IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. III (Sep. - Oct. 2016), PP 23-28
www.iosrjournals.org

Various Techniques of Clustering: A Review

Rubina[1], Prince Verma[2]
Department of CSE, CTIEMT Shahpur, Jalandhar

Abstract: *Data mining is a method that is used to select the information from large datasets and it performs the principal task of data analysis. The Clustering is a technique that consist groups of data and elements into disjoined clusters of data. The same cluster data are related to similar cluster and different cluster data belong to different cluster. Clustering can be done different methods like partition, Hierarchical, density based etc. In this paper the problem with k-mean i.e. initialization is discussed and the possible solutions have been elaborated. The spherical k-means is the approach currently used for clustering which have provided best results till date.*

Keywords: *Data mining, clustering, K-mean, Spherical K-mean, Methods of initializing, Spherical initialized.*

I. Introduction

Data mining is a tool selecting the information from large datasets because it is very difficult to get important information and provide that information within time limit. In data mining clustering is cluster analysis of data performing principal task.[3] The term data mining is appropriately named as "Knowledge mining from data" or "Knowledge mining". Data mining is the study of exploration and analysis, by automated or systematic means of large quantities of data in order to invent patterns and rules in meaningful form.

1.1. Data mining elements

Data mining contain five major elements

1. Extract, transform, and load transaction data to the data warehouse system.
2. Data manage and Store in a multidimensional database system.
3. It also Provide data access to business analysis and Analyzing the data by application software.
4. Present the data in a useful meaning, such as a graph flow chart or table.
5. Data mining functionalities are used to specify the patterns in data mining tasks. Data mining tasks can be defined in two parts -descriptive and predictive. Descriptive mining tasks characterize the data in general properties of the database.

1.2.Data mining process

Data mining is a information gathering process which includes

1. State the problem for example:-classification or numeric calculation.
2. Assemble the data relevant that is also related with stated problem.
3. Perform preprocessing on data and represent the data in the form of labels.
4. Study a model and predictor.
5. Assess the model.
6. Well adjust the model as requirement.[6]

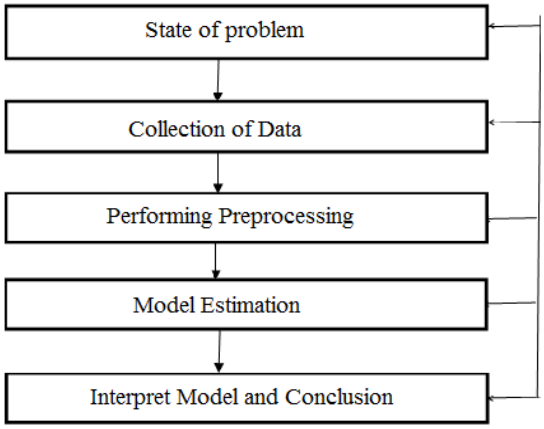

Figure 1.1 Data mining process

II. Clustering

Clustering is a data mining process for grouping and collection of data objects into disjointed clusters of data so that the same cluster data have similar properties but dissimilar data belongs to different cluster. Clustering is collection of object of data that are similar to object in similar cluster and dissimilar to the objects in other clusters. Cluster analysis is a tool that is used to observe the cluster characteristics and to focus on a particular cluster of data for further analysis. Clustering is an unsupervised learning technique which does not rely on redefined classes of cluster data. In clustering we measure the dissimilarity between the objects by measuring the distance between pair objects of data [4]. The main advantage of clustering is that interesting arrangement and found the structures directly from very large data sets with little of the background knowledge .[2] The cluster results are subjective and implementation dependent. The quality of a clustering is depends on:

- The similarity measure used by the method and its implementation.
- Its ability to invent some or all of the hidden patterns.
- K-means clustering and quality measures.

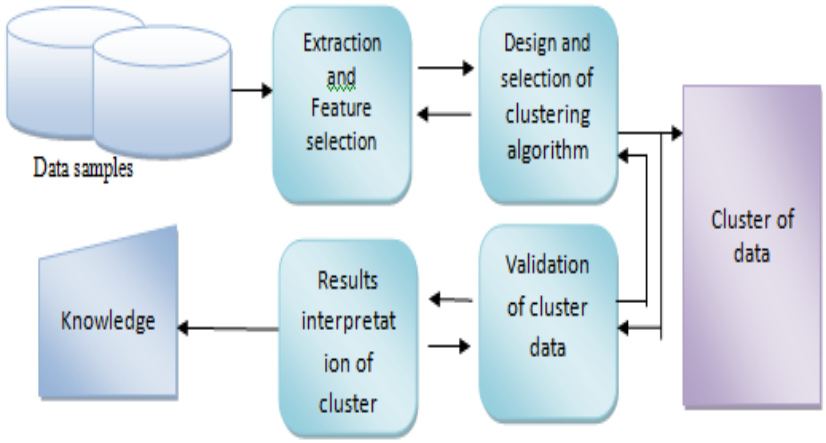

Figure2.1: Clustering procedure steps[2]

Many clustering algorithms are used for clustering the data. The major primary clustering methods can be defined into following categories.

