PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS IN DATA MINING IN WEKA

Prakash Singh[1], Aarohi Surya[2]
[1]Department of Finance, IIM Lucknow, Lucknow, India
[2]Department of Computer Science, LNMIIT, Jaipur, India

ABSTRACT
Data mining is the extraction of intriguing (relevant, constructive, previously unexplored and substantially valuable) 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. In order to perform the analysis, we need software or tools. Weka is a tool, which allows the user to analyze the data from various perspective and angles, in order to derive meaningful relationships. In this paper, we are studying and comparing various algorithms and techniques used for cluster analysis using weka tools. Our aim is to present the comparison of 9 clustering algorithms in terms of their execution time, number of iterations, sum of squared error and log likelihood. Finally on the basis of the results obtained we analyse and judge the efficiency of the algorithms with respect to each another.

KEYWORDS *– cluster analysis, weka tools, data mining algorithms, clustering techniques*

I. INTRODUCTION

At present, the 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-labelled 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 responsible for finding a structure in a group of unlabeled data. There is some sort of similarity among the objects which are present in the same cluster and at the same time, dissimilarity is observed among the objects belonging to different clusters. Clustering algorithms are used to organize, model, categorize and compress data [1]. During the evaluation, the input datasets and the clusters used are varied in number to measure the performance of Clustering algorithms. 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 analysed few algorithms in terms various factors.

II. RELATED WORK

Few of the researchers have improved the data clustering algorithms while others have implemented new ones, and there are few others who have analysed and compared the already existing clustering algorithms. [12] applied various indices to determine the performance of various clustering techniques. The indices were separation scores and homogeneity, WADP, and redundant scores.[13] In their work, they have discussed the technique and have showed the performance of the algorithms with respect to the execution time and speed. [14].In [15] the researchers have used agglomerative technique in order to build dendogram and used simple heuristic method to partition the data. They studied about the similarity based agglomerative clustering algorithms and presented its effectiveness.

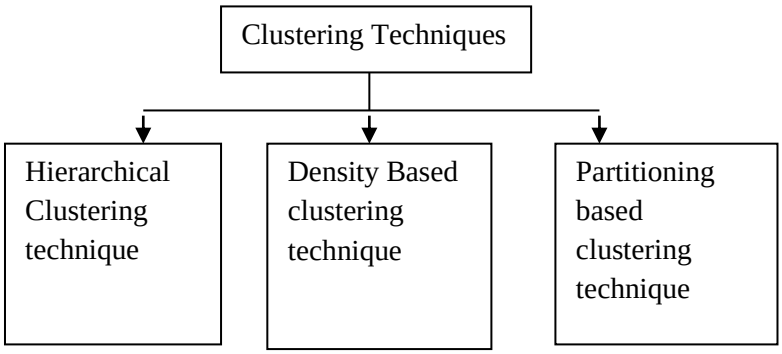
[16] Compared the different clustering algorithms according to the following factors- size of dataset, number of clusters, type of data set and the software used for clustering.

III. CLUSTER ANALYSIS IN WEKA

Data Mining is not exclusively performed by the application of expensive tools and software, as a matter of fact, there is a tool which acts a counterpart to these expensive tools. This tool is called Weka[2], and is the sole toolkit which has been widely used and has remained for an extended period of time. The software is written in JAVA language, and, consists of GUI which is used to connect with the data set file and generate visual outputs, like graphs, tables etc. The clustering technique are of three kinds, namely, Hierarchal methods, Partitioning methods, and density based methods.

IV. CLUSTERING CONCEPTS

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.
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.



4.1. Hierarchical Clustering

In hierarchal clustering, hierarchy of objects is built. There are two types of hierarchal clustering: Agglomerative (bottom up) and divisive (top down). In agglomerative clustering, we start with one data object and gradually build the cluster while in divisive technique; we start with the whole data set and then split the data objects into clusters. The agglomerative technique consists of the following steps [3]:
STEP1: Assign each data object to a cluster, such that each object is associated with only 1 cluster. If we have N data objects, then N clusters are formed, each containing 1 data object.
Step 2: Find the nearest pair of cluster and merge them together to form a pair, so that we are left with N-1 clusters.
Step3: Calculate the distance between the new cluster and each of the old ones.
Step4: Repeat steps 2 and 3 until all the data objects have been clustered into cluster of size N.
Step 3 can be performed by two different methods, linkage technique and metric technique. While linkage techniques specify how the distance between two clusters is measured, metric technique indicates how the distance between two data objects id measured. Linkage technique is further classified into three broad categories: Single Linkage technique, Complete linkage technique and average linkage technique. In single linkage technique the distance between the two clusters is considered to be equal to the shortest distance between any data object of one cluster and any data object of the second cluster, while in complete linkage technique, the greatest distance is considered instead of the shortest distance. In average linkage, the average distance between any object of one cluster and any object of the second cluster is considered.

