ISSN No.: 2455-6130
Volume 1, Issue 4, 2016, pp. 29-35

website: www.ijoetr.com

International Journal of Emerging Trends in Research

An Observed Study of Clustering in Data Mining

Saloni Chaudhary[1] , Tarun Kumar[2]

[1,2]Department of Computer Science & Engineering, Geeta Institute of Engineering and Technology, Kanipala Haryana, INDIA

Abstract

The aim of this paper is to study the various clustering concept in the field of data mining. Clustering is the necessary task in Data Mining process which is used for the purpose to make groups of cluster of the given data based on the basis of their similarity. Clustering is used in number of active research areas such as statistics, pattern recognition, machine learning, data analysis etc. We can observe that Clustering is the one of the data mining techniques in which data is divided into the groups of similar objects. This paper includes many clustering and data mining techniques and also covers the behavior of clustering concept and basic approaches.

Keywords: Clustering; Type of Clustering; Data Mining; Classification.

1. Introduction

Data mining is the concept to extract the knowledge or data from large amount of data. To extract useful information is the objective of data mining. A variety of analytic computer models have been used in data mining. The standard model variety in data mining contains decay (normal decay for prediction, logistic degeneration for classification), neural networks, and decision trees. The following definition is given: Data mining is the process of study and examination, by regular or repeated means, of bulky quantity of data in order to discover meaningful patterns and rules [1]. The beginning of information technology in various fields of human life has direct to the large volumes of data storage in different formats like records, documents, images, sound recordings, videos, scientific data, and many latest data formats. The data collected from different applications require appropriate method of extracting knowledge from big database for better decision making. Knowledge discovery in databases (KDD), often called data mining, goal at the detection of useful information from bulky cluster of data [2]. Data mining is also recognized as fundamental method where intelligent methods are applied in order to extract the data patterns. Data mining consists of five major elements:

1. Extract, change, and stack transaction data onto the data warehouse system.

2. Store and manage the data in a multidimensional database system.
3. Offer data access to business analysts and information technology professionals.
4. Evaluate the data by application software.
5. Demonstrate the data in a useful format, such as a graph or table.

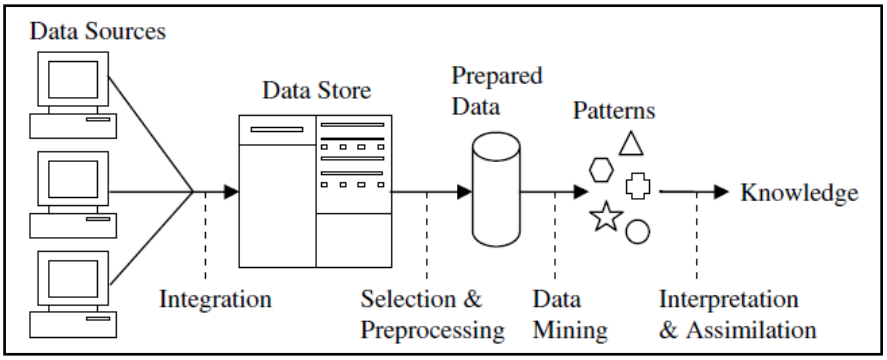

Figure 1 The Knowledge Discovery Process

2. Literature survey

Muhammad Husnain Zafar and Muhammad Ilyas [3] in their paper represents the grouping of data objects such that the objects within a cluster are related to one another and unrelated from the objects in other groups. Many of clustering algorithm is obtainable to analyze data. In their paper they intend to study and compare different clustering algorithms. These algorithms consist of K-Means, Farthest First, DBSCAN, CURE, Chameleon algorithm. Every algorithm is evaluated on the basis of their pros and cons, comparison measure, their working, functionality and time complexity.

Amit Kumar Kar, Shailesh Kumar Patel and Rajkishor Yadav [4] describes in their paper that the data mining process is to extract information from a huge data set and convert it into an different utilizable form for further use. They say clustering is extremely essential in data analysis and in different data mining applications. Clustering is a partition of data into groups of related objects. All groups, called cluster, consists of objects that are similar among themselves and dissimilar to objects of other groups. Clustering deals with finding a structure in a gathering of unlabeled data. At this time there are number of clustering algorithms are used to organize data, categorize data, for data compression and model production etc. They investigates the four major clustering algorithms namely: Partitioning process, Hierarchical process, Grid-based process and Density-based process and comparing the performance of these algorithms on the basis of appropriately class wise cluster building capability of algorithm.

