Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.273 – 277

AN EXTENSIVE ANALYSIS ON VARIOUS CLUSTERING ALGORITHM IN DATA MINING

D.Osmond Niranjan Giftson[1], R.Jemina Priyadarshini[2], M.C.A., M.Phil., Ph.D

[1]Research Scholar, Department of Computer Science, Bishop Heber College, Trichy, Tamil Nadu, India

[2]Associate Professor, Department of Computer Science, Bishop Heber College, Trichy, Tamil Nadu, India

[1] niranjand26@gmail.com; [2] jemititus@gmail.com

Abstract—Data analysis is used as a common method in modern science research, which is across communication science, computer science and biology science. Clustering as the basic components of data analysis plays significant role. The grouping of similar data items into cluster is called data clustering. Clustering is one of the unsupervised learning techniques. A clustering algorithm segregates the data set into several clusters. In this survey work the different types of clustering techniques are deeply dealt in different perspectives of its merits and demerits. The various clustering techniques considered in this work are k-means, DB Scan, Density Based, Optics and EM algorithms. All the algorithms were tested using particular data sets and the results are calculated and presented. Finally, from the study show that k-means clustering using effective time complexity produces better result in efficient way.

Keywords—Data Clustering, Groups, k-means, DBScan, Density Based, Optics and EM

I. INTRODUCTION

Cluster is a group of objects that belong to the same behavior. The similar objects are grouped in one cluster and dissimilar objects are grouped in other cluster [1]. Clustering Analysis is broadly used in many applications such as market analysis, recognition of pattern, analysis of data and image processing. Clustering is a process of grouping objects with same properties [2]. Any cluster should contain intra-class similarity. The clustering techniques are one of the unsupervised learning algorithms i.e. it learns by observation rather than examples. Even though there are many clustering algorithm exists, there is no single algorithm can handle all type of requirements. It is very difficult to select particular clustering algorithms for particular task.

This paper describes about the genaral working principles, the methodologies to be and the parameters which is used in these clustering algorithms. This paper also presents the comparison of the some popular clustering algorithms to provide the idea for selecting particular clustering algortithm for a specific task.

II. LITERATURE REVIEW

Data mining is a multi-step procedure. It needs accessing and preparing data for a data mining algorithm, mining the data, analyzing results and taking appropriate action. The accessed data that can be stored in one or more operational databases, a data warehouse or a flat file. In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised clustering. A cluster is a group of data objects that are similar to one another within the same cluster and are different to the objects in other clusters. A good clustering algorithm is able find out the clusters regardless of their shapes.

PallaviPurohit and Ritesh Joshi et.al [1] proposed an enhanced approach for traditional K-means clustering algorithm due to its certain limitations. The poor performance of traditional K-means clustering algorithm is selection of initial centroid points randomly. The proposed algorithms deal with this problem and improve the performance and cluster quality of traditional k-means algorithm. The enhanced algorithm selects the k initial centroids in an efficient manner rather than randomly selecting. It first discovers the closest data objects by calculating Euclidian distance between each data objects and then these data points are deleted from population and forms a new data set. The enhanced algorithm provides more precise results and also reduces the mean square distance. But the proposed algorithm works better for dense dataset rather than sparse data set.

The enhanced k-means clustering algorithm basically consists of three steps. The first step talks about the construction of the dissimilarity matrix. Secondly, Huffman algorithm is used to create a Huffman tree according to dissimilarity matrix. The output of the Huffman tree gives the initial centroids. Wang Shunyeet et. al [2] proposed finally the k-means clustering algorithm is being appropriate to initial centroid to get k clusters as output. Wine and Iris datasets are selected from UIC machine learning repository to test the enhanced algorithm. Proposed algorithm gives better accuracy rates and results than the traditional k-means clustering algorithm.

