Exploring Cluster Analysis

Mini Singh Ahuja
Punjab technical University
Jalandhar, India
Minianhadh {at} yahoo.co.in

Jatinder Singh Bal
KC Group of institutes
Nawashahr, India

Abstract— **Data mining is a process of analyzing data from different fields and then summarizing it to get useful information out of it. Clustering is one of the important activities in data mining. Clustering is a process of grouping data into classes or clusters so that objects within a cluster have more similarity than between different clusters. The aim of this paper is to give a brief idea about clustering and give future scope of it in the research field.**

***Keywords*- data mining, clustering, community detection.**

I. Introduction

In recent years data mining has attracted a lot of attention in the information industry and society due to the huge availability of data. This data can then be used to extract useful information and knowledge which is used for various applications like market analysis, fraud detection, and science exploration. The three main tasks in data mining are regression, classification [14] and clustering.

In classification we group the data in to a set of predefined classes and want to know about the class of the new object. But in clustering we try to group the set of objects without prior knowledge of classes. So classification can be classified as supervised learning where as clustering can be classified as unsupervised learning. In data mining lots of efforts have been made on finding efficient and effective cluster analysis in large databases.

II. Cluster Analysis

Clustering is an important area within data mining. Clustering aims at partitioning the data into groups such that the data objects assigned to a common group called cluster, are as similar as possible and the objects assigned to different clusters differ as much as possible Clustering helps users in understanding the structure in a data set. Cluster analysis can be used as standalone tool to get knowledge about data distribution or it can be used as preprocessing step for other algorithms. Cluster analysis can also be used for summarizing data rather than for finding “natural” or “real” clusters. This type of cluster analysis is called dissection. We can use any type of data in cluster analysis such as interval, ordinal or categorical. But if we use a mixture of different types of variable the analysis will be more complicated. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis which can be used in various applications like market research, image processing, bioinformatics, pattern recognition etc. Clustering can help marketers discover different group of customers based on purchasing patterns. Later this information about the different groups can be used to make recommendation systems for the viral market. Clustering can also be used for outlier detection. It is a very challenging field of research in recent years. These days a lot of research is going on the scalability of clustering methods, effectiveness of different clustering methods for complex data etc.

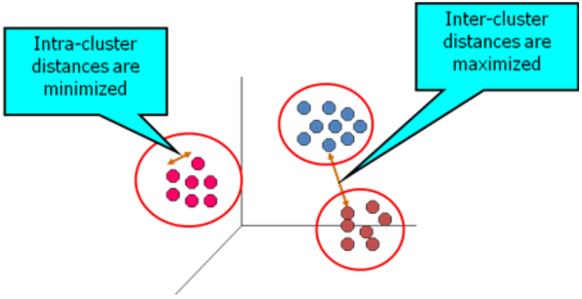
Intra-cluster distances are minimized
Inter-cluster distances are maximized

Fig1: clusters

III. Community Detection vs Cluster Analysis

Community detection is one of the widely discussed problems in network science and it has a wide range of applications. These applications include detecting friend circles in online groups, grouping similar types of proteins in protein-interaction networks, and uncovering clusters that are separated by characteristics such as geographic location. Sometimes terms Cluster Analysis and community detection are misunderstood as different terms. Both terms are similar with a little difference. Community detection is a type of graph clustering where the data to be used for detection of clusters deals with relationships instead of features. Community structure is one of the important properties of complex networks. So we can say community detection is a clustering problem with a specific attributes. Community detection is also an active field of research in network science these days.

IV. Clustering Techniques

Many clustering algorithms have been proposed till today which can be categorized into partitioning methods, hierarchical methods, density-based methods and grid-based methods.

A. Hierarchal methods:

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be either agglomerative or divisive, based on how the hierarchical decomposition is formed. Agglomerative approach also called the bottom-up approach starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one, or until a termination condition holds. The divisive approach also called the top-down approach starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. This technique is used in biological, social and behavioral science where there is a need to construct taxonomies. Few hierarchical clustering based algorithms are SLINK [11], COBWEB, CURE [4] and CHAMELEON [2].