Namrata S Gupta, Bijendra S.Agrawal and Rajkumar M. Chauhan [5] in their survey paper, they describes various types of clustering techniques in data mining is done. They states that data mining refers to remove useful information from huge amounts of data. It is the procedure of find out interesting knowledge from big quantity of data stored either in databases, data warehouses, or other information collections. A significant method in data analysis and data mining applications is Clustering. It break up data into groups of similar objects. Every group, called cluster, consists of objects that are similar between themselves and unrelated to objects of other groups. Data mining has two forms of nature: Predictive and other is descriptive. There are

different types of clustering algorithms such as hierarchical, partitioning, grid, density based, model based, and constraint based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centered based clustering; the value of k-mean is set. Density based clusters are distinct as area of higher density then the remaining of the data set. Grid based clustering is the greatest processing time that typically depends on the size of the grid in its place of the data. Model based clustering put forward for each cluster and find the best fit of data to the specified model. Constraint based clustering is performed by inclusion of user or application oriented constraints.

Srinivas Sivarathri and A.Govardhan [6] in their paper explores clustering algorithms in terms of computational efficiency, measure of similarity, speed and performance. An overabundance of algorithms exists for clustering to discover actionable knowledge from large data sources. Given un-labeled data objects, clustering is an unconfirmed learning to find natural groups of objects which are alike. Each cluster is a subset of objects that show signs of high similarity. Quality of clusters is far above the ground when they feature highest intra-cluster similarity and lowest inter-cluster similarity. The quality of clusters is prejudiced by the similarity calculate being employed for grouping objects. The clustering quality is measured the ability of clustering technique to uncover latent trends distributed in data. The usage of data mining technique clustering is all over the place in real time applications such as market research, discovering web access patterns, document classification, image processing, pattern recognition, earth observation, banking, insurance to name few. Clustering algorithms differ in type of data, measure of similarity, computational efficiency, and linkage methods, soft or hard clustering and so on. Make use of a clustering technique correct depends on the technical knowledge one has on various kinds clustering algorithms and suitable circumstances to apply them.

V. Sarala and Dr. V.V.Jaya Rama Krishnaiah [7] “Empirical Study Of Data Mining Techniques In Education System” in their paper describes the role of Data mining Techniques in Education system. They states that educational institutions are important parts of our society and playing a vital role for enlargement and development of nation. In earlier days the information flow in education field was relatively straightforward and the application of technology was limited. However, as we progress into a more integrated world where technology has become an essential part of the business processes, the process of transport of information has become more complicated. Today, one of the biggest challenges that educational institutions face the volatile growth of educational data and to use this data to develop the quality of managerial decisions. Data mining techniques are methodical tools that can be used to extract meaningful knowledge from large data sets. In their paper the applications of data mining in educational institution to extract useful information from the huge data sets and providing methodical tool to view and use this information for decision making processes by taking real life examples.

3. Clustering analysis

Cluster Analysis is a regular process to find comparable objects from a database. It is a fundamental operation in data mining. In this section, clustering analysis is done. Cluster analysis is an important data mining technique which is used to discover data segmentation and sample information. By clustering the data, user can get the data division, examine the character of all clusters, and make further study on particular clusters.

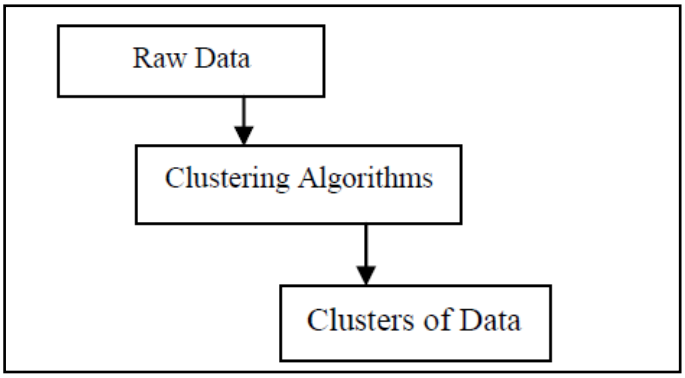

Figure 2 Stages of Clustering

In adding up, cluster analysis regularly performs as the pre processing of other data mining operations. The goal of cluster analysis is that the objects in a group be supposed to be similar to one another and dissimilar from the objects in other groups. Clustering is much enhanced when there is superior similarity within a group and improved the difference between the groups.

