International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0020
ISSN (Online): 2279-0039

International Journal of Engineering, Business and Enterprise Applications (IJEBEA)

www.iasir.net

INTRODUCTION TO CLUSTERING, ITS APPLICATIONS AND STACK

Neha Pathak[1], Imran Hashim[2], Anand Muni Mishra[3]
[1]Student (M.Tech. CSE), [2]HoD, [3]Research Guide
[1,2,3]Department of Computer Science & Engineering, Jawaharlal Nehru College of Technology (JNCT), Ratahara, Rewa, Madhya Pradesh 486001, INDIA

Abstract: *In this paper we will discussed about basics of clustering, applications, its usage, and clustering base. We are talking about a brief introduction to clustering, and its layer architecture. Also we have discussed about a challenges in data mining clustering and requirements for data mining clustering.*
Also we have a brief introduction of requirements of clustering for data mining techniques.

Keywords: *clustering, data mining, process, challenges, requirement.*

I. INTRODUCTION

Cluster analysis, or a clustering, refers to a set of mathematical techniques for sorting data observed in groups to maximize the similarity of the observation in the same cluster and minimize similarities observed between the different groups. These techniques may be used to explore the association and structures within the data file that may have been known. Cluster analysis was widely used in biological and social sciences to help define the classification system or taxonomy. It has also been used to suggest new ways to describe the population in enterprise applications and marketing. Clustering is a division of data into similar groups of objects. Every group, called the cluster contains the objects that are similar and different objects from other groups. In other words, the goal of good document aggregates is minimized within the cluster distance between documents, while maximizing inter-cluster distance (using a suitable distance measurement between documents).
The measure of distance (or doubling measure of similarity) is the core of the clustering. Clustering is a milestone for many areas in machine learning. Every rule has its own algorithmic bias due to the improvement of various criteria. Unsupervised machine learning is in itself an optimization process; one try to adapt the best model for sample data. The definition of "best" is unconditional; generalized with significance to the full universe of data points. However, machine learning algorithms does not understand it a priori, rather than relying on heuristics given estimates of their norms and the answer, as good conformity with relevance for experimentation facts and figures, model parsimony, and so forth. Optimization is a way to get the simplest outcome or benefit under a given set of mitigating factors. Enterprise conclusions were eventually adapted to maximize / minimize objective or benefit. The size and complexity of the issues for improvement, which can be explained in a reasonable time was advanced by the advent of technology exceeding the date of calculation.

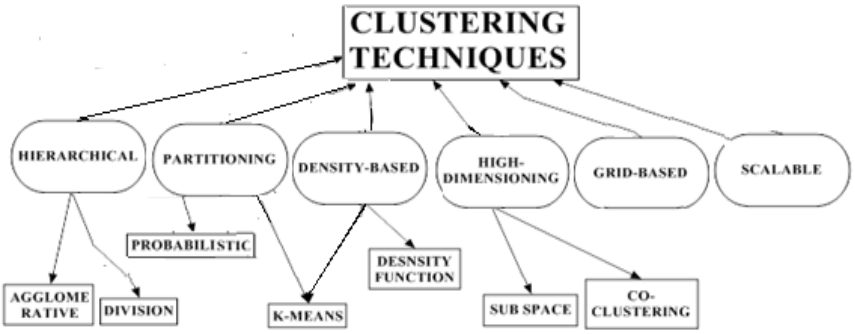

Figure 1.1: Represents various Clustering techniques.

II. APPLICATIONS OF CLUSTERING

Clustering is the most common form of supervised learning and is an important tool in a variety of applications in many fields of business and science. Thus, we summarize the basic directions that clustering is applied.
Find Similar Documents: This feature is mostly used when the user has spotted a "good" document in a search result and wants more like it. The interesting feature here is that clustering is able to find documents that are

conceptually opposed alike to seek approaches that only are able to detect whether the documents share many of the same words.

