ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Review Paper on Clustering and Validation Techniques

Jyoti, Neha Kaushik, Rekha

***Abstract*— Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters).**

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering, the value of k-mean is set. Clustering has been applied to serve various purposes like, to gain insight to data distribution, generate hypotheses, to observe the characteristic and find anomalies. The intension of this paper is to provide a categorization of some well known clustering algorithms. It also describes the clustering process and overview of the different clustering methods. The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Validation comparing the results of two clusters and find out the best cluster.

***Index Terms*— Data mining, clustering process, Categorization of Clustering, validation etc.**

I. INTRODUCTION

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized into partition-based algorithms,hierarchical-basedalgorithms,density-based algorithms and grid-based algorithms.Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid Density based algorithm uses the multi resolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time.In this survey paper, an analysis of clustering and validation techniques. In data mining data can be mined by passing these steps.

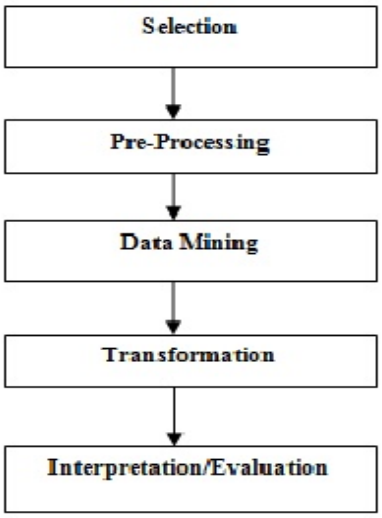

Figure 1 : Phases of Data Mining

In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised

learning.

a) Supervised Learning

In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees.

b) Unsupervised Learning

The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self organizing maps.

II. CLUSTERING PROCESS:

The overall process of cluster analysis is shown in fig.2 It involves four basic steps as explained below.

A. Feature Selection or Extraction

Feature selection is the process of identifying the most effective subset of the original features to use in clustering, whereas the feature extraction is the process of transforming one or more input features to produce new salient feature. Clustering process is highly dependent on this step. Improper selection of features increases the complexity and may result into irrelevant clusters, too.

B. *Clustering Algorithm Design or Selection*

The impossibility theorem [12] states that, "no single clustering algorithm simultaneously satisfies the three basic axioms of data clustering, i.e., scale-invariance, consistency and richness". Thus it impossible to develop a generalized framework of clustering methods for the application in the different scientific, social, medical and other fields. It is therefore very important to select the algorithm carefully by applying domain knowledge. Generally all algorithms are based on the different input parameters, like number of clusters, optimization/construction criterion, termination condition, proximity measure etc. This different parameters and criteria are also designed or selected as a prerequisite of this step.

C. Cluster Validation

As there is no universal algorithm for clustering, different clustering algorithm applied to same dataset produce different results. Even the same algorithm, with the different values of parameter produces different clusters. Therefore it becomes necessary to validate or evaluate the result produce by the clustering method. The evaluation criteria are categorized as:

1) Internal indices: The internal indices generally evaluate the clusters produces by the clustering algorithm by comparing it with the data only.

2) External indices: The external indices evaluate the clustering results by using the prior knowledge, e.g. class labels.

3) Relative indices: As the name suggest, this criteria compares the results against various other results produced by the different algorithms.

D. Results Interpretation

The last step of clustering process deals with the representation of the clusters. The ultimate goal of clustering is to provide users with meaningful insights from the original data, so that they can effectively analyze and solve the problems. This is still an untouched area of research.

III. CATEGORIZATION OF CLUSTERING METHODS

There is difference between clustering method and clustering algorithm]. A clustering method is a general strategy applied to solve a clustering problem, whereas a clustering algorithm is simply an instance of a method. As mentioned earlier no algorithm exist to satisfy all the requirements of clustering and therefore large numbers of clustering methods proposed till date, each with a particular intension like application or data types or to fulfil a specific requirement. All clustering algorithms basically can be categorized into two broad categories: partitioning and hierarchical, based on the properties of generated clusters. Different algorithms proposed may follows a good features of the different methodology and thus it is difficult to categorize them with the solid boundary. The detail categorization of the clustering algorithm is given in figure 2. Though we had tried to provide as much clarity as possible, there is still a scope of variation. The overview of each categorization is discussed below.

A. Hierarchical Methods

ISSN: 2321-9653

As the name suggest, the hierarchical methods, in general tries to decompose the dataset of n objects into a hierarchy of a groups. This hierarchical decomposition can be represented by a tree structure diagram called as a dendrogram; whose root node represents the whole dataset and each leaf node is a single object of the dataset. The clustering results can be obtained by cutting the dendrogram at different level. There are two general approaches for the hierarchical method: agglomerative (bottom-up) and divisive (top down).

Hierarchical Clustering is classified as

A. Agglomerative

B. Divisive

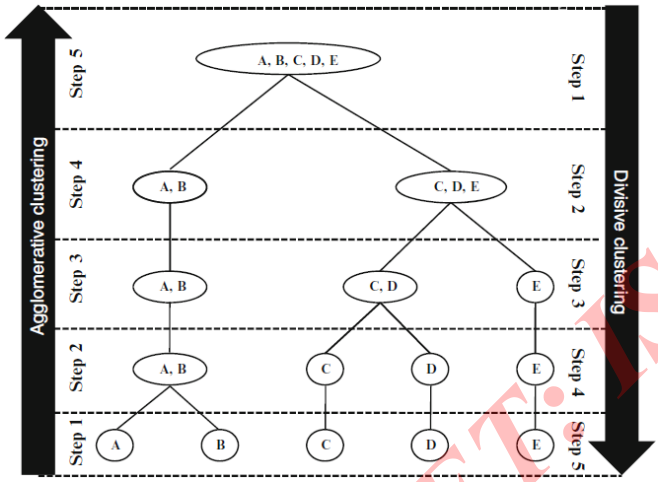

Fig 2. Agglomerative and Divisive Clustering

Agglomerative clustering

It is also known as AGNES. It is bottom-up approach. This method construct the tree of clusters i.e. nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance, center distance. The steps of this method are:

(1) Initially all the objects are clusters i.e. leaf.