Fahim A.M, Salem et. al [3] suggested an efficient k-means clustering algorithm to prevail over problems in traditional k-means. Traditional k-means is well-known due to its easiness, straightforward and flexibility to sparse data. Even though traditional k-means clustering algorithm has its large number of advantages, it has certain disadvantages also. The final result depends on the depends on the initial centroids. The improved algorithm initially allots datasets to its nearby centroid and then calculates distance with other centroids. Further, compare the distance between the two data objects and then the new distance is smaller than it is allocated to same cluster. This procedure will reduce the time and improve the efficiency of the traditional k-means clustering algorithm. The proposed method uses two functions. Initially the distance () function that is used to calculate the distance between data object and its nearest cluster head. Next, the distance-new () function can be used to calculate distance between data objects and other remaining clusters. The experimental results demonstrate that the proposed k-means clustering algorithm is much faster and efficient than the traditional k-means clustering algorithm.

An improved k-means clustering algorithm to deal with the problem of outlier detection of traditional k-means clustering algorithm. The enhanced algorithm makes use of noise data filter to deal with this problem. Outliers can be detected and removed by using density based outlier detection method. The purpose of this method is that the outliers may not be engaged in competition of initial cluster centers. The Factors used to test or clustering time and clustering accuracy. The drawback of the enhanced k-mean clustering algorithm is that while dealing with large scale data sets, it takes more time to produce the results.

Md.SohrabMhmudet. al [4] proposed an algorithm uses heuristic method to calculate initial k centroids. The proposed algorithm yields accurate clusters in lesser computational time. The proposed algorithm initially calculates the average score of each data objects that has multiple attributes and weight factor. Next, the Merge sort is applied to arrange the output that was generated in first phase. The data points are then divided into k cluster. Finally, the nearest possible data point of the mean is taken as initial centroid. Although the proposed algorithm still deals with the problem of assigning number of desired k-cluster as input.

III. OVERVIEW OF VARIOUS CLUSTERING ALGORITHMS

Clustering can be considered the most important unsupervised learning methodology [5]; so, as every other problem of this kind, it deals with finding a pattern in a collection of unlabeled data items. Clustering is a division of data into groups of similar objects [6]. In this part a deep analysis of various clustering techniques is presented. The result of each technique is shown in following sections.

3.1 K-means Clustering

K-means clustering is a portioning method **K-means clustering** is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean [7].

The k-means algorithm has the following characteristics:

- There are always K clusters.
- There is always at least one data item in each cluster.
- The clusters are not in hierarchical and they do not overlap with one another.
- Every member in a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

Advantages:

- It is faster if we keep k smalls.
- It creates tight clusters than hierarchical clustering, specifically if the clusters are globular.

Disadvantages:

- Difficult to find out K-Value.
- With global cluster, it didn't work well.
- Different initial centroid selection can show the different cluster results.
- If the cluster size and density are different then it will not work well .

3.2 Hierarchical Clustering

Hierarchical clustering algorithm groups' data objects to form a tree shaped structure. It can be divided into agglomerative hierarchical clustering and divisive hierarchical clustering [8][9].

3.2.1 Agglomerative hierarchical Clustering:

Agglomerative clustering is a bottom-up clustering process. At the beginning, every input object forms its own cluster. In each subsequent step, the two 'closest' clusters will be merged until only one cluster remains [10].

3.2.2 Divisive hierarchical Clustering:

This is a top down approach and all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy [11].

Advantages:

- Algorithm can never undo what was done previously.
- Time complexity of at least O ($n^2 \log n$) is required, when 'n' is the number of data points.
- Objective function is not minimized directly.
- Sometimes it is difficult to find the exact number of clusters by using the dendogram.

3.3 DB Scan Clustering Algorithm:

DBSCAN is a well-known clustering algorithm, which is easy to implement. Basically, a point q is directly density-reachable from a point p if it is not farther away than a given distance and if p is surrounded by sufficiently many points such that one may consider p and q are part of a cluster [6]. For practical observations, the time complexity is mostly governed by the number of get Neighbors queries. The overall complexity of DBSCAN is O (n log n).

Advantages:

- Not require number of cluster specification in advance.
- While clustering we can find out the noise data.
- It can find out the arbitrarily size and arbitrarily shaped clusters.

Disadvantages:

- The complexity of algorithm is still very high.
- It cannot cluster data items well with large differences in densities.
- It can only result in a good clustering as good as its distance measure is in the function get Neighbors.
- It works well in case of low dimensional data.

