Volume 2, Issue 6, June 2017 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Data Mining Clustering Techniques: - A Comparative Study

Deepika Patidar
M. Tech.(CSE) IV Semester
Lord Krishna College of Technology
Indore M.P. India

Vijay Kumar Verma
Asst. Prof CSE Dept.
Lord Krishna College of Technology
Indore M.P. India

Abstract: - : Clustering can be used to partition data set into a number of “interesting” clusters. Cluster analysis is applied to the data set and the resulting clusters are characterized by the features of the patterns that belong to these clusters. Clustering techniques are widely nowadays. Artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing are some common clustering problem. But improving accuracy and efficiency are to issue are always arises for finding a new algorithm and process for extracting knowledge for. . These issues motivated us to develop new algorithm and process for clustering problems. In this paper we presented a study of some data mining clustering techniques.

Keywords: - Cluster, Accuracy , Efficiency, Partition, Features

1. INTRODUCTION

A Cluster is a set of entities which are alike, and entities from different clusters are not alike. Clusters may be described as connected regions of a multidimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points. Clustering is based on some properties like density, variance, dimension, shape, and separation. The cluster should be a tight and compact high-density region of data points when compared to the other areas of space. The shape of the cluster is not known a priori. It is determined by the used algorithm and clustering criteria.

Clustering is unsupervised learning because it doesn’t use predefined category labels. A clustering algorithm attempts to find natural groups of components based on some similarity. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster [1, 2].

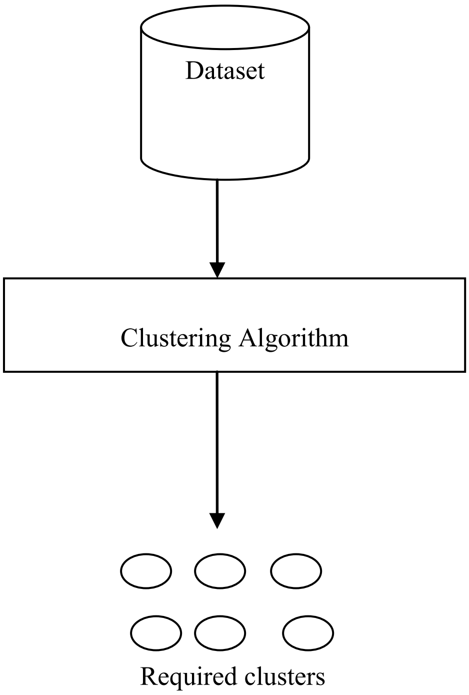


Figure 1. Clustering of raw data

II. ATTRIBUTES OF A CLUSTER

Clustering of object is a difficult task. To construct a cluster there should be some properties that have to be considered. Some of the properties are size, depth breath, etc [11].