B. Partitioning methods

Partitioning methods are the most simple and fundamental method of clustering. It organizes the objects of a set into several exclusive groups or clusters. Number of clusters is known in advance which acts as a starting point for partitioning method. In this technique a given data set D of n objects and k no of clusters are organized in such a way that partition represents a cluster. The clusters are formed to optimize an objective partitioning criteria such as dissimilarity function based on distance. This technique is used in engineering applications where single partitions are important. More over portioning methods are also used in efficient representation and compression of large databases. This technique is further divided into probabilistic Clustering (EM framework, algorithms SNOB, AUTOCLASS, MCLUST), k-medoids methods (algorithms PAM, CLARA, CLARANS), and k-means methods (different schemes, initialization, optimization, harmonic means, extensions).

C. Density-Based method

This technique is based on discovering dense connected components of data, which are flexible in terms of their shape. Density-based connectivity is used in the algorithms like DBSCAN [12], OPTICS [13], DBCLASD, GDBSCAN, WaveCluster [6]. These algorithms are less sensitive to outliers and can discover clusters of irregular shapes. They usually work with low-dimensional data of numerical attributes, known as spatial data.

D. Grid based methods

Grid-based methods work with attributes of different types. This technique takes a space driven approach by partitioning the embedding space into cells independently of the distribution of the input objects. The important grid-based algorithms are STING [3], CLIQUE [5] and MAFIA.

V. CHALLENGES IN CLUSTER ANALYSIS

A. Identification of clusters:
Identifying the number of clusters is a difficult task if the number of class labels is not known beforehand.

B. Selection of distance measure:
There are many distance measures like eucledian, manhattan, and maximum distance measure, but the difficulty lies in selecting which measure is best for which type of data.

C. Structure of database:
Real life data may not always contain clearly identifiable clusters. Also the order in which the tuples are arranged may affect the results when an algorithm is executed if the distance measure used is not perfect.

D. Types of attributes:
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.

E. Selecting the starting cluster:
Starting cluster plays an important role in partitioning method, so we need to a solid algorithm which selects a initial cluster.

F. Choosing the best method of clustering:
There are many algorithms for clustering till today. But the question is which one gives the best result.

G. Cluster validity:
Different clustering methods can generally produce different solutions on the same data, the question arises whether the clusters have ”reality” or validity vis-a-vis the data

H. Standard benchmarks:
There are no standard benchmarks for comparing the partitions. So, the authors of new techniques are reduced to devising their own comparisons to previous techniques. Usually authors obtain a few of the data sets used by previous authors and run comparisons based on those. Even the criteria to be measured are not standard.

VI. APPLICATIONS

Clustering algorithms can be used in a large variety of applications. These areas are: image segmentation, object and character recognition, document retrieval, data compression, data mining etc.

- Clustering can also help marketers discover distinct groups in their customer basis and they can characterize their customer groups based on purchasing patterns.

- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- Cluster Analysis acts as a tool to gain insight into the distribution of data to observe characteristics of each cluster.
- In field of biology it can be used to derive plant and animal taxonomies, categorize genes with similar functionality and gain insight into structures inherent in populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according house type, value, and geographic location.
- Clustering can be used to reduce the number of patterns that need to be considered in pattern recognition [15].
- Cluster analysis can be used in crime analysis to identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.
- Cluster analysis can be used to study the patterns in atmospheric pressure of Polar Regions and oceans which play important role on the land climate.
- In the study of social networks, clustering can be used to recognize communities within large groups of people.

VII. CLUSTERING TOOLS AND PACKAGES

There are many commercial and open source tools available for cluster analysis [1]. Few of these are:

Clustan: Clustan includes a collection of procedures for performing cluster analysis. It helps in designing software for cluster analysis, data mining, market segmentation, and decision trees.

CLUTO: CLUTO is software used for low- and high-dimensional datasets and for analyzing the characteristics of the various clusters.

XLMiner: It includes various statistical and machine learning techniques for classification, prediction, affinity analysis and data exploration and reduction.

Weka: Weka has a collection of machine learning algorithms for data mining tasks and is capable of developing new machine learning schemes.

Matlab Statistical Toolbox: It has a collection of tools which are built on the MATLAB for performing numeric computations.

CViz: Cluster Visualization tool is used for analyzing large high- dimensional datasets. It also provides full-motion cluster visualization.

