Performance Evaluation of Clustering Algorithms

Sharmila[#1], R.C Mishra[#2]
[1]PG Student, CSE Department, M.D.U Rohtak, Haryana, India
[2] Professor, CSE Department M.D.U Rohtak, Haryana, India

***Abstract*—Data mining is the process of analysing data from different viewpoints and summarizing it into useful information. Data mining tool allows users to analyse data from different dimensions or angles, categorize it, and précis the relations recognized. Clustering is the important aspect of data mining. It is the process of grouping of data, where the grouping is recognized by finding similarities between data based on their features. Weka is a data mining tool. It provides the facility to classify and cluster the data through machine leaning algorithms. This paper compares various clustering algorithms.**

Keywords— Data mining algorithms, Weka tool, K-means algorithm, Clustering methods.

I. INTRODUCTION

Data mining is the use of automatic data analysis techniques to uncover previously undetected relationships among data items. Data mining frequently involves the analysis of data stored in a data warehouse. The major data mining techniques are classification, clustering and regression,. In this research paper we are working on clustering algorithms because it is most important process, if we have a very large database. Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of data mining, and a common technique for statistical data analysis used in many fields, with machine learning, pattern recognition, information retrieval, image analysis, and bioinformatics. I am using weka tools for clustering. It provides a better interface to the user than compare the other data mining tools. The main thing is that we can work in weka easily without having the deep knowledge of data mining techniques. Clustering Algorithms are:

A. Small dataset clustering algorithms

- K-means clustering algorithm.
- EM clustering algorithm.

B. Large dataset clustering algorithms

- Farthest First clustering algorithm.
- Hierarchical clustering algorithm.

II. K-MEANS CLUSTERING ALGORITHM

In data mining, K-means clustering is a method of cluster analysis which aims to partition *n* observations into k clusters in which each statement belongs to the cluster with the nearest mean. This results into a splitting of the data space into Verona cells. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple way to cluster a given data set through a certain number of clusters (assume k clusters) fixed a priority. The main idea is to define k centroids, one for each cluster. These centroids must be placed in a calculating way because of different location causes different result. So, the superior choice is to place them as much as possible far away from each other. The following step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is waiting, the first step is completed and an early group phase is done. At this point we need to recalculate k new centroids of the clusters resulting from the previous step. After these k new centroids, a new obligatory has to be done between the same data set points and the nearest new centroid. A loop has been produced. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move. The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent early group centroids.
2. Assign each object to the group that has the nearby centroid.
3. When all objects have been allocated, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

This creates a separation of the objects into groups from which the metric to be minimized can be calculated.

The problem is computationally difficult (NP-hard), however there are efficient heuristic algorithms that are commonly employed that converge fast to a local optimum. These are similar to the expectation maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Moreover, they both use cluster centres to model the data, however *k*-means clustering inclines to find clusters of comparable spatial extent, whereas the expectation maximization mechanism allows clusters to have different shapes.

III. EM ALGORITHM

EM algorithm [3] 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 hidden variables. The EM [11] 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 used to determine the distribution of the hidden variables in the next E step .EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to generate by cross validation, or you may specify a priori how many clusters to generate.

IV. FARTHEST FIRST ALGORITHM

Farthest first is a modified of K-Means [3] that places each cluster center in turn at the point further most from the existing cluster center. This point must lies within the data area. This greatly increases the clustering speed in most of the cases since less reassignment and modification is needed.

V. HIERARCHICAL CLUSTERING

Hierarchical clustering is a method of cluster analysis which pursues to build a hierarchy of clusters. Approaches for hierarchical clustering generally fall into two types:

A. Agglomerative: This is a "bottom-up" approach. In this approach analyses starts from individual cluster, and pairs of clusters are combined as one moves up the hierarchy.
Agglomerative (bottom up)

1. start with 1 point (singleton).
2. recursively add two or more suitable clusters.

The process stops when k number of clusters is achieved.

B. Divisive: This is a "top-down" approach. In this approach analyses start in one cluster, and separations are performed recursively as one move down the hierarchy. In general, the merges and splits are determined in a greedy manner. The complexity of agglomerative clustering makes them slow for large data sets. Divisive clustering with an extensive search is even worse.
Divisive (top down)
1. Start with a big cluster.
2. Recursively divides into smaller clusters.
The process stops when k number of clusters is achieved.

C. General Steps of Hierarchical Clustering

Given a set of N items to be clustered, and an N*N distance matrix, the basic process of hierarchical clustering is this:

- Start by assigning each item to a cluster, so that if you have N items, you have N clusters, each having just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
- Find the nearby (most similar) pair of clusters and merge them into a single cluster, so that you have one cluster less.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 up to all objects are clustered into K number of clusters.

VI. RELATED WORK

We studied various journals and articles concerning performance evaluation of Data Mining algorithms on various different tools, some of them are described here, Ying Liu et all worked on Classification algorithms while Osama abu abbas worked on clustering algorithm, and Abdullah compared various classifiers with different types of data set on WEKA, we presented their result as well as about tool and data set which are used in performing evaluation.

Ying Liu,wei-keng Liao et al [39] in his article “performance evaluation and characterization of scalable data mining algorithms” explored data mining applications to identify their features in a sequential as well as parallel execution environment .They first establish Mine bench, benchmarking suite containing data mining applications. The selection principle include categories & applications that are commonly used in industry and are likely to be used in the future, there by achieving a truthful representation of the existing applications. Minebench can be used by programmers & processor designers for effective system design.