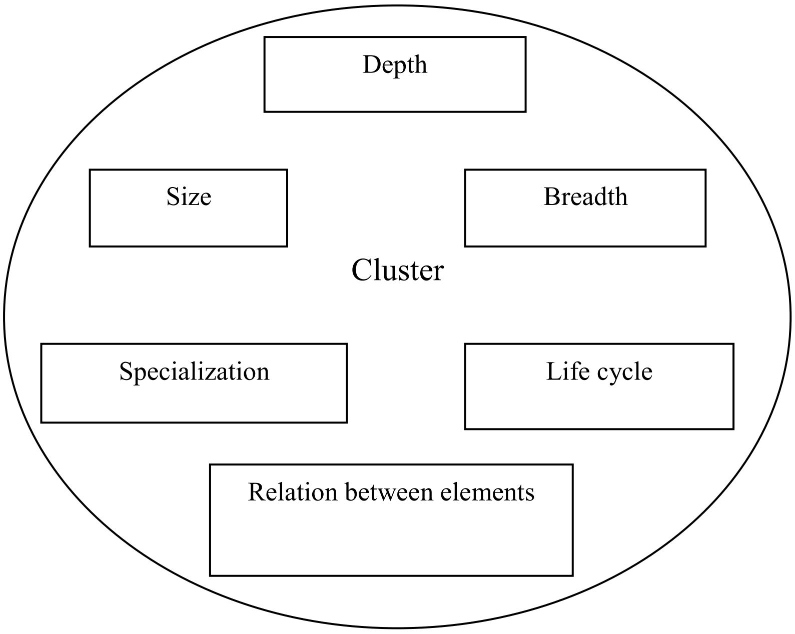
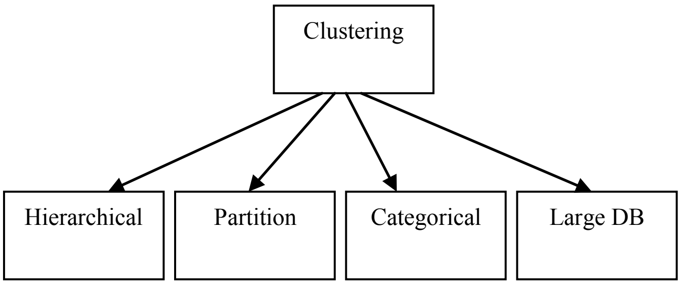
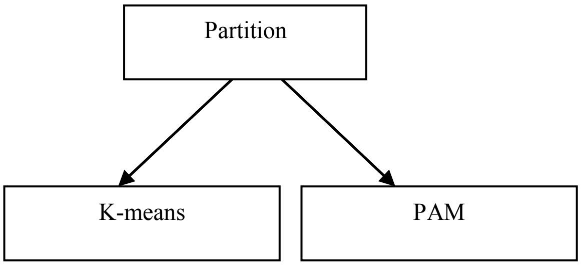


Figure 2. Properties of a cluster

III. CLUSTERING TECHNIQUES

There are several clustering methods have been developed in past year. Each of which uses a different techniques and process induction principle. Farley and Raftery suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid based methods. An alternative categorization based on the induction principle of the various clustering methods is presented in (Estivill-Castro, 2000). We discuss some of them here[12,15].





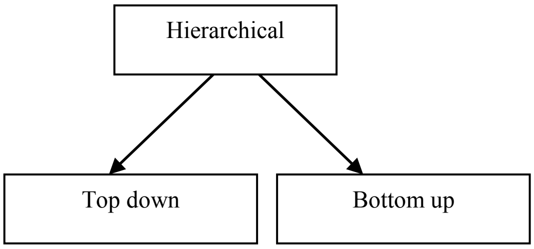


Figure 3. Categories of clustering methods

III. LITERATURE REVIEW

In 2011 K. Ranjini proposed "Performance Analysis of Hierarchical Clustering Algorithm" They explain the implementation of agglomerative and divisive clustering algorithms by using various types of data. They implements and analysis running time of the algorithms using different linkages (agglomerative) to different types of data are taken for analysis[4].

In 2012 Akshay Krishnamurthy proposed "Efficient Active Algorithms for Hierarchical Clustering". They show that a family of active hierarchical clustering algorithms has strong performance. They show that clustering can be improved by using statistical properties. We propose a general framework for active hierarchical clustering [5].

In 2013 K. Sasirekha, P. Baby proposed "Agglomerative Hierarchical Clustering Algorithm- A Review". They showed that data mining hierarchical clustering method are used to build a hierarchy of clusters. They also show that hierarchical clustering generally fall into two types: Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy [6].

In 2014 Archana Singh and Avantika Yadav proposed "Hybrid Approach of Hierarchical Clustering". They proposed a hybrid approach of clustering based on AGNES and DIANA clustering algorithms, an extension to the standard hierarchical clustering algorithm. In the proposed algorithm, they used single linkage as a similarity measure. The proposed clustering algorithm provides more consistent clustered results from various sets of cluster centroids with tremendous efficiency [7].

In 2015 Olga Tanaseichuk "An Efficient Hierarchical Clustering Algorithm for Large Datasets". They show that Hierarchical clustering is a widely adopted unsupervised learning algorithm. Standard implementations of the exact algorithm for hierarchical clustering require O(n)2 time and O(n)2 memory and thus are unsuitable for processing datasets with large object. They present a hybrid hierarchical clustering algorithm requiring less time and memory [8].