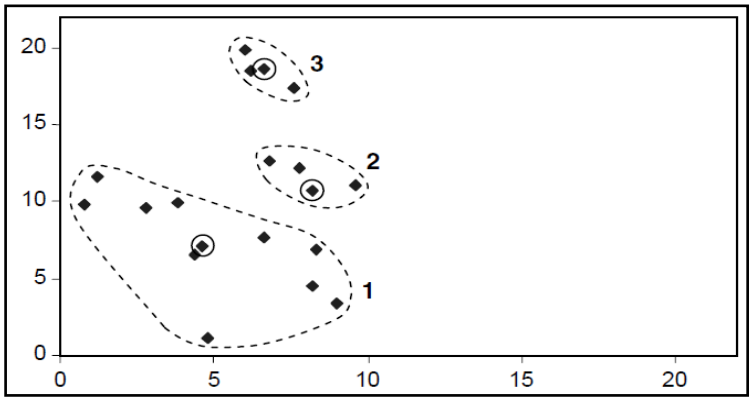

Figure 3 Stages of Clustering of Data

3.1 Classification of clustering

Clustering is the fundamental operation of Data Mining. And it is done by the number of algorithms. There is no. of algorithms used in clustering such as: Partitioning, Hierarchical, Density and Grid based algorithms.

3.1.1 Partitioned clustering

Partition-based process builds the clusters by building different partitions of the dataset. So, partition gives for each data object the cluster index pi. The user proposed the preferred number of clusters M, and some standard function is used in order to estimate the proposed partition or the solution. This examine of quality could be the average distance between clusters; for example, some well-known algorithms under this category, such as PAM, k-means and CLARA. One of the most popular and widely studied clustering methods for substance in Euclidean space is known k-means clustering. The k-means algorithm is very easy iterative method to partition a known dataset into the user-precise number of clusters, k [8]. This algorithm has been discovered by several researchers i.e. Gray and Neuhoff offers a superior historical Back ground for k-means

placed in the larger context of hill-climbing algorithms. The algorithm works on a set of d-dimensional vectors, D = {xi | i = 1. . . N}, where xi belongs to d represent the ith data point. The algorithm is in progress by picking k points in d as the initial k cluster legislature. Here's shown how k-mean algorithm works:

Algorithm:

Input: k= the number of clusters. D= a data set that contains n items.
Output: Set of k clusters.
Method: 1. Randomly select k objects from D as the initial cluster centre.
2. Repeat.
3. Shift each object to the cluster to which the object is most similar based on the mean value of the objects in the cluster.
4. Bring up to date the cluster means, i.e. analyze the mean value of the objects for each cluster.
5. **until** no change.

3.1.2 Hierarchical clustering

Hierarchical clustering process constructs a cluster hierarchy. A tree diagram frequently used to characterize the results of a cluster analysis. Hierarchical clustering process is categorized into bottom-up and top-down as shown in Figure 4. A bottom-up clustering starts with one-point clusters and recursively combine two or more most suitable clusters. In distinction, a top-down clustering starts with one cluster of all data points and recursively divide into non overlapping clusters.

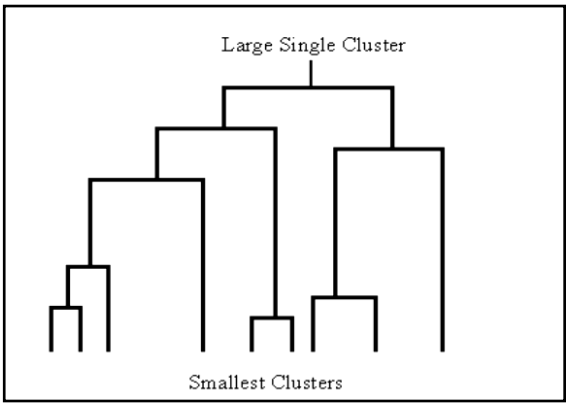

Figure 4 Hierarchy of clusters

3.1.3 Density-based and Grid-based clustering

The process of density-based methods is that for each object of a cluster the region of a given radius has to include a certain number of objects; i.e. the density in the region has to go above some threshold. The nature of a region is determined by the selection of a reserve purpose for two substances. These algorithms can capably divide noise. DBSCAN and DBCLASD are the well-known methods in the density based category. The fundamental theory of grid-based clustering algorithms is that they quantize the gap into a predetermined number of cells that form a grid structure. And then these algorithms do all the operations on the quantized gap. The foremost advantage of this process is its fast processing time, which is typically not dependent on the

number of objects, and depends only on the number of grid cells for each measurement. Well-known methods in this clustering category are STING and CLIQUE.