Organized Large Document Collections: Document Retrieval focus on the relevant documents for a particular query to find, but it does not; of solving the problem make sense to a large number of uncategorized documents. The challenge is to organize these documents into taxonomy identical to humans, sufficient time would create and use as a browser interface to the genuine collection of documents.

Duplicate Content Detection: In many applications, it is necessary to find a large number of duplicate or near-duplicate files. Clustering is used for plagiarism detection to organize grouping of related news and search results ranking new (higher diversity ensure the top documents). Note that the description of clusters in such applications is rarely necessary.

Recommendation System: In this application, a user item is based on the recommended items the user has already read. Clustering of products enables in real time and improves the quality of a lot.

Search Optimization: Clustering helps improve the quality and efficiency of search engines like user query can only be compared with clusters instead of comparing it directly to the documents and search results can also be arranged easily.

It is important to emphasize that getting out of a collection of documents to a clustering of the collection, is not only a single operation, but is a process in several stages. These stages contain traditional information retrieval operations such as crawling, indexing, weighing, filtration, etc. These other processes play an essential role in the quality and performance of most clustering algorithms, and should therefore be examined together with these stages of a clustering algorithm exploit true potential. We will present a detailed overview of the clustering process before we begin the study and analysis. We divided the offline process of clustering into the four stages as follows:

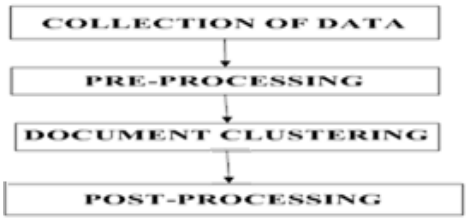

Figure 1.1: The Stages of the Process of Clustering.

Collection of Data: This includes processes such as crawling, indexing, filtering, etc. which collect the documents should be clustered index them to recover in a better way, and the filter is to remove unnecessary data, for example, stop words.

Preprocessing: It happens to represent the data in a format that can be used for grouping. There are many ways of representing documents, such as the Vector Model, graphical models, etc. Many measures are also used documents and similarities weight.

Document Clustering is the main focus of this paper, and discussed in detail.

Post-processing includes the main place where the document is used for clustering, such as the application of the recommendation, which uses clustering results to suggest news article for users.

III. CHALLENGES IN DATA MINING CLUSTERING

Clustering in high-dimensional spaces is a very difficult problem because of the curse of dimensionality phenomenon and the presence of irrelevant characteristics.

Curse of Dimensionality: By the time the curse came to refer dimensionality to a problem with the data analysis, the results of a large number of variables. For the purpose of clustering, the most important aspects of the curse of dimensionality, the influence of dimensionality to the point of proximity and density should be increased. In particular, a distance-based clustering technique depends critically on distance measure, and requires that the objects closer within the cluster are generally collectively referred to as objects in other clusters. Density-based clustering algorithms require that the density point within the group must be significantly higher than the ambient noise areas. In the mid- to high-dimensional spaces almost all pairs of points are approximately as far as the diameter and the density of the dots is within the range of the predetermined volume to the diameter. Under these circumstances, the data of "Lost in Space" and the effectiveness of the clustering algorithms, which are critical to rise depending on the distance or degree of density deteriorates rapidly dimensionality.

Irrelevant Features-Subspace Clustering: Often, that maybe not all the dimensions relevant data along such dimensions to a particular cluster are attached. The presence of irrelevant elements reduces any tendency in the data to be grouped. Intuitively, if all functions are not needed are pruned away points in each set closer to each other, which facilitates the discrimination of clusters a distance based criterion or density using. However, the feature selection techniques are susceptible to considerable loss of information, because different types of cross-

correlation attributes can be present in various subsets of dimensions in different locations of data. Therefore, it is important that each clustering algorithm to operate at full dimensional space.

IV. REQUIREMENTS FOR DATA MINING CLUSTERING

Emerging data mining applications where special requirements on clustering techniques like.

