ISSN No:-2456-2165

A Review of Clustering, its Types and Techniques

M. Jayaprabha
Asst. Professor
Department of Computer Applications
T.John College, Bangalore

Dr.P. Felcy Judith
Associate. Professor
Department of Computer Applications
T.John College, Bangalore

Abstract:- Data Mining is a technique for organizing, analyzing and making decisions with raw data. Statistical data analysis plays a vital role in understanding and decision making. However analysis can be made using a technique called Clustering in data mining. Clustering is grouping or segregating a large data set into small groups for analysis. Clustering is used in various fields like image recognition, database management system, data analysis, augmented reality etc. This paper gives a review of the steps involved in converting a raw data into a knowledgeable data set, types of clustering and different stages involved in clustering. Also this paper involves a complete study of clustering with all the techniques and comparisons involved between different techniques. This study will be helpful for people to select a specific method and apply it for suitable environment.

I. INTRODUCTION

Cluster Analysis is a technique to segregate a set of related objects that are more similar to each other than other objects. It is one important tool for data analysis to place objects into similar groups. Clustering is a fundamental operation in data mining for analyzing the data for a specific purpose. The data can be viewed in various perspectives and decisions can be made accordingly. Clustering is used in many research communities to group related or similar objects.

Identify the problem → select the data → remove the unwanted data → segregate the data into clusters → identify the closest data related to the problem → Process data (extraction) →Analyze the data (Data mining)→ display data in graph format.

Data mining involves detecting data inconsistency, association rule learning that is understanding the relationship between different variables, classification, regression that is showing the relation between the mean variable and other variables, summarization that gives a detailed description of the data and clustering. In this paper, clustering is explained as an integral process of data mining. The procedure followed is: the data is mined using two learning approaches i.e. supervised learning and Unsupervised clustering.

In supervised learning technique,a set of related data sample is taken and a new function is found which can be matched with another new data set. This method is accurate and fast when compared to unsupervised learning because it involves finding a hidden structure or feature of the data first and then making conclusions using a relatively unrelated data set. Some examples of supervised learning are neural networks, multilayer perceptron, decision trees etc. For unsupervised learning it is K-means, hierarchical clustering, self-organizing maps, distances and normalization. Clustering can be considered the most important unsupervised learning problem as it is the basic step in segregating and analyzing the data set. So this kind of problem helps in finding of this kind deals with finding a structure in a collection of unlabeled data. In other words it is an undirected data mining process.

Cluster Name	Description	Diagram
Hard Cluster	Each data point may either belongs to a cluster completely or it may not.	
Soft Cluster	Is a probability that the data point may coincide with the other clusters assigned	
Well separated cluster	In a cluster, every one data is closer to every other data in the cluster	
Center based cluster	Is the object in a cluster that is close to the center of its cluster than the center of another cluster.	
Contiguous cluster	The object in the cluster should be closer to one or more points in the same cluster than to the points in other cluster.	
Density based cluster	Clusters are placed in high density. There may some points outside the high density area which is the noise.	
Property or conceptual cluster	Cluster that share the common property	

Table 1. Different types of clusters

II. STAGES IN CLUSTERING PROCESS

Select data (loop)→ Find patterns (loop)→ extract the patterns (loop)→ group/ arrange them based on requirement (loop)

Analyze at every step if the cluster is based on user's requirement for decision making.

In pattern representation and recognition, different types of patterns and scale of their features available in the clustering algorithms are represented. In feature selection, first identification of the effective subset of the cluster is done using extraction. Transformation of the input features are done to produce the original feature of the cluster. Then the inter pattern similarity is measured by the distance between the patterns using different distance measures. Anderberg, Jain and Dubes and Diday and Simon uses different types of distance measures. Euclidean measurement is a famous method used to bring out the dissimilarity between two or more patterns. The grouping of the elements can be done in two ways. 1. Hard way which is partitioning data into groups. 2. Fuzzy where each pattern has some amount of similariy or variable degree of membership in each output cluster. However, the selection of grouping method depends clustering to be performed. Now data is represented in a well-mannered and readable form. The loop in every step is used to make refinement of data in every step for consistency and improvement.