4. Issues in cluster analysis research

The most important concerns of cluster analysis research can be collected in the following extensive categories [9-11] such as: Characteristic of data and Characteristic of cluster, Cluster Algorithm Characteristic, Data transformation issues, Cluster solution issues, Validity issues and Variable selection issues and described the related description Table.1

Table 1 Issues in cluster analysis research

Issues	Description
Data Characteristic	• High dimensionality • Data size • Sparseness • Blare and Outlier • Nature of traits and data set • Scale • Mathematical properties of data space
Cluster features	• Data division • Nature • Conflicting size • Contradictory densities • Poorly divided clusters • Correlation among clusters • Subspace grouping
Cluster Algorithm features	• Order dependence • Nondeterministic • Scalability • Constraint Selection • Converting the clustering problem to another field
Data renovation features	• What measure of comparison/variation should be used? • Be supposed the data be identical? • How should non correspondence of metric among variables be deal with? • How should interdependencies in the data be addressed?
Solution concern	• How many clusters should be achieved? • What group method should be used? • Should all cases be included in a cluster analysis or should some subset be ignored?
Validity concern	• Be the cluster decision different from what may be predicted by chance? • Is the cluster decision consistent or secure across samples? • Are the cluster connected to variables other than those used to originate them? Are the clusters practical?
Variable choice concern	• What is the most excellent set of variables for create a cluster analytic decision?

5. Conclusion

In this paper we have presents a brief introduction to cluster analysis in the data mining. We have also described the literature survey from last three years i.e. 2013 to 2016, analysis of clustering, classification and issues related to clustering analysis research. And we conclude that clustering is that technique of data mining which is used to extract the useful information from large data sets or un-labeled data and convert it into an understandable form that work for real world applications such as Engineering, Biology, Libraries, Educational results and database, Railways, Insurance, and Marketing etc. On the basis of this paper we can say that unprocessed data is useless without the diverse clustering techniques.

References

[1]. Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September 2010.

[2]. Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. and Weimin, Xiao, "A Visual Data Mining Framework for Convenient Identification of Useful Knowledge", ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no.-1, pp. - 530- 537, Dec 2005.

[3]. Muhammad Husnain Zafar and Muhammad Ilyas "A Clustering Based Study of Classification Algorithms" International Journal of Database Theory and Application Vol.8, No.1, pp.11-22, 2015.

[4]. Amit Kumar Kar, Shailesh Kumar Patel and Rajkishor Yadav "A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining" IJCSIT pp 139-142, ACEIT Conference Proceeding, 2016.

[5]. Namrata S Gupta, Bijendra S.Agrawal and Rajkumar M. Chauhan "Survey on Clustering Techniques of Data Mining" American International Journal of Research in Science, Technology, Engineering & Mathematics, 9(3), pp. 206-211, December 2014-February 2015.

[6]. Srinivas Sivarathri and A.Govardhan "Analysis of Clustering Approaches for Data Mining In Large Data Sources" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 9, pp 2590 – 2595, September 2014.

[7]. V. Sarala and Dr. V.V.Jaya Rama Krishnaiah "Empirical Study Of Data Mining Techniques In Education System" International Journal of Advances in Computer Science and Technology (IJACST), ISSN 2320 – 2602, Vol. 4 No.1, Pages : 15 –21, 2015.

[8]. Wei-keng Liao, Ying Liu, Alok Choudhary, "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and Engineering Datasets, pp.1-9, 2004.

[9]. Rama. B et. Al, "A Survey on clustering Current status and challenging issues", International Journal on Computer Science and Engineering, Vol. 02, No. 09, 2976-2980, 2010.

[10]. Fuyuan Cao et al, "A Framework for Clustering Categorical Time-Evolving Data", IEEE Transactions On Fuzzy Systems, Vol. 18, No. 5, pp 872-882, October 2010.

[11]. Gupta, Sachin. "A Comparative Analysis of Wired and Wireless Network Architecture." International Journal of Emerging Trends in Research 1, no. 1 (2016): 05-11.