1.2. Partitioning Methods:

The general criteria of partioning, is combining the high similarity of the samples inside clusters with high dissimilarity between separate clusters. Most of the partitioning process is distance based process. Given k, the number of partitions to create, it takes an initial partitioning and then uses an technique of iterative relocation that attempts to improve the partitioning by transferring objects from one group to another. In a good partitioning, the objects close to each other are in one cluster while object for away one in different clusters. The majority applications implement popular heuristic methods such as greedy approaches i.e k-means or k-medoids algorithms which gradually progress the clustering quality and approach a local optimum. These clustering methods are generally used to develop spherical size clusters and used for small to medium size datasets [7].

Figure 2.2: Partition based Clustering[9]

1.3. Hierarchical Methods:

In this method, Hierarchical breakdown of the given set of data objects is created. It can be defined into two approaches agglomerative and divisive. Agglomerative approach is the bottom up approach. This approach starts with each object forming a separate group. It combines groups close to one another until all the groups are merged into one group. Divisive approach is top down approach. cluster is divide into smaller clusters until each object is in one cluster [7,9].

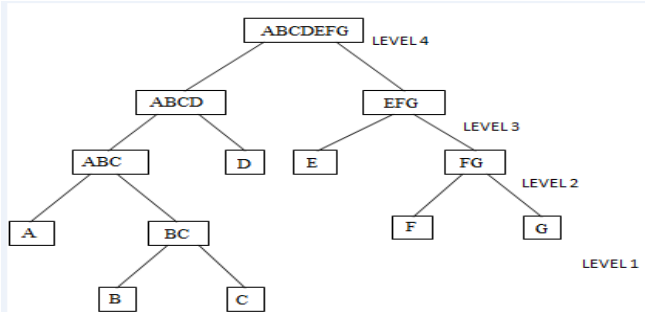

Figure 2.3: Hierarchal Clustering

1.4. Density Based Methods:

Generally in partitioning methods, cluster objects are based on distance between objects. While Spherical shaped clusters can be discovered by density based method and come across difficulty in inventing Clusters of random and arbitrary shapes. In Arbitrary shapes new methods are used called as density-based methods which are based on the notion of density. In these methods the clusters continue to produce as long as the density in the neighborhood crosses specified threshold. DBSCAN is an example of density based clustering method.[7,9]

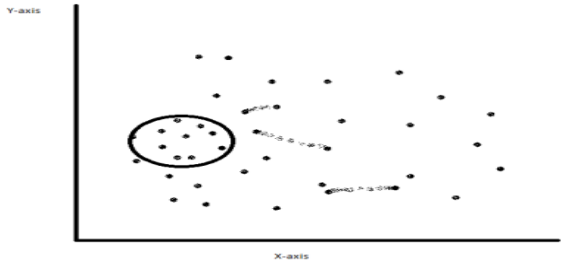

Figure 2.4: Density based clustering

III. K-Means Algorithm

K-means algorithm is a partitioning based clustering algorithm that is based on the concept of the data items with similar object of data properties in one cluster.. It is an unsupervised learning process useful for solving the clustering problem. It is the most popular and the simple algorithm for clustering It contains very simple and easy procedure to cluster a given data set [8]. It works well with small data set and takes large execution time. In there are also create the cluster in simple circular form.

The Properties of K-Means Algorithm are:-

1. There are always K clusters.
2. There is at least one item in each cluster always.
3. The clusters are non-hierarchical method and they do not overlapping.
4. Every member of a cluster is closer to the cluster object than any other cluster because closeness of cluster does not always involve the 'center' of clusters object. [7]

1.5. K-means Algorithm:

The K-means algorithm is one of partitioning algorithm, in which each clusters centre represented by the mean value of all cluster objects .

Input:

K: number of clusters objects,

D: data set contain n objects.

Output:

Number. Of clusters "k" should be generated.

