Applications Of Clustering Techniques In The Different Research Areas Of Applied Sciences

Ulhaskumar Patki, Jayprakash Duve and Pradeep Khot

January 18, 2020

Applications Of Clustering Techniques In The Different Research Areas Of Applied Sciences

Mr. U.S. Patki[1] Mr. J.P. Duve [2] Dr. P.G. Khot[3]

[1]Asst. Professor, Dept of Computer Science, NES Science College Nanded.

[2]Asst. Professor, Dept of Computer Science, NES Science College Nanded .

[3]Ex-Professor, Dept of Statistics, RSTM University, Nagpur.

e-mail: *patkiulhas@gmail.com*, *, duvejayprakash007@gmail.com*, *pgkhot@yahoo.com*

Corresponding Author: patkiulhas@gmail.com, +91 9860056449

Abstract:

In today's high speed and digital human life, Internet plays a imperative role. Peoples working in different research areas of applied sciences are searching their required information on the Internet. Due to advancements in web technologies, amount of data available has grown tremendously. Information retrieval from this voluminous data has become most difficult but very needy operation. Data mining is the computing process of discovering patterns in large data sets involving different methods with the integration of machine learning, statistics, and database systems. Clustering is often referred as the first steps in data mining. It identifies groups of related records that can be used as a starting point for exploring further relationships. Clustering techniques are used in different research areas of applied sciences. This paper attempts to focus application of clustering techniques in different areas of applied sciences. In the first part of the paper we have introduced clustering and clustering techniques. Later we have presented applications of clustering in different areas of applied sciences.

Keywords: information retrieval ,data mining ,clustering ,Text Mining ,BioMedical Text ,Image Segmentation ,GIS

1. Introduction

Searching is a major part in the knowledge discovery. Searching becomes easier if objects to be search are classified. From childhood, we have learned to classify objects by color, size and shapes. Human beings are skilled to divide objects into groups and assign a particular object to a particular group. Clustering or cluster analysis is a process of grouping the objects having similar attributes. Main objective of clustering is to minimize intra-cluster distance and to maximize inter-cluster distance. In knowledge discovery, search plays very important role. Grouping of objects having similar attributes will increase efficiency of searching process. In different areas of applied sciences clustering plays very significant role. For example, in chemical sciences, Periodic Table is the result of a clustering of the elements in groups that presented similar physical properties. We can easily locate any element within a periodic table. In life science also researches classify the similar species having common attributes into one group. Many times the words classification and clustering are used interchangeably. Though the purpose of classification and clustering is similar, these two are different methods.

In classification you have a set of predefined classes and want to know which class a new object belongs to. Clustering tries to group a set of objects and find whether there is some relationship between the objects. In the context of machine learning, classification is supervised learning and clustering is unsupervised learning. Clustering has many different names. In biology, clustering analysis called taxonomy". In pattern recognition, it is called "unsupervised learning."

In this paper we have discussed applications of clustering techniques in different research areas of applied sciences. Initially we have given brief introduction of clustering techniques and then we have discussed different applications of clustering in applied sciences.

2. Clustering Techniques

Basically clustering techniques are divided into two categories, Namely Hierarchical Clustering and Partitioned Clustering.[1]

A hierarchical clustering technique divides the given data set into smaller subsets in hierarchical fashion. On the other hand, a partition clustering algorithm partitions the data set into N number of Clusters. Hierarchical technique further divided into two types namely agglomerative and divisive clustering. An agglomerative clustering start with singleton point into a cluster and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits into the most appropriate clusters. The process continues until a stopping criterion which is defined by the user is achieved.

There are two important issues in clustering. First, to find the possible number of clusters in a given dataset and second, for given two sets of clusters, to compute a relative measure of goodness between them. The most commonly used partitioning algorithm is the K-means algorithm. K-means algorithm iteratively move objects between partitions to optimize objective function. The objective function is usually a sum of distances (or sum of squared distances) between objects and their cluster's centers and the objective is to minimize it[2]. Partitioned clustering is further divided into two categories namely Hard Clustering and Soft Clustering.

Hard clustering and Soft Clustering:

In Hard clustering, single document belongs to exactly one and only one cluster. In Soft clustering overlapping is permitted i.e. single object may belongs to multiple clusters. The two types of classic hard Clustering techniques are Hierarchical Document Clustering and Partitional Document Clustering using K-Means. Soft Computing techniques are Fuzzy Logic, Neural Network, Multilayered perception, Radial Bias function network, Genetic algorithms etc. Soft Computing techniques make computers to behave like human beings.[3]

The fig. bellow shows a clustering process.

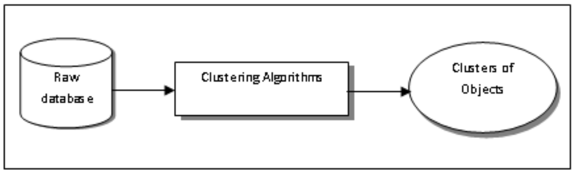

Fig. 1 Clustering Process

3. Applications of Clustering

In the literature clustering techniques are used in variety of applications in applied sciences. In this section we discuss some important areas of research in which clustering techniques plays a vital role.

a) Application of Clustering in Text Mining:

Text mining, which is also termed as text data mining, is defined as the process of deriving high-quality information from textual database. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Document clustering is one of the most important text mining methods that are developed to help users effectively navigate, summarize, and organize text documents.

