Analysis and Significance Study of Clustering Techniques

S.Saradha
Research Scholar, Vels University, Chennai, Tamilnadu, India.

Dr. P. Sujatha
Head & Associate Professor, Vels University, Chennai, Tamilnadu, India.

Abstract – Clustering is a division of data into groups of similar objects. Each group is called as a cluster. Data in the same cluster are similar and to the other are dissimilar the aim of this paper is intended to study and compare different clustering algorithms and to find out the appropriate one. The algorithms under investigation are K-Means, Farthest First, and DBSCAN. Merits and Demerits of above algorithms are also discussed here.

Index Terms – Cluster, K-Means, DBSCAN.

1. INTRODUCTION

Cluster is a classification of data objects into different groups. The objects in a same cluster is similar to one another, dissimilar to other cluster objects. Cluster Analysis is to find the similarities between the objects. Clustering is an unsupervised learning problem. The rest of the paper is organized as follows. Section 2 discusses the related survey work. All Clustering Algorithms described in section 3. In Section 4 compares all algorithms according to their data, application and use. Section 5 describes about the Weka tool. Finally Conclusion and future work is given in Section 6.

2. RELATED WORK

Narendra Sharma[1] says weka tool is the simplest tool for classify the data various types. It is the first model for provide the graphical user interface of the user. And shows how to perform clustering in weka. This paper shows advantages and disadvantages of various algorithms. Concludes k-means clustering algorithm is the simplest algorithm as compared to other algorithms.

W. Sarada [2] compares various clustering techniques like hierarchical, partitioning, density-based based on methodology, structure, model, application and use. Concludes that approaches regardless of performance, each approach have its own benefits.

Chandrakant Mahobia [3], this paper compares two algorithms, viz., the weighted fuzzy c-mean clustering algorithm (WFCM) and weighted fuzzy c-mean-adaptive cluster number (WFCM-AC) for 1000 data chunks. The mean absolute error (MAE) is compared for two algorithms and concludes WFCM-AC is performed better than WFCM.

Maryam Bakshi [4], compares and implements K-Means, Single Linkage, DBSCAN and self-organizing maps. According to this paper, SOM algorithm achieves higher.

Osama Abu Abbas [5], K- Means algorithm, hierarchical clustering algorithm, Self Organization Map algorithm and Expectation Maximization (EM) clustering algorithm are analyzed. And concludes algorithms K-Means and EM performs better than hierarchical algorithm. The performance of SOM algorithm becomes lower in which k- the number of clusters, becomes greater.

Sonam Narwal [6], provides detailed introduction of weka clustering algorithms. shows the advantages and disadvantages of each algorithm. Weka is more suitable tool for data mining application. Found that K-Means is the simplest algorithm as compared with all algorithms.

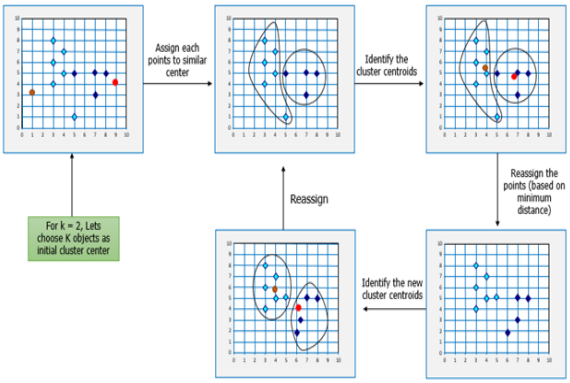
Muhammad Husnain Zafar [7], this paper study and compares five different clustering algorithms like K-Means, Farthest First, DBSCAN, CURE, Chameleon algorithms. They evaluate these algorithms for different data sets and results produced through tables.

3. CLUSTERING ALGORITHMS

3.1 K-Means Algorithm

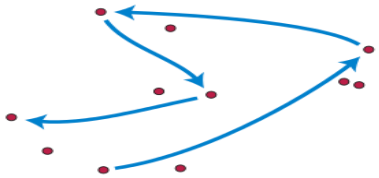
This algorithm is an unsupervised algorithm. It assigns each point to the cluster whose centroid is nearest. It is an algorithm to group the objects based on attributes or features into K number of group, where K is a positive integer. This grouping is done by the minimizing the sum of squares of distances between data and the corresponding cluster centroid. Simplicity and speed with the large number of variables is the main advantages of this algorithm. Another advantage is, it may produce tighter clusters for globular clusters. The disadvantage of this algorithm is difficult to compare quality of clusters produced each time. Because it produce different clusters at each run. Fixed number of clusters also make difficult to predict the K value. This algorithm is does not suitable for non-globular clusters. The steps involved in K-Means algorithm are:

1. Place K objects as group centroids. These are being clustered.
2. Assign all the objects to the groups to its closest centroid.
3. When assigning all the objects, recomputed the centroid of all the cluster.
4. Repeat the step 2 and 3, until the centroids has no longer move.