Method:

1. Select any k data items randomly as cluster centroids.
2. Assign cluster, with closest centroid, to each data item.
3. Calculate the new centroid point in each cluster, by arithmetic mean of the data items belonging to that cluster.
4. If any of the data item change its centroid then go to step no. 2, else go to step no. 5
5. Output the final clusters

1.6. Drawback of K-mean Algorithm:

1. Sensitive to the selection of initial cluster center.
2. In there are no rule for decision of value of k and sensitive to initial value , for different initial value in K-mean also contain different result.
3. This algorithm is easy to be effected by unusual points.
4. It also consist dead unit problem.

IV. Spherical K-Means

Spherical K-mean clustering the text documents clustering is fundamental task and approach in modern data analysis that require the approaches which perform well both in the terms of computational efficiency or quality solution. Spherical K- mean clustering approach addresses both issues, employing cosine dissimilarities to performing prototype based partitioning of the term weight representations of the text documents cluster.
In the text clustering there are:-

1. One can introduce measure the suitable dissimilarity in between the perform clustering or texts.
2. On the other hand second, one can perform model based clustering with using probabilistic models for the text generation, like as topic models for uncovering the latent semantic document. This structure of document collection is also based on hierarchical Bayesian analysis of texts.[11]
3. All problems in the K-mean clustering are overcome by spherical K-mean clustering. It works with the large data set and gives better performance as compare to K-mean. It contain less time when records are large or small. It is used when we have max number. of members in cluster. In spherical K-mean the cluster can be made broad and we can increase the members of clusters that can change the shape of cluster which is not possible in K-mean. In spherical K-mean we can reduce the number of cluster and increase the members of clustering files.
4. In spherical we can add the max members in the cluster .When max members are added then there are cluster elements that can occur on the another cluster due to which some points are invisible. We have two solutions to avoid this problem. Visibility of cluster members is clear not points are placed to the point. In K-mean visibility of cluster is not clear.

Methods using to remove the invisibility of the cluster are following

1. To increase the size of cluster and change the shape
2. To convert the 2D to 3D view of every member of cluster and also clear its position.

V. Initialization

K means algorithm use partition based algorithm for data clustering. There are many algorithms proposed for data clustering using with K-Means algorithm due to its efficiency & simplicity but it consist of some drawbacks like as initialization of cluster centers. In initialization a new method is proposed to address the initial cluster centers problem that occur in K-Means algorithm [10].This is based on binary search technique. In it, the initial cluster centers obtained by using the binary search property. The performance of the new algorithm is tested by using two benchmark dataset. This data set is downloaded from the UCI machine learning repository. The new technique is used on the Minkowaski weighted K-Means (MWKM) algorithm that prove its significance and effectiveness.[15]

1.7. Random method Initialization:

A random method is used to generate the cluster center in the clustering, i.e., rand ().The Rand function is used to find the similarity in between the two clustering. The Random Partition method first randomly assigns a cluster to each data and proceeds cluster to the update step, thus calculate the initial mean of the centroid of cluster's randomly assigned points. While Random Partition places all of them close to the center of the data set.

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Where TP = number of true positives condition, TN= number of true negatives condition
FP= number of false positives, FN = number of false negatives.

1.8. Hartigan method Initialization

This method is used to generate the cluster center using the following equation:
1+(k-1)*[N/K]
Where, k=1, 2, 3 …….. K
N = N is the numbers of instance and that must be sorted according to center of gravity. [15]

1.9. Ward method Initialization

Generate the initial center based on the ward criteria:

$$dist(a,b) = \frac{n_a n_b}{n_a + n_b} \|x_a - x_b\|^2$$

where, n_a = number of points in cluster a, n_b = number of points in cluster b

x_a =cluster points of cluster a, x_b = cluster points of cluster b [15]

VI. Spherical Initialized

In spherical initializing is used to calculate input space with well distributed seeds. It is a new technique for calculating vectors measurement. it also used the measure for cluster's compactness. The proposed initialization is compared with the classical K-means – where initial seeds are specify or define randomly or arbitrarily. It is based on three measures: intra cluster similarity measure, cluster compactness and time converge. The proposed algorithm (called initialized K-means) better perform the classical (random) K-means. Initialized K-means clustering faster than the random K-means clustering technique for large number of clusters as it need of the time to calculate the initial center.

The spherical K-means is a variant of the K-means that uses the cosine similarity. In there the input space is also subdivided in the equivalent intervals. This subdivision contains Anderburg's observation that is distributed as possible of a good initial means. Clustering algorithms also use different metrics that are related to the quality metrics. The spherical K-means' clusters estimate technique that is uses a maximization objective function. This function measures the intra-cosine similarity for each cluster. In spherical initialize algorithm introduced a new assessment metric that measure the simulates cluster's compactness.