The second technique to compute the distance is metric technique. It can be implemented in many ways but Manhattan distance and Euclidean distance are the most used techniques. While Manhattan distance considers the sum of the differences of the corresponding data objects, the Euclidean distance is the shortest distance beteen the two data objects, the formulas for both the metric is given below.

$$d = \sum_{i=1}^{n} |x_{i-} y_i|$$

Manhattan formula

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Euclidean Formula

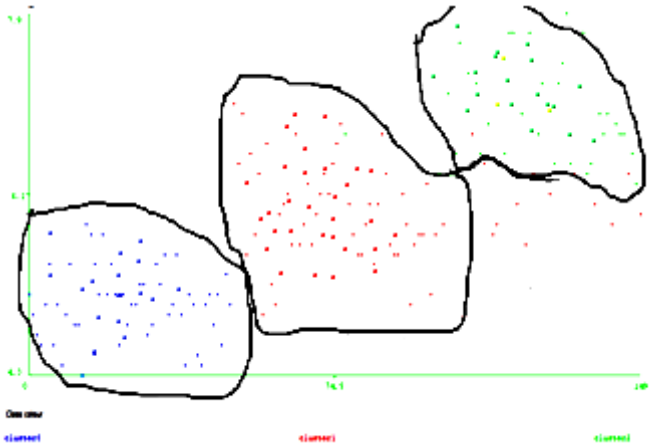


Figure 1. Hierarchical clustering performed on iris data set on weka. X axis- instance number, y axis- sepal length. Blue dots- cluster1, red dots- cluster 2, green dots- cluster 3.

4.2 Partition Based Clustering

It is based on the concept of iterative relocation of the data points between the clusters [4]. The quality of the cluster is measured by the clustering criterion. After each iteration, the iterative relocation algorithm reduces the value of clustering criterion until the point it converges. One of the algorithms based on partition based technique is K-Means algorithm. It is one of the simplest clustering techniques, where mutually exclusive clusters of spherical shaped are built. The clustering process ensures an easy and simple way to cluster the data objects in N number of clusters (specified by the user).

The principal concept of this clustering technique is to designate N clusters to each k data objects. The position of these centroids is very important, since the result may vary if the location of these centroids is changed. So for the best results, the centroids should be placed as far as possible from each other. Next we take each point of the data set and associate it with the nearest centroid. We continue doing it until the time when there is no points pending. After the initial phase of grouping we determine the new centroid of each n clusters. Once we have N new centroids, we start a new process of binding between the original data points and the new centroids. Hence a loop is formed. As a result of the loop formation, the position of N centroids keep on changing until no more change in the position occurs. The goal of the K-Means algorithm is to lessen the objective function. Here the objective function is squared error function [5].

$$z = \sum_{j=1}^{n}\sum_{i=1}^{k} ||x_i^{(j)} - c_j||^2$$

Squared error function

In the formula, $||x_i^{(j)} - c_j||^2$ is the distance between the data point $x_i^{(j)}$ and the cluster centroid c_j.

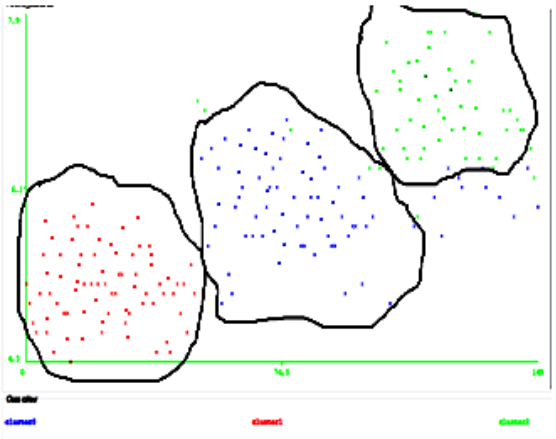


Fig 2. K-Means technique performed on iris data set on weka. X axis- instance number, y axis- sepal length. Blue dots- cluster1, red dots- cluster 2, green dots- cluster 3.

4.3 Density 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. ε- Maximum Neighborhood radius
b. Min points- Min number of points in the ε neighborhood of that point.

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

Density Reachability - A point "a" is density reachable from a point "b" if the point "a" is within a distance of ε from point "b" and "b" has enough number of data points in its neighborhood which are within a distance of ε.