Osama Abu Abbas [38] in his article “comparison between data clustering algorithms” compared four different clustering algorithms (K-means, hierarchical, SOM, EM) according to the size of the dataset, number of the clusters, type of S/W. The reasons for selecting these algorithms are:

- Popularity
- Flexibility
- Applicability
- Handling High dimensionality

Abdullah et al [41] in his article “A comparison study between data mining tools over some classification methods” showed a comparison study between a number of open source data mining S/W and tool depending on their capability for classifying data correctly and accurately. The methodology of the study constitute of collecting a set of free data mining & knowledge discovery tools to be tested, specify the datasets to be used, and choosing a set of classification algorithm to test the tool performance. For testing, each dataset is described by the used data type, the types of attributes, whether they are certain, real, or integer, the number of instances stored within the data set, the number of attributes that describes each dataset, and the year of data set creation. After selecting the dataset , a number of classification algorithm are chosen that are Naïve Bayes, K-nearest, SVM,C4.5 and also some classifiers are used that are Zero R, One R, and Decision Tree classifier. For estimating purpose two test level modes were used; the K-fold cross validation mode and the percentage splitting mode. After running the four tools, they have obtained some results regarding the ability to run the selected algorithm on the selected tools.

T. velmurgun [27] in his research paper "performance evaluation of K-means & Fuzzy C-means clustering algorithm for statistical distribution of input data points" studied the performance of K-means & Fuzzy C-means algorithms. These two algorithms are implemented and the performance is examined based on their clustering result quality. The performance of both the algorithms depended on the number of data points as well as on the number of clusters. The input data points are produced by two ways, one by using normal distribution and another by applying uniform distribution. The performance of the algorithm was examined during different execution of the program on the input data points. The execution time for each algorithm was also analysed and the results were compared with one another, both unsupervised clustering methods were examined to analyse based on the distance between the various input data points. The clusters were formed according to the distance between data points and clusters centers were designed for each cluster. The implementation strategy would be in two parts, 1.One in normal distribution. 2. Uniform distribution of input data points. The data points in each cluster were shown by different colors and the execution time was calculated in milliseconds. Velmurugan and Santhanam chose 10 (k=10) clusters and 500 data points for experiment. The algorithm was repeated 500 times to get efficient output. The cluster centroid were calculated for each cluster by its mean value and clusters were designed depending upon the distance between data points

VII. CONCLUSION

CLUSTERIG ALGORITHM	**ADVANTAGES**	**DISADVANTA GES**
K-MEANS	1. For large number of variables, K-Means algorithm may be faster than hierarchical clustering, when k is small. 2. K-Means may produce constricted clusters than hierarchical clustering, if the clusters are globular.	1. Difficulty in comparing quality of the clusters formed. 2. Fixed number of clusters can make it difficult to forecast what K should be. 3. Does not give good result with non-globular clusters. Different primary partitions can result in different final clusters.
EM	1.Gives very useful result for the real world data set. 2. Use this algorithm when you want to perform a cluster analysis of a small scene or region-of interest and are not satisfied with the results obtained from the k-means algorithm.	Algorithm is extremely complex in nature.
HIERARCHICAL	1. Does not require the number of clusters to be identified in advance. 2. Calculates a whole hierarchy of clusters. 3. Good result visualizations Joint into the methods.	1. May not scale well: runtime for the standard methods: O(n^2log n) 2. No obvious clusters "flat partition" can be derived. 3.No automatic discovering of optimum clusters.

REFERENCES

[1] Narendra Sharma, Aman Bajpai , Mr Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools" 2012.

[2] Dr.N.Rajalingam, K.Ranjini, "Hierarchical Clustering Algorithm - A Comparative Study" 2011.

[3] Bharat Chaudhari, Manan Parikh "A Comparative Study of clustering algorithms using weka tools" 2012.

[4] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta "A Comparative Study of Various Clustering Algorithms in Data Mining", 2012.

[5] Shi Na, L. Xumin, G. Yong, "Research on K-Means clustering algorithm- An Improved K-Means Clustering Algorithm", "IEEE Third International Symposium on Intelligent Information Technology and Security Informatics",2010.

[6] D. Napoleon and P. G. Laxmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", "IEEE Trendz in Information science and computing", Feb.2011.

[7] Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, and Andre C. Ponce Leon F. de Carvalho ," A Survey of Evolutionary Algorithms for Clustering", IEEE Transction, 2009.

[8] Jiawei Han, Micheline Kamber,"Data Mining:Concepts and Techniques", 2006.

[9] Zhang, T., Ramakrishnan, R., Linvy, BIRCH: "An Efficient Data Clustering Method for Very Large Databases". ACM SIGMOD International Conference on Management of Data, 1996.

[10] Guha, S., Rastogi, R., Shim, K,"An Efficient Clustering Algorithms for Large Database"ACM SIGMOD International Conference on Management of Data, 1998.

[11] Yedla M, Pathakota SR and Srinivasa Enhancing "K-means clustering algorithm with improved initial center" Intl Journal of Computer Science, 2010.

[12] R.Xu and D. Wunsch, "Survey of Clustering Algorithms", "IEEE Transactions on Neural networks", May 2005.

[13] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li, "Automated Variable Weighting in k-Means Type Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.

[14] Ran Vijay Singh, M.P.S Bhatia, "Data Clustering with Modified K-means Algorithm" IEEE-International Conference on Recent Trends in Information Technology, 2011.