Process of initializing

1. Set the points
2. X={Xi----------Xd}
3. Create two hypothetical vectors
4. hv_{j1}=min(xj_i),1<=i<=d (contains minimum weights)
5. hv_{j2}=max(xj_i),1<=i<= (contain maximum weights)
6. Create vectors for means M={m_1……..m_k} in space where
7. mij =h vi_1+ j*(hv_{i2}-hvi_1)/k+1 , 1<=j>=k
8. Find C^o= {C_1…….C_k} (by applying cosine normalization to the means
9. M= {m_1….m_k}
10. i.e. $c_j=m_j/|m_j|$
    let centroid are initialize
11. t=0
12. Assign each vector $x_i \in x$ to c_j (with maximum cosine similarity)
13. J=arg max($x_i^t c_l$), $c_l \in c^{(t)}$, l≤i≤d
14. $x_{i \in \pi_j}$
15. Calculate the new means $c^{(t)}$ for new means$c^{(t)}$ for each cluster
16. $$S_j = \sum_{x_i \in \pi_j} x_i \text{ , } 1 \le j \le k$$
17. $$c_j = \frac{S_j}{|S_j|}$$
18. If $c^{(t+1)} = c^{(t)}$ (clusters are stable)
19. Else repeat 5 & 6 point

VII. Conclusion

It has been concluded that clustering is the efficient technique to analyze complex data. To analyze data in the efficient manner various techniques of clustering has been proposed so far. K-mean clustering is the best technique of clustering in terms of execution time and accuracy. In the recent times, various improvements has been proposed in k-mean clustering to improve its performance. In this paper, various improvements of k-mean clustering has been discussed and recent improvement in spherical clustering are elaborated.

References

[1] Fillippone, M. Camastra ,F. Masulli, F. Rovetta S, "A survey of kernel and spectral methods for clustering", Pattern Recognition, vol 41, pp. 176–190, (2008)

[2] Lei Gu1, "A Locality Sensitive K-Means Clustering Method Based on Genetic Algorithms" Springer-Verlag Berlin Heidelberg, Part II,vol 7929, pp. 114–119, (2013).

[3] Anwiti Jain Anand Rajavat Rupali Bhartiya," Design Analysis and Implementation of Modified K-Mean Algorithm for Large Data-set to Increase Scalability and Efficiency", Fourth International Conference on Computational Intelligence and Communication Networks, IEEE , pp. 627-631, (2012).

[4] Jyoti Yadav, Monika Sharma," A Review of K-mean Algorithm" International Journal of Engineering Trends and Technology" (IJETT) ,Vol 4, pp. 2972-2976, (2013).

[5] Nikita Jain1,Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER", International Journal of Research in Engineering and Technology(IJRET), Vol 02 , pp. 116-119, (2013)

[6] Sandeep Kaur, Sheetal Kalra," Comparison of Various Enhancements in K-means Clustering Algorithm" International Conference on Sciences, Engineering & Technical Innovation

[7] L.V. Bijuraj, "Clustering and its Applications" Proceedings of National Conference on New Horizons in IT – NCNHIT, pp.169-172, (2013).

[8] Purvashi Mahajan, Abhishek Sharma,"Role of K-Means Algorithm in Disease Prediction" International Journal Of Engineering And Computer Science (ISSN) 2319-7242, Vol 5, pp. 16216-16217, (2016).

[9] Saroj Tripti Chaudhary, " Study on Various Clustering Techniques", International Journal of Computer Science and Information Technologies(IJCSIT) , Vol. 6 (3) , pp. 3031-3033 , (2015).

[10] Rehab Duwairi, Mohammed Abu-Rahmeh, " A novel approach for initializing the spherical K-means clustering algorithm" vol. 54, pp. 49–63, Elsevier (2015).

[11] Kurt Hornik Ingo Feinerer, "Spherical k-Means Clustering" Journal of Statistical Software September, Volume 50, (2012) .

[12] Ting Su, Jennifer Dy" A Deterministic Method for Initializing K-means Clustering "16th international conference, IEEE, pp. 784-786, (2004)

[13] P. S. Bradley, Usama M. Fayyad" Refining Initial Points for K-Means Clustering"

[14] Bapusaheb B. Bhusare1, S. M. Bansode2, " Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm" , International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) ,Vol. 3, pp.1317-1322, (2014).

[15] Yugal Kumar and G. Sahoo, " A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm " International Journal of Advanced Science and Technology, (IJAST), Vol.62, pp.43-54, (2014).

[16] Raed T. Aldahdooh, Wesam Ashour," DIMK-means Distance-based Initialization Method for K-means Clustering Algorithml", Intelligent Systems and Applications" (MECS), Third International Joint Conference on Artificial Intelligence,vol.02,pp.41-51,(2013).