By organizing a large amount of documents into a number of meaningful clusters, document clustering can be used to browse a collection of documents or to organize the results returned by a search engine in response to a user's query. It can significantly improve the precision and recall in information retrieval systems, and it is an efficient way to find the nearest neighbors of a document. Hence by using document clustering we can find out similar text document very easily.[4]

b) Bio Science Applications and Biomedical knowledge based search engine for biomedical text:

MEDLINE is a major biomedical literature database repository that is supported by the U.S. National Library of Medicine (NLM). It has now generated and maintained more than 15 million citations in the field of biology and medicine, and incrementally adds thousands of new citations every day. Researchers can no longer keep up-to-date with all the relevant literature manually, even for specialized topics. As a result, information retrieval tools play vital roles in facilitating researchers to search and access relevant papers. These Information retrieval tool uses integration of text mining and different clustering methods to find out required information. Frequently, biomedical researchers query the MEDLINE database and retrieve lists of citations based on given keywords. PubMed, an information retrieval tool, is one of the most widely-used interfaces to access the MEDLINE database. It allows Boolean queries based on combinations of keywords and returns all citations matching the queries. Many advanced retrieval methods, such as GoPubMed and Textpresso, also use natural language processing methods to better identify documents relevant to a query. Even with these improvements, significant challenges remain to efficient and effective utilization of ad hoc information retrieval systems such as PubMed. [5]

In biotechnology, cluster analysis was used to categorized animal and plant populations according to population and to obtain the latent structure of knowledge;

c) *Image Segmentation in Image Processing:*

Segmentation refers to a technique in which an image in digital form is partitioned into multiple segments basically groups of pixels. In the literature, there are number of techniques available for segmenting an image. These techniques are classified as follows:

- Region Based Segmentation
- Edge Based Segmentation
- Threshold Segmentation
- Clustering Based Segmentation

Clustering in image processing is basically defined as the technique in which groups of similar image primitive are identified. Clustering is a method in which objects are combined into groups based on their attributes. Data clustering algorithms are built over the whole image and these algorithms are applied to study data distances.

In image segmentation, basically partitioned clustering technique is most popular. In Partition techniques, K-means is a Hard clustering while Fuzzy C-means is a Soft clustering technique. [6]

d) Business Applications:

In traditional markets, customer clustering / segmentation is one of the most significant methods used in studies of marketing. This study classifies existing customer cluster/segmentation methods into methodology-oriented and application-oriented approaches. Customer clustering is the most important data mining methodologies used in marketing and customer relationship management (CRM). Customer clustering would use customer-purchase transaction data to track buying behavior and create strategic business initiatives. [7]

Some Additional application of clustering in Business and Marketing are Human resource management, Customer Relationship Management and Market Analysis.

e) GIS Applications:

The clustering is a very important and effective thing in the field of GIS. The main aim of clustering is to group data sets. The grouping of data is a fundamental part of every GIS system. The application of clustering in GIS ranges from soil type grouping to crop clustering and many more. Clustering in GIS is even more challenging because GIS deals with huge data. So, the clustering as well as data structure for storing the knowledge about cluster is one of the major concerns of literally every GIS system.[9]

In addition to above applications, cluster analysis has been applied to many different areas. In geography, clustering can help biologists to determinate the relationship of the different species and different geographical climate; in the banking sector, by using cluster analysis to bank customers to refine a user group; in the insurance industry, according to the type of residence, around the business district, the geographical location, cluster analysis can be used to complete an automatic grouping of regional real estate, to reduce the manpower cost and insurance company industry risk [8]

4. Conclusion

Pattern recognition is a promising field in the modern computations. It is very usefully in almost all research areas of applied sciences. Clustering is a field of pattern recognition. Clustering is applicable in Information retrival like text mining, in life science it is used to classify animals and plants to obtain the latent structure of knowledge. It is also used in image segmentation an emerging research filed in computer science. Clustering is also applicable in Business applications like Human resource management, Customer Relationship Management and Market Analysis. If overlapping does not exists in patterns to be recognized, one can use hard clustering like K-means, while if a possibility of overlapping exists, Soft clustering like Fuzzy c-Means techniques are helpful

5. References

[1] Nisha, IEEE, 2nd International Conference , New Delhi, 11-13 March 2015, “A survey of clustering techniques and algorithms”

[2] Sowmya P , et.al IJAECS ISSN: 2393-2835, Special Issue, Sep.-2016 , “Survey on Algorithms used for Text Document Clustering”

[3] Dibya Jyoti Bora,et. al, International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 2 – Apr 2014 Page108 “A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm”

[4] J.Sathya Priya ,et. al, (IJCSIT), Vol. 3 (1) , 2012, 2943-2947 “Clustering Technique in Data Mining for Text Documents”

[5] Yongjing Lin, et. al. Journal of the American Medical Informatics Association JAMIA . 2007 Sep-Oct; 14(5): 651–661. “A Document Clustering and Ranking System for Exploring MEDLINE Citations”

[6] Priyansh Sharma , et.al. International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.9, No.5 (2016), pp.209-218 “A Review on Image Segmentation with its Clustering Techniques”

[7] Dr. Sankar Rajagopal, International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011, “Customer Data Clustering Using Data Mining Technique”

[8] R.Roseline ,et. al, International Journal of Computing Algorithm,Volume: 03, May 2014, Pages: 910-912
“Analysis and Application of Clustering Techniques in Data Mining”
9] Parthajit Roy,Annual Progress Report , 2013 The University of Kalyani,” Clustering and its application to GIS”