In 2016 K.Jeyalakshmi, S.Shanmugapriya "An Efficient Hierarchical Clustering Algorithms Approach Based on Various- Widths Algometric Clustering". They proposed method by initially assigning each point to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all inclusive cluster. The key parameter in agglomerative algorithms is the method used to determine the pair of clusters to be merged at each step[9]
.

In 2017 Shaoning Li , Wenjing Li proposed "A Novel Divisive Hierarchical Clustering Algorithmfor Geospatial Analysis". They proposed a new method, cell-dividing hierarchical clustering (CDHC), based on convex hull retraction. They used following steps a convex hull structure is constructed to describe the global spatial context of geospatial objects. Then, the retracting structure of each borderline is established in sequence by setting the initial parameter. The objects are split into two clusters with the borderlines. Finally, clusters are repeatedly split and the initial parameter is updated until the terminate condition is satisfied [10].

In 2012 Akhil jabbar et al. proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart disease prediction. The main advantage of genetic algorithm is the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. The proposed method helps in the best prediction of heart disease which even helps doctors in their diagnosis decisions[1].

In 2013 Akhil Jabbar et al. proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They proposed a new feature selection method using ANN for heart disease classification. For rank the attributes which contribute more towards classification of heart disease they applied different feature selection methods, and indirectly reduce the no. of diagnosis tests to be taken by a patient. The proposed method eliminates useless and distortive data[2] .

In 2014 N. S. Nithya et al . proposed "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface". They showed that earlier model based on information gain and fuzzy association rule mining algorithm for extracting both association rules and membership functions are not feasible. They used large number of distinct values. They modify gain ratio based fuzzy weighted association rule mining and improve the classifier accuracy[3] .

In 2015 S. Olalekan Akinola, O. Jephthar Oyabugbe proposed “Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study”. They proposed study was designed to determine how data mining classification algorithm perform with increase in input data sizes. They used three data mining classification algorithms Decision Tree, Multi-Layer Perceptron (MLP) Neural Network and Naïve Bayes were subjected to varying simulated data sizes. The time taken by the algorithms for trainings and accuracies of their classifications were analyzed for the different data sizes[4].

In 2015 Jaimini Majali, Rishikesh Niranjan & Vinamra Phatak proposed “Data Mining Techniques for Diagnosis and Prognosis of Cancer”. They used data mining techniques for diagnosis and prognosis of cancer. They presented a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. They used FP algorithm in Association Rule Mining to conclude the patterns frequently found in benign and malignant patients[5] .

In 2016 Nikhil N. Salvithal & R.B. Kulkarni proposed “Appraisal Management System using Data mining Classification Technique”. The proposed assorted classifier algorithms applied on Talent dataset to spot the talent set so as to judge the performance of the individual. Finally counting on accuracy one best suited classifier is chosen this method has been used to construct classification rules to predict the potential talent that for promotion or not[6].

In 2016 Tanvi Sharma & Anand Sharma proposed “Performance Analysis of Data Mining Classification Techniques on Public Health Care”. The proposed study focused on the application of various data mining classification techniques using different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analyzing the health care system. The percentage of accuracy of every applied data mining classification technique is used as a standard for performance measure. The best technique for particular data set is chosen based on highest accuracy[7].

IV. PROBLEM WITH CLUSTERING

The important problems with cluster analysis that this work have identified are as follows:
1. The identification of distance measure: For numerical attributes, distance measures can be used. But identification of measure for categorical attributes in strength association is difficult.
2. The number of clusters: Identifying the number of clusters & its proximity value is a difficult task if the number of class labels is not known in advance. A careful analysis of inter & intra cluster proximity through number of clusters is necessary to produce correct results.
3. Types of attributes in a database: The databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other types like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.
4. Merging decision in not given: Hierarchical clustering tends to make good local decisions about combining two clusters since it has the entire proximity matrix available. However, once a decision is made to merge two clusters, the hierarchical scheme does not allow for that decision to be changed. This prevents a local optimization criterion from becoming a global optimization criterion [13,14].