3.4 OPTICS Clustering Algorithm:

Ordering points to Identify the Clustering Structure (OPTICS) is an algorithm for predicting density-based clusters in data. Its basic thing is similar to DBSCAN, but it focuses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of different density [12]. The points of the database are sequentially ordered such that points which are spatially closest become neighbors in the ordering. Further, a special distance is saved for every point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster [4]. This can be represented using dendogram.

3.5 EM Clustering Algorithm:

The EM algorithm is used to find the maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Primarily these models contain many latent variables accompanied with unknown parameters and predictable data observations. That is, at times the values are missing with the data; alternatively, the model can be created more concisely by presuming the presence of additional unobserved data points [12]. The EM algorithm proceeds from the observation that the

following is a way to solve these two sets of equations numerically [13]. It simply picks arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a best estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points.

Advantages:

- Gives extremely useful result for the real world data set.

Disadvantages:

- Algorithm is highly complex in nature.

IV. COMPARATIVE STUDY

There are various clustering algorithms had been developed. A clear comparative analysis of various clustering algorithms using Iris dataset in WEKA Tool is shown in the Number of iterations, Time Complexity and Execution time for various clustering algorithms.

TABLE I: COMPARITIVE ANALYSIS OF VARIOUS CLUSTERING ALGORITHMS BASED ON EXECUTION SPEED AND TIME COMPLEXITY

S.no	Algorithm	Cluster Instances	Execution Time	Time Complexity	Cluster Shape
1	K-means clustering	0:100(67%)	0.14 sec	O(nkt)	Spherical
2	Hierarchical Clustering	0:50(33%) 1:100(67%)	0.25 sec	$O(n^{2)}$	Tree
3	DBScan Clustering	0:50(33%) 1:50(33%) 2:50(33%)	0.31 sec	O(nlogn)	Arbitrary
4	OPTICS Clustering	0:50(30%) 1:100(70%)	0.91 sec	O(nlogn)	Arbitrary
5	EM Clustering	0:48(32%) 1:50(33%) 2:29(19%) 3:23(15%)	6.67 sec	O(nlogn)	Arbitrary

V. CONCLUSION

The paper describes different methodologies and parameters associated with different clustering algorithms used in larger sets. Since clustering is a common method for data analysis, the paper gives an overview of different clustering algorithms used in large data sets. It also describes about the general working principles, and the methodologies followed on these approaches and the parameters which used in these algorithms with large data sets.

The future work planned is to enhance the efficiency of K-means clustering algorithm on the basis of time and accuracy.

References

[1] Pallavipurohit "A new Efficient Approach towards k-means Clustering Algorithm", International journal of Computer Applications, Vol 65-no. 11, March 2013.

[2] Wang Shunye, "An Improved K-means Clustering Algorithm Based on Dissimilarity" 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22,2013, Schenyang, China IEEE.

[3] FAHIM, SALEM A.M, TORKEY F.A, RAMADAN M.A ,"An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, ISSN 1009-3095 (Print): ISSN 1862-1775.

[4] Md. Sohrab Mahmud, Md. MostafizerRahman, and Md.NasimAkhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weughted average" 2012 7th International Conference on Electrical and Computer Engineering 20-22 December ,2012, Dhaka, Bangladesh, 2012 IEEE.

[5] O.A.Abbas, "Comparison between Data Clustering Algorithms", International Journal of Info. Tech., vol. 5, pp.320-325, jul.2008.

[6] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on:4/18/2010.

[7] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491.
[8] NehaSoni , AmitGanatra, "Categorization of Several Clustering Algorithms from different Perspective:Review ", International Journal of Advanced Research in Computer Science and Software Engibeering, vol. 2,no. 8,pp.63-68, Aug.2012.
[9] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from:http://www.improvedoutcomes.com/docs/WebsiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm [Accessed 22/02/2013].
[10] Sanjoy Dasgupta , "Performance guarentees for hierarchical Clustering" ,Department of Computer Science and Engineering University of Califirnia, San Diego.
[11] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta," A Comparitive Study of Various Clustering Algorithms in Data Mininig," International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1348, 2012.
[12] Glenn Fung, "A Comprehensive Overview of Basic Clustering Algorithms", June 22, 2001.
[13] Juntao Wang &Xiaolong Su , " An Improved K-Means Clustering algorithm",2011 IEEE.