III. DIFFERENT TECHNIQUES OF CLUSTERING

Clustering techniques can be broadly classified into two.

1. Hierarchical clustering 2. Partitioning clustering.

The broad classification as given by Jain and Dubes is Hierarchical and partition clustering. Hierarchical Clustering is a data analysis tool to build a binary tree by merging successive closely related points. It helps in creating clusters that have a predetermined values from top to bottom. For example, all files and folders in the computer are arranged in hierarchical fashion. The pairs which are closest to each other are merged into single cluster. Repeat the steps until all items are clustered. Hierarchical clustering can be done in two ways. 1. Agglomerative (Bottom upapproach) 2. Divisive (Top down approach). In Agglomerative approach the hierarchical clustering starts with the closest pair of clusters satisfying some similarities. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. This method is also called as agglomerative clustering.During agglomeration,the closest cluster are merged into a single cluster. This process continues until all data are merged into a single cluster. However there is one disadvantage that is there is no provision made for a relocation of objects that may have been incorrectly grouped.
Process: Assign an object to a separate cluster → evaluate the distance between the clusters → construct a distance matrix → analyze the points in matrix having short distance→now merge the points → repeat the process until you arrive at a single point.

The Divisive approach works like the agglomerative approach but in the reverse order. This one starts with a single cluster and later splits into many related clusters until single object remain.

IV. PARTITIONING CLUSTERING METHODS

There are few methods of clustering under Partitioning method like the K-means, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering LARGE Applications) and the Probabilistic Clustering. The name itself tells that the data is divided into a number of subsets. These methods are used for large data sets and doing complex computations is not an easy task. To overcome this problem rank values are assigned using statistical method to the data in the cluster. K-means algorithm uses this method for finding mutually exclusive clusters by assigning rank values to them. This method is efficient for processing large data and always terminates with optimum results.

V. STEPS IN K-MEANS ALGORITHM

Data is divided into many clusters → Clusters have same rank values are grouped again → calculate the distance between each data point in the cluster → if the data points are not closest to each other then move it the closest cluster → repeat the steps until all data points are fixed in some cluster → the process ends.

VI. FUZZY C-MEANS CLUSTERING METHODS

All the above mentioned methods are arranged in clusters for the given data and the elements in the cluster are not very absolute in real world problems as they may have some characteristic that belong to other cluster. This limitation can be overcome using Fuzzy logic theory. The fuzzy logic theory considers this uncertainty of data samples where the data has more than one same characteristics or relationships of different cluster reflecting the real world situation. Zadeh has concluded that fuzzy clustering has a feature which is supported by a membership function. One of the best fuzzy methods is Fuzzy C-means (FCM) algorithm which is depicted by Bezdek.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This algorithm works by assigning membership values to each data corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. It also works on optimization of specific cost function and when the

clusters are identical. It has many advantages in real world situations like, agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis as it is an unsupervised method for data analysis and construction of models. It also shows relationship between patterns of different clusters.

ANFIS- Adaptive Neuro Fuzzy Inference System is a technique developed by Roger Jang in 1990's to overcome the limitations of Artificial Neural Networks and Fuzzy logic systems. The core idea of ANFIS technique is used to merge both Fuzzy logic and Neural Networks approaches simultaneously for better performance. This is achieved by reducing the optimization search space and highlighting the prior findings as constraints. ANFIS is best used to keep track of old data and make new decisions and conclusions from the past. It is constructed by using both input and output variables, fuzzy rules (are a set of if-then rules that define how the output must be member or a specific value in the membership set of inputs), membership functions (defines the fuzzy data set) and inference methods. Steps involved in the implementation of ANFIS is:

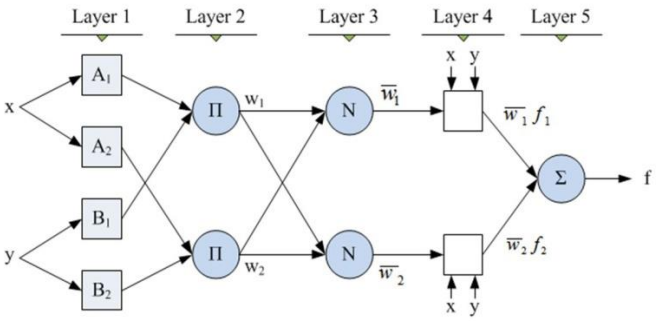
1. Initialize the problem statement → 2. Give parameter or constraints for learning → 3. Start the learning process → 4. if successful – end (validate the data) else – go to step 2.