V. ADVATAGE AND DISATVATGE

There are several clustering algorithm. Each and every clustering approach has some advantage and disadvantage. Selection of a particular clustering method is depend on the data set other factors. We give some advantage and disadvantage of these methods.

Clustering Techniques	Benefits	Drawbacks
Partition based appaoch	clusters are mostly identical size	Initial value not known
Hierarchical Approach	Produce more accurate hierarchies	Local patterns are important
Density based Approach	Does not require number of clusters	Depends on the distance measure

CONCLUSION

We represent a study of some clustering method. We give a study over some cluttering techniques on the basis of their properties.ome method has good scalability and some has good performance. Selection of a method depends on the nature of the data object and other factors.
Hierarchical clustering tends to make good local decisions about combining two clusters since it has the entire proximity matrix available. However, once a decision is made to merge two clusters, the hierarchical scheme does not allow for that decision to be changed. This prevents a local optimization criterion from becoming a global optimization criterion.

Agglomerative algorithms is a type of hierarchical clustering it places each object in its own cluster and then it merges these atomic cluster into larger and larger clusters until all objects are in a single cluster or until termination condition holds. Most commonly used hierarchical agglomerative clustering methods are Single linkage and complete linkage. It is very difficult to decide to select a method for a given objects because each method has its own advantage and disadvantage. In our proposed work our main focus is this problem. We use Dendogram distance between two object and correlate with original distance matrix. This correlation between two matrixes gives a value between 0 and 1. The value near to give more accurate cluster.

REFERENCE

[1]Data Mining: Concepts and Techniques Jiawei Han and Micheline Kamber Simon Fraser University Note: This manuscript is based on a forthcoming book by Jiawei Han and Micheline Kamber, c 2000 (c) Morgan Kaufmann Publishers.

[2] Arun K. Pujari "Data Mining Techniques" Universities Press, 2001

[3]Margaret H. Dunham "Data Mining: Introductory and Advanced Topics" Publisher : Pearson 2002-09-01 ISBN-13 : 9780130888921

[4] K.Ranjini Performance Analysis of Hierarchical Clustering Algorithm Performance Analysis of Hierarchical Clustering Algorithm" Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011 (2011)

[5] Akshay Krishnamurthy "Efficient Active Algorithms for Hierarchical Clustering" Appearing in Proceedings of the 29 th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.

[6] K.Sasirekha, P.Baby Agglomerative Hierarchical Clustering Algorithm- A Review "International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013 1 ISSN 2250-3153

[7] Archana Singh and Avantika Yadav "Hybrid Approach of Hierarchical Clustering"World Applied Sciences Journal 32 (7): 1181-1191, 2014 ISSN 1818-4952 © IDOSI Publications, 2014

[8] Olga Tanaseichuk, Alireza Hadj "An Efficient Hierarchical Clustering Algorithm for Large Datasets" Austin J Proteomics Bioinform & Genomics - Volume 2 Issue 1 - 2015 ISSN : 2471-0423

[9] K. Jeyalakshmi, S. Shanmugapriya "An Efficient Hierarchical Clustering Algorithms Approach Based on Various-Widths Algometric Clustering" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 7, July 2016".

[10] Shaoning Li , Wenjing Li "A Novel Divisive Hierarchical Clustering Algorithm for Geospatial Analysis" ISPRS Int. J. Geo-Inf. 2017, 6, 30; doi:10.3390/ijgi6010030

[11] Parul Agarwal et al. "Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes" International Journal of Innovation, Management and Technology, Vol. 1, No. 2, June 2010 ISSN: 2010-0248

[12] Ashish Jaiswal et al. " Hierarchical Document Clustering: A Review" 2nd National Conference on Information and Communication Technology (NCICT) 2011 Proceedings published in International Journal of Computer Applications® (IJCA).

[13] Sudesh Kumar "Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 10 3161 – 3166

[14] Shraddha Shukla A Review ON K-means DATA Clustering APPROACH" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1847-1860 © International Research Publications House http://www. irphouse.com

[15] Shuhie Aggarwal et al. " Hierarchical Clustering- An Efficient Technique of Data mining for Handling Voluminous Data" International Journal of Computer Applications (0975 – 8887) Volume 129 – No.13, November2015