Handling High Dimensionality: Often complex, real world concepts is accompanied by a large number of functions. As a result of this sparsely filled space - the number of available point cannot grow exponentially with the dimension; it is often a full three-dimensional space very poor discrimination between clusters.

Irrelevant Features: Subspace Clustering: Often, especially in high dimensional spaces, not all dimensions are relevant data are attached along such dimensions, a given cluster. It is necessary that the clustering method, in order to detect clusters is inserted into subspaces possibly created using various combinations of dimensions of the various data locations.

Scalability: Massive data sets, both in size and dimensionality associated with Data Mining applications require highly scalable clustering algorithms. Sampling and parallelization can potentially be used to improve scalability.

Clusters of Arbitrary Shape, Size, Density, and Data Coverage: Distance-based clustering algorithms tend to find the spherical agglomerates of similar size and density. It is important to develop clustering algorithms which can detect clusters of any shape, size, density and data coverage. This would help us gain a deeper insight into the various correlations between functions, which in turn can greatly facilitate the next step of knowledge discovery in databases, eg. The decision-making processes.

Interpretability of the Results: Even the most advanced rendering techniques do not work well in large dimensional spaces, simply because the human eye-brain system is able to make a rough clustering only in three dimensions. For this reason, it is necessary to produce the cluster descriptors that can be easily assimilated by the end user, such as IF-THEN rules, decision trees.

Insensitivity to Noise: Most real-world database contains noise and outliers that do not fit into any of the clusters. The quality of clustering results may not be affected by the presence of noise and outliers.

Insensitivity to Initialization and Order of Input: It is imperative to develop clustering algorithms that produce similar results to quality regardless of the initialization phase and the order in which the input data are processed.

Minimal Requirements for Domain Knowledge:

Clustering algorithm should have a minimum of auxiliary knowledge domain to determine the input parameters, since the former is rarely complete and consistent. Furthermore, quality of the results must be relatively insensitive to the adjustment inputs. Finally, the clustering algorithm may not assume any canonical distribution data for the input data.

Handling of Different Types of Features: Given the diversity of species that data stored in the current database, eg. Numerical, categorical, multimedia, real-world applications may require the pooling of data, which consists of a mixture of data types.

V. CONCLUSION

In this paper we discussed about basics of clustering, applications of clustering, its usage, stack, and clustering base. We are talking about a brief introduction to clustering, and its layer architecture. Also we have discussed about a challenges in data mining clustering and requirements for data mining clustering.

VI. REFERENCES

[1] Pasi Franti, mohammad Rezaei, Qinpei Zhao "Centroid index: Cluster level similarity measure" Pattern Recognition, Elsevier ltd. Vol- 47, 2014, Pp 3034–3045.

[2] Alexander Solovyov, W Ian Lipkin "Centroid based clustering of high throughput sequencing reads based on n-mer counts" Solovyov and Lipkin BMC Bioinformatics 2013, Pp 1-21.

[3] Francesco Gullo, Andrea Tagarelli "Uncertain Centroid based Partitional Clustering of Uncertain Data" 38th International Conference on Very Large Data Bases, 2012. Pp 610-621.

[4] Kelvin Sim, Ghim-Eng Yap, David R. Hardoon, Vivekanand Gopalkrishnan, Gao Cong, Suryani Lukman "Centroid-based Actionable 3D Subspace Clustering" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2012. Pp 1-14.

[5] Nenad Tomasev, Milos Radovanovic´, Dunja Mladenic, Mirjana Ivanovic "The Role of Hubness in Clustering High-Dimensional Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, REVISED JANUARY 2013, Pp 1-15.

[6] A. Lourenco, A.L. Fred, A.K. Jain "On the scalability of evidence accumulation clustering" in Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10), 2010, Pp 782–785.

[7] J. Wu, H. Xiong, J. Chen "Adapting the right measures for k-means clustering" ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2009, Pp 877–885.