The initialization of fuzzy system takes place with a set of previous data (as reference data commands). After a number of iterations, there may be errors (or tolerance) with a given set of inputs. Now AFNIS commands are given and wait for the results or tolerance level. The result obtained is validated with the independent data.

Fuzzy Inference System (FIS) is a collection of logical statements that describe how the FIS should make a decision regarding classifying an input and controlling an output. The rule may be *if* (input1 is membership function1) and/or (input2 is membership function2) and/or then ($output_n$ is output membership $function_n$). E.g.if temperature is high and humidity is high then room is hot.

VII. STEPS INVOLVED TO COMPUTE THE OUTPUT USING FIS

Take input →Frame a set of Fuzzy rules → select the inputs using membership function → apply fuzzy rules on the input and establish the rule strength → find the consequence of the set of rules applied on the inputs → combine the consequences to get the output →finally using defuzzification process quantify the output depending on the degree of membership which is mainly used in fuzzy control system.The below diagram has been taken from [3]



First layer :- Is called fuzzification. Each node generates a membership grade.

Second layer:- Weight for every node is computed using fuzzy AND operation.

Third layer :- Values are normalized i.e. each node calculates the ratio of its weight with the sum of all the weights in the layer.

Fourth layer :- The nodes compute a parameter function using the output of third layer.

Fifth layer :- The overall output of the system is obtained.

VIII. CONCLUSION

This paper explains all the techniques and its implementation methods on clustering. Also this paper involves a complete study of clustering with a comparative study involved between different techniques. However, ANFIS and FCM methods can be used in real life situations when compared to other methods. Methods are chosen for different applications based on their tolerance level.

REFERENCES

[1]. A.K. JAIN,M.N. MURTY&P.J. FLYNN,"Data Clustering: A Review ",ACM Computer Survey, vol. 31, no. 3, pp. 264–323, 1999.
[2]. FarhatRoohi," NEURO FUZZY APPROACH TO DATA CLUSTERING: A FRAMEWORK FOR ANALYSIS", European Scientific Journal March 2013 edition vol.9, No.9 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431.
[3]. Amandeep Kaur Mann ,Navneet Kaur ,"Survey Paper on Clustering Techniques ",International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.
[4]. Anderberg, " Cluster Analysis for Applications."Academic Press, Inc., New York, NY.
[5]. DUBES, R. C.& JAIN, A. K. ,"Clustering techniques: The user's dilemma", Pattern Recogn. 8, 247–260.
[6]. DIDAY, E. AND SIMON, J. C.," Clustering analysis. In Digital Pattern Recognition"K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, 47–94.
[7]. JAIN, A. K. AND DUBES, R. C. ," Algorithms for Clustering Data",Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

[8]. Aastha Joshi, RajneetKaur ," A Review: Comparative Study of Various Clustering Techniques in Data Mining " Volume 3, Issue 3, March 2013,International Journal of Advanced Research in Computer Science and Software Engineering.
[9]. Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierar chical_Clustering_Overview.htm [Accessed 22/02/2013].
[10]. Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491.
[11]. Patnaik, Sovan Kumar, SoumyaSahoo, and Dillip Kumar Swain, "Clustering of Categorical Data by Assigning Rank through Statistical Approach," International Journal of Computer Applications 43.2: 1-3, 2012.
[12]. ZADEH, L. A.," Fuzzy sets" Inf. Control 1965, 8,338–353.
[13]. J. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Ithaca, NY: Cornell University, 1974.
[14]. BEZDEK, J. C.," Pattern Recognition With fuzzy Objective Function Algorithms" Plenum Press, New York, NY.1981.
[15]. Shing, Roger and Jhang,"AFNIS- Adaptive Network Based Fuzzy Inference System",IEEE transaction On systems,man and cybernetics, vol 23, no. 3, JUNE 1993.