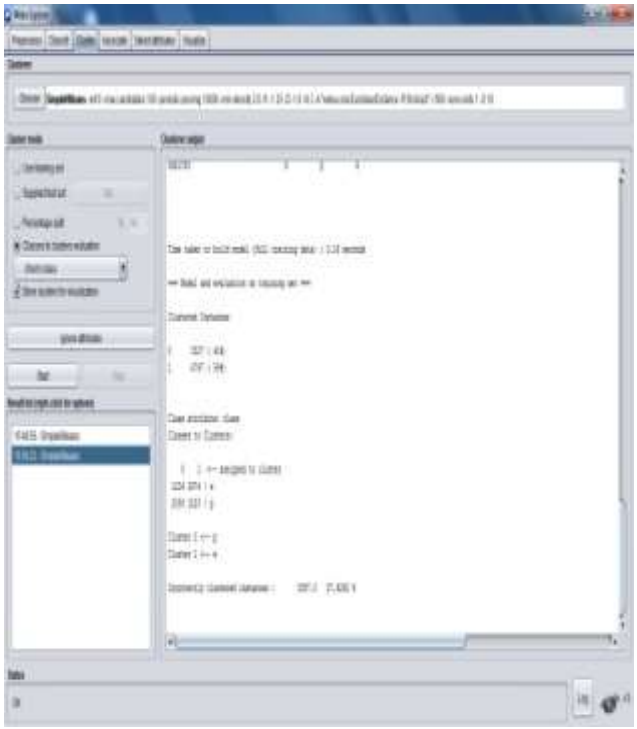

Fig.3.3 Result for K-Means Algorithm

IV PERFORMANCE METRICS:

Performance of Various Clustering Algorithms for Mushroom Data Sets

S.No	Algorithm	Time	Correctly Clustered Instances
1	Expectation–Maximization (EM)	1017.75 sec0	42.5406%
2	Farthest Fast	.08 sec	60.6105%
3	K-Means	0.16 sec	62.3708%

V. CONCLUSION

The performance of the various clustering algorithms is compared based on the correctly clustered instances and time taken to form the estimated clusters. The experimental results of various clustering algorithms to form clusters are depicted as a tabular column. When comparing to the time taken to build the model farthest fast algorithm takes very less time than expectation-maximization (EM) and k-means. Correctly clustered instances refer as a eatable mushrooms. When comparing to correctly clustered instances K-means algorithm gives the best result than farthest fast and expectation-maximization for this mushroom dataset.

VI. REFERENCES

[1] Ameer Rashed Khan and Dr. S. Shajun Nisha, "Comparison Of Classification Techniques Using Mushroom Datasets", Sadakath Research Bulletin, Vol. 0 No.0 Feb- 2018.

2] S.Anitha Elavarasi and J. Akilandeswari, "Survey On Clustering Algorithm And Similarity Measures For Categorical Data", ICTACT- Journal Of Soft Computing, Vol. 04 No.02 Jan- 2014.

[3] Bharat Chaudhari and Manan Parikh, "A Comparative Study Of Clustering Algorithm Using WEKA Tool", International Journal Of Application Or Innovation Engineering And Management, Vol. 01 No.02 Oct- 2012.

[4] Garima Sehgal and Dr. Kanwal Garg, "Comparison of Various Clustering Algorithms", International Journal of

Computer Science and Information Technologies (IJCSIT), Vol. 05 No.03 2014.

[5] Namita Bhan and (Dr.) Deepti Mehrotra, "Comparative Study Of Em And K-Means Clustering Techniques In Weka Inter-Face", International Journal of Advanced Technology & Engineering Research (IJATER), Vol. 03 No.04 Jul - 2013.

[6] Narendra Sharma, Aman Bajpai and Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, Vol. 02 No.05 May – 2012.

[7] Pallavi and Sunila Godara, "A Comparative Performance Analysis of Clustering Algorithms", International Journal of Engineering Research and Applications, Vol. 01 No.03 pp 441-445.

[8] Priyanka Sharma, "Comparative Analysis Of Various Clustering Algorithm Using WEKA", International Research Journal Of Engineering And Technology, Vol. 02 No.04 Jul-2015.

[9] D.Ramya and D.T.V.Dharmajee Rao, "Performance Evaluation of Learning by Example Techniques over Different Datasets", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 03 No.09 Sep- 2014.

[10] S. Revathi and Dr.T.Nalini, "Performance Comparison of Various Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 03 No.02 Feb- 2013.

[11] Shivanjli Jain and Amanjot Kaur, "Performance Evaluation Of Different Clustering Algorithm On Different Datasets", International Journal Of Advanced Research In Computer Science And Software Engineering, Vol. 05 No.10 Oct- 2015.

[12] Suman and Mrs.Pooja Mittal, "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set Using the weak tool", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Vol. 03 No.02 Mar-Apr- 2014.

[13] k. Vanitha and G. Roch libia rani, "Analysis Of Classification And Clustering Algorithm Using WEKA For Banking Data", International Journal Of Advance Research In Computer Science, Vol. 01 No.04 Nov-Dec – 2010.

BIOGRAPHIES .

A.Ameer Rashed Khan,M.Phil Research scholar,currently pusiung at sadakathullah appa college,I had completed my PG at Manonmaniam sundaranar University in Statistics & IT and im a gold medalist and completed my B.Sc., Computer Science at sadakathullah appa college.I had a certification of NPTEL courses.My research area is in Data Mining

Dr.S.Shajun Nisha, Assistant Professor and Head of the PG & Research Department of ComputerScience, Sadakathullah Appa College, Tirunelveli. She has completed M.Phil. (Computer Science) M.Tech (Computer and Information Technology) in Manonmaniam Sundaranar University, Tirunelveli and She had completed Ph.D (Computer Science) in Bharathiyar university,Coimbatore. She has involved in various academic activities. She has attended so many national and international seminars, conferences and presented numerous research papers. She is a member of ISTE and IEANG and her specialization is Image Mining

Dr.M.Mohamed Sathik,Principal Sadakathullah Appa college,Tirunelveli.He has completed Ph.D(Computer science & engineering) Ph.D(Computer science), M.Phil. (Computer Science),M.Tech(Computer Science and Information Technology) in Manonmaniam Sundaranar University, Tirunelveli.He has so far guided more than 35 research scholars. He has published more than 100 papers in International Journals and also two books. He is a member of curriculum development committee of various universities and autonomous colleges of Tamil Nadu.He is a syndicate member Manonmaniam Sundaranar University,Tirunelveli.His specializations are VRML, Image Processing and Sensor Networks.