3.2 Farthest first Algorithm

Farthest first algorithm proposed by Hochbaum and Shmoys 1985, a variant of K-means that places each cluster centre in turn at the point farthest from the existing cluster centres. The point must lie within the data area. This greatly sped up the clustering. In most cases, it needs less reassignments or adjustments. It also defines initial seeds and then on basis of "k" number of cluster which we need to know prior. In farthest first it takes point P i then chooses next point P i which is at maximum distance. P i is centroid and p1, p2, ..., pn are oints or objects of dataset belongs to cluster. Min {max dist(pi,p1),max dist(pi,p2)…….}. The advantage of Farthest-point heuristic based method is fast and suitable for large-scale data mining applications. Disadvantages of Farthest First are same as K-means algorithm .



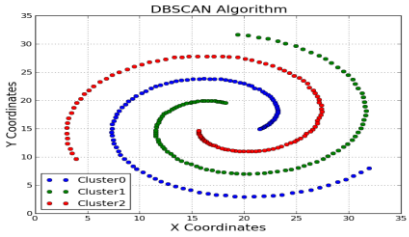
3.3 DBSCAN Algorithm

Density Based spatial clustering for application with noise, proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. This algorithm is based on density of clusters and also have ability to handle noise. Regions with high density points depict the existence of clusters and regions with low density points indicate the noise or outlier of clusters. Grouping of data objects into meaningful subclasses is the main task of this algorithm. Two global parameters of this algorithm are

- Minpts : Minimum number of points in an Eps-neighbourhood of that point.
- Eps: Maximum radius of the neighbourhood.

Steps of DBSCAN is as follows:[7]

- Arbitrary choose an point p.
- Retrieve all points density reachable from p w.r.t Eps and Minpts.
- If p is a core point, then cluster is formed.
- If p is a border point, no points are density reachable from p and DBSCAN visits the next point of the database.
- Continue this process until all the points have been processed.



Advantage of this algorithm is, it does not require any priori specification of the number of clusters, as opposed to K-Means algorithm. It is able to identify the noise while clustering. And also identify the arbitrarily shaped clusters and arbitrarily sized clusters. Disadvantage is, it cannot clusters data sets well with larges differences in densities and in the case of neck type of data set. The algorithm is not partition able for multiprocessor system[7].

4. WEKA TOOL

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java. Weka is a collection of machine learning algorithms for data mining tasks. Weka is used for research, education and applications. Weka is easy to use and to be applied at several different levels. It is a Plaform independent. It consists of four buttons i.e., Explorer, Experimenter, Knowledge Flow, Simple CLI. In Clustering, the records in data base is divided into several groups.

Advantages of Weka include:

1. Free availability under the GNU (General Public License).
2. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
3. A comprehensive collection of data Pre processing and modeling techniques.
4. Ease of use due to its graphical user interfaces.

S.No	Algorithm	Type	Methodology	Structure	Model	Application
1.	K-Means	Partitioning	The center is the average of all the points /objects in the cluster	spherical shaped clusters in small to medium sized data sets	Centroid	Large datasets
2.	Farthest First	Partitioning	It takes point P then chooses next point P1 which is at maximum distance.	Combined shape of spherical and arbitrary clusters	Centre	Large Datasets
3.	DBSCAN	Density-Based	Density of points	Arbitrary-shaped clusters	Density	Density based notion of clusters

5. CONCLUSION

In this paper, various clustering algorithms and also weka tool are analyzed. Weka method is choosing based on choosing the data and attribute. Our work is extended to utilize the implementation of different data set.

REFERENCES

[1] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering
[2] W. Sarada, "A Review on Clustering Techniques and their Comparison", IJARCET, ISSN: 2278 – 1323.
[3] Chandrakant Mahobia, M. Kumar, "Performance Comparison of Two Streaming Data", IJCIT
[4] Maryam Bakshi, Mohammad-Reza Feizi-Derakhshi, Elnaz Zafarani,
[5] Osama Abu Abbas, " Comparison between clustering Algorithms", The International Arab Journal of Information Technology, vol.5, No.3, July 2008.
[6] Information Technology, Vol.5, No.3, July 2008.
[7] Sonam Narwal, Kamaldeep Mintwal, "Comparison the Various Clustering and Classification Algorithms", ijarcsse, ISSN:2277 128X.
[8] Muhammad Husnain Zafar, Muhammad Ilyas, "A Clustering Based Study of Classification Algorithms", International Journal of Database Theory and Application, Vol.8, No.1, pp: 11-22.
[9] Michael Steinbach, "A Comparison of Document Clustering techniques",
[10] Namrata. M, Prajwala T.R, "A Comprehensive Overview of Clustering Algorithms in Pattern Recognition", IOSRJCE, ISSN : 2278-0661.
[11] Rui Xu, Donald Wunsch, " Survey of Clustering Algorithms", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.
[12] Dr. Sudir B. Jagtap, Dr.Kodge B.G., "Census Data Mining and Data Analysis Using Weka", ICETSTM-2013.
[13] Dr. P. Sujatha, S. Saradha, "A Study of Data mining concepts and techniques", IJAER, ISSN 0973-4562,Vol.9, No.27(2014).