Density Connectivity - A point "a" and "b" are said as density connected if there exists a point "c" which has enough numbers of data points in its neighborhood and both the points "a" and "b" are within the distance of ε. This process is also known as chaining. DBSCAN is one of the clustering techniques which follows the concept of density base notion of cluster: A cluster is a maximal set of density connected data points. It mainly has the following steps[7]:-

1) Start with an arbitrary initial point that has not been visited yet.
2) Take out the neighborhood of this point using ε.
3) If there is adequate neighborhood around this point then the process of clustering begins and point is labelled as visited else this point is marked as noise. (Later this point can even become the part of cluster).
4) If a data point is found to be a part of any cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated iteratively for all ε neighborhood points. The iteration happens until all points in the cluster is defined.
5) A fresh unvisited point is processed, pertaining to the analysis of a further cluster or noise.
6) This process keeps on going until all points are labelled as visited.

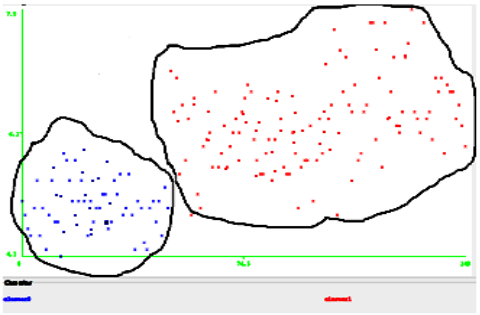


Fig 3: DBSCAN technique performed on iris data set on weka. X axis- instance number, y axis- sepal length. ε- 0.27, min points- 2. Blue dots- cluster1, red dots- cluster 2

V. EXPERIMENTAL SETUP

In our work for the comparison of various clustering algorithms we have used Weka tool. Weka is a data-mining tool which consists of a set of machine learning algorithms. Weka consists of tools for pre-processing, classification, regression, clustering, association rules, and visualization of data. We have chosen three datasets, namely, iris, German credit and Labor data set. In our work we have compared nine clustering algorithms (based on K-mean, Hierarchal, EM, and Density) on the basis of Number of cluster, Cluster instances, Square error, and time taken to build model and Log likelihood.

5.1 Input Data set

The data used in our experiment is real world data obtained from UCI data repository. During evaluation multiple data sizes were used, each dataset is described by the types of attributes, the number of instances stored within the dataset, also the table demonstrates that all the selected data sets are used for the clustering task. These datasets were chosen because they have different characteristics and have addressed different areas.

5.2. Details of the Data Set

We have used 3 data sets which are archived from the UCI ML repository. Table shown below shows the number of attributes and the number of instances in each of the databases.

Table1. Description of the Databases used for the Experiment

Name of the database	Number of attributes	Number of Instances
Iris data set	4	150
German credit data set	20	1000
Labor data set	16	57

5.3. Evaluation:

For evaluation purpose, a test percentage split (holdout method) mode is used.

VI. EXPERIMENT RESULTS

Table 2 below shows the experimental results obtained while comparing the clustering algorithms. Except for the hierarchical clustering, we have to specify the k value (the number of clusters for each algorithm).

Name	**Data set name**	**No. Of clusters**	**Cluster distribution**	**No of iterations**	**Sum of squared error**	**Log likelihood**	**Time taken to build model(sec)**
K means	iris	3	61 (41%) 50 (33%) 39 (26%)	6	6.998		0.02
	German credit	2	643(64%) 357(36%)	5	5365.99		0.03

	Labor	2	50(88%) 7(12%)	3	125.425		0
Hierarchal clustering (birch)	iris	3	49(33%) 1(1%) 100(67%)				0.02
	German credit	2	999(100%) 1(0%)				7.12
	Labor	2	27(47%) 30(53%)				0
EM	iris	3	64(43%) 50(33%) 36(24%)	10		-2.055	0.02
	German credit	2	613(61%) 387(39%)	7		-32.04	0.10
	Labor	2	6(11%) 51(89%)			-18.17	0.08
MTree	iris	3	36(24%) 24(13%) 90(63%)		16.524		1.3
	German credit	2	520(51%) 480(49%)		6382.7		1.8
	Labor	2	50(87%) 7(13%)		89.3		1.6
Farthest first	iris	3	41(27%) 50(33%) 59(39%)				0
	German credit	2	781(78%) 219(22%)				0.02
	Labor	2	42(73%) 15(27%)				0.01
canopy	iris	3	72(48%) 50(33%) 28(19%)				0
	German credit	2	383(38%) 617(62%)				0.05
	Labor	2	37(64%) 20(36%)				0.02
LVQ	iris	3	16(11%) 10(7%) 124(83%)				0.11
	German data	2	836(84%) 164(16%)				2.75
	Labor	2	51(89%) 7(11%)				1.46
Cascading k-mean	iris	3	50(33%) 61(41%) 39(26%)				0.09
	German data	2	525(53%) 475(48%)				7.48
	Labor	2	36(63%) 21(37%)				1.07
DBscan	iris	3	13(25%) 20(39%) 18(35%)				0.01
	German Data	2	336(99%) 4(1%)				0.35
	labor	2	27(87%) 4(13%)				0.01