(2) It recursively merges the nodes (clusters) that have the maximum similarity between them.

(3) At the end of the process all the nodes belong to the same cluster i.e. known as the root of the tree structure.

Devise clustering

It is also known as DIANA. It is top-down approach. It is introduced in Kaufmann and Rousseeuw (1990). It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own.

Advantages of hierarchical clustering [2]

(1) Embedded flexibility with regard to the level of granularity.

(2) Ease of handling any forms of similarity or distance.

(3) Applicability to any attributes type.

Disadvantages of hierarchical clustering [2]

(1) Vagueness of termination criteria.

(2) Most hierarchical algorithm does not revisit once constructed clusters with the purpose of improvement.

Validation

The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Validation comparing the results of two clusters and find out the best cluster.In validation step we use the BSS stands for Between Sum of Squared Error and WSS stands for Within Sum of Squared Error techniques.

WSS:

ISSN: 2321-9653

Cluster Cohesion: Measures how closely related are objects in a cluster.

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

WSS stands for Within Sum of Squared Error.

BSS:

Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters

$$BSS = \sum_i |C_i|(m - m_i)^2$$

– Where $|C_i|$ is the size of cluster i

BSS stands for Between Sum of Squared Error.

Validation techniques is used to find out the best cluster.

Different Aspects of Cluster Validation

Determining the clustering tendency of a set of data, i.e. distinguishing whether non-random structure actually exists in the data.

Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.

Evaluating how well the results of a cluster analysis fit the data without reference to external information.

- Use only the data.

Comparing the results of two different sets of cluster analyses to determine which is better.

Determining the 'correct' number of clusters. For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

AIM AND OBJECTIVE:

Main objective of my thesis is to delivery of document using hierarchical clustering technique.

Increase document retrieval efficiency

Compare the set of cluster and validate the cluster using BSS/WSS technique

Solve outliers problem

Ranking the document according to their relevancy

And deliver to user.

Conclusions and Directions of Future Research

For the explosion of information in the World Wide Web, this thesis proposed a new method of summarization via soft clustering algorithm. It used Google search engine to extract relevant documents, and mixed query sentence into document set which segmented from multi-documents set, then this paper created efficient hierarchical clustering to cluster all the documents. Also, there are a lot of rooms for improvement. For example, readability is an important aspect in the performance of multi-document summarization. In future work, we will consider new soft cluster algorithm to more improve the efficiency of clustering. Cluster Analysis is a process of grouping the objects, called as a cluster/s, which consists of the objects that are similar to each other in a given cluster and dissimilar to the objects in other cluster. With the application of clustering in all most every field of science and technology, large number of clustering algorithms had been proposed which satisfy certain criteria such as arbitrary shapes, high dimensional database, and domain knowledge and so on. It had been also proved that it is not possible to design a single clustering algorithm which fulfils all the requirement of clustering. Therefore, number of methods had been proposed such as partitioning, hierarchical, density based, model based and so on. Different algorithms may follow good features of one or more methods and thus it is difficult to categorize them with the solid boundary. In this paper we had tried to provide a detail categorization of the clustering algorithms from our perspective. Though it had been tried to cover as much clarity as possible, there is still a scope of variation. In this paper we had covered the detailed categorization of the different clustering methods with the representative algorithms under each. The future work planned is to perform a detailed analysis of major clustering algorithm and find out the best algorithm for document deliver to the user

ACKNOWLEDGEMENT

I would like to thanks Ms. Neha Kaushik and the Department of Computer Science & Engineering of DITM Institute, Gannur , Sonepat,India.

REFERENCE

[1] A.K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, vol. 31, pp. 264-323, Sep. 1999.

[2] O. A. Abbas, "Comparisons between Data Clustering Algorithms", *The Int. Journal of Info. Tech.* ,vol. 5, pp. 320-325, Jul. 2008.

[3] Dr. E. Chandra, V. P. Anuradha, " A Survey on Clustering Algorithms for Data in Spatial Database Management Systems", *International Journal of Computer Application*, vol. 24, pp. 19-26

[4] Oded Maimon, Lior Rokach, "DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK", Springer Science+Business Media.Inc, pp.321-352, 2005.

[5] Bharati M Ramager, "Data Mining techniques and Applications", "International Journal of Computer Science and Engineering Vol. 8", December 2009.

[6] Accessible from Sonali Agarwal, Neera Singh, Dr. G.N. Pandey, "Implementation of Data Mining and Data Warehouse in E-Governance", "International Journal of Computer Applications (IJCA) (0975-8887), Vol.9- No.4," November 2010.

[7] Calinski, T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27

[8] Ng, R. and Han, J.(1994). Efficient and Effective Clustering Methods for Spatial Data Mining. In Proceeding's of the 20th VLDB Conference, Santiago, Chile.

[9] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Turkey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.

[10] Willet, Peter "Parallel Database Processing, Text Retrieval and Cluster Analyses" Pitman Publishing, London, 1990.