VII. CONCLUSIONS

Recently data mining techniques have encompassed every field in our life. Data mining techniques are being used in the medical, banking, insurances, education, retail industry etc. Prior to working in the data mining models, it is very important to have the knowledge of the existing essential algorithms.[8] Every algorithm has their own significance and we use them on the nature of the data, but on the basis of this research we concluded that k-means clustering algorithm is simplest algorithm as compared to other algorithms and its performance is better than Hierarchical Clustering algorithm. Density based clustering algorithm is not suitable for data having very huge variations in density and hierarchical clustering algorithm is more susceptible to noisy data. EM algorithm takes more time to build cluster as compared to K- Mean, hierarchical, density based clustering algorithms, that's why k-mean and density based algorithm are better than EM algorithm. Density based algorithm takes relatively less time to build a cluster but it's not better than the k-mean algorithm since density based algorithm has high log likelihood value, if the value of log likelihood is high then it makes bad cluster. Hence k-mean is best algorithm because it takes very less time to build a model. Hierarchal algorithm take more time than k-mean algorithm and cluster instances are also not good in hierarchal algorithm.
Clustering is a vivid method. The solution is not exclusive and it firmly depends upon the analysts' choices. Clustering always provides groups or clusters, even if there is no predefined structure. While applying cluster analysis we are contemplating that the groups exist. But this speculation may be false. The outcome of clustering should never be generalized**.** [9]

VIII. FUTURE WORK

The motive of this paper was to compare some of the clustering algorithms in terms of the execution time, number of iterations, log likelihood and sum of squared error. As a future work, we would attempt to compare all of the algorithms above in terms of different factors other than those mentioned in this paper. One of the approaches could be Normalization which can affect the performance of a clustering algorithm, since we know that the normalized data would produce different result in comparison to the data which is not normalised.

REFERENCES

[1]Aastha Joshi and Rajneet Kaur .: "A Review: Comparative Study of Various Clustering Techniques in Data Mining" IJARCSSE, 2013
[2] Swasti Singal and Monika Jena: "A Study on WEKA Tool for Data Preprocessing, Classification and lustering", IJITEE, 2013.
[3] Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides" http://www2.cs.cmu.edu/~awm/tutorials/kmeans.html.
[4]Alan D Gordon: "An iterative Relocation Algorithm for classifying Symbolic Data". Studies in Classification, Data Analysis, and Knowledge Organization, Springer Link, 2000, pp 17-23.
[5] T hayasaka, N. Toda, S Usui, K Hagaiwara — "least square error and prediction square error of function representation with discrete variable basis", Proceedings of the 1996 IEEE Signal Processing Society Workshop.
[6] Olaf Sporns: "Graph Theory Methods For The Analysis Of Neural Connectivity Patterns", Chapter 12.
[7] Xin Zhou, Richard Luo, Prof. Carlo Zaniolo: "DBSCAN & Its Implementation on Atlas", Spring, 2002.
[8] Ms.Sunita N.Nikam: "THE SURVEY OF DATA MINING APPLICATIONS AND FEATURE SCOPE",ASM's IIBR.
[9] Reza Bosagh Zadeh and Shai Ben-David "A Uniqueness Theorem for Clustering",Stanford.edu.
[10] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis, 14:315–332.
[11] Z. Huang."Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery,2:283–304, 1998.
[12]Chen G JaradatS.,Banerjee N., Tanaka T., Kom and Zhang M: "Evaluation and comparison of clustering Algorithms in analyzing ES cell Gene expression Data ", Statistica Sinica, vol 12.
[13]Masciari E., Pizzuti C, and Raimondo G., "Using an out of core technique for clustering Large Data sets", proceedings of 12th international workshop of database and expert system Application, Munich, Germany.
[14]Li C and Biswas G: "Unsupervised learning with mixed Numeric and Nominal Data",IEEE transactions on Knowledge and Data Engineering, vol. 14.

[15]Osama Abu Abbas: "Comparisons between data clustering Algorithms", IAJIT, Vol. 5, 2008.

AUTHORS

Prakash Singh received his B.Tech degree (Mechanical Engineering) from HBTI, Kanpur (Uttar Pradesh), MBA degree from Lucknow University and PHD degree from BITS, Pilani, Rajasthan. He is currently working as an associate professor at IIM Lucknow. He is also currently the Chairman of Financial Aid and International Linkages Division and also the Chairman of Purchase and Service Contracts Committee at IIM, Lucknow.

Aarohi Surya is a student of B.Tech, Computer Science, LNMIIT, Jaipur (Rajasthan) and is expected to graduate in 2015.