Viscovery: It is used for explorative data mining modules, with visual cluster analysis, segmentation, and assignment of operational measures to defined segments.

Cluster3: It is an open source clustering software which contains clustering routines that can be used to analyze gene expression data.

VIII. RECENT ADVANCES IN CLUSTER ANALYSIS

In recent years information exploration has created large amounts of data and too of different types namely: structured and unstructured. *Unstructured data* is a collection of objects that do not follow a specific format such as images, text, audio, video, etc. On the other hand, in *structured data,* there are semantic relationships within each object that are important. This increase in data has attracted many researchers to explore useful information out of this data. Below there are few research areas where further work can being carried out in the field of cluster analysis:

A. Ensembles of clustering algorithms

These days a lot of research is going on in combining different clustering methods for unsupervised learning. The basic idea is that by taking multiple looks at the same data, one can generate multiple partitions of the same. Combining multiple clustering algorithms is a more challenging problem than combining multiple classifiers. Cluster ensembles can be done in different ways

- The use of a number of different clustering techniques
- The use of a single technique many times with different initial conditions
- The use of different partial subsets of features or patterns.

B. Distributed clustering

Due to the increasing size of current databases, constructing efficient distributed clustering algorithms has attracted considerable attention. Distributed Clustering assumes that the objects to be clustered reside on different sites. So instead of sending all objects to a central site (server) we can apply standard clustering algorithms to analyze the data at their local sites (clients).

C. Multi-way clustering

Sometimes objects which are to be clustered consist of heterogeneous data. In such situations objects are converted into a pooled feature vector of its components prior to clustering, but it is not a natural representation of the objects and may result in poor clustering performance.

D. Dynamic data

A lot of research is carried being carried out in handling dynamic data such as web pages and blogs where data keep on modifying with course of time so clustering must also be updated accordingly.

E. Graph Data:

Graph mining [8] has been popular area of research in recent year because several objects, such as chemical compounds, social networks, protein structures, etc. are represented most naturally as graphs. Recently efforts are going on in extracting graph features to allow existing clustering algorithms to be applied to the graph feature vectors. The features can be extracted based on patterns such as frequent sub-graphs, shortest paths, cycles, and tree-based patterns.

IX CONCLUSION

Clustering analysis techniques is an interdisciplinary subject. Social scientists, psychologists, biologists, statisticians, mathematicians, engineers, computer scientists, medical researchers, and others have all contributed to clustering methodology. It is a useful and challenging problem which has many potential applications like information filtering, image segmentation, software bug localization etc. Cluster analysis is still an active field of research. In this paper our aim was to produce a critical review of clustering and to study the future scope of clustering techniques.

References

[1] Data Mining And Knowledge Discovery Software Tools", SIGKDD Explorations, Vol. 1, No. 1, P. 20-33, June 1999.

[2]Karypis G., Han E. H. and Kumar V. (1999), CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, Computer 32(8): 68-75, 1999.

[3] STING: A Statistical Information Grid Approach to Spatial Data Mining, Proceedings of the 23rd VLDB Conference Athens, Greece, 1997

[4] Guha, S., Rastogi, R., Shim K. (1998), "CURE: An Efficient Clustering Algorithm for Large Data sets", Published in the Proceedings of the ACM SIGMOD Conference.

[5] [1] Agrawal R., Gehrke J., Gunopulos D. and Raghavan P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In Proc. of the 1998 ACM-SIGMOD Conf. On the Management of Data, 94-105

[6] Sheikholeslami, C., Chatterjee, S., Zhang, A. (1998), "WaveCluster: A-MultiResolution Clustering Approach for Very Large Spatial Data set". Proc. of 24th VLDB Conference

[7]Michael Steinbach, George Karypis, and Vipin Kumar, *A Comparison of Document Clustering Techniques*, University of Minnesota, Technical Report #00-034, 2000

[8] George Karypis and Vipin. Kumar, (1998), *METIS 4.0: Unstructured graph partitioning and sparse matrix ordering system*, Technical report, Department of Computer Science, University of Minnesota, 1998.http://www-users.cs.umn.edu/~karypis/metis/

[9] Richard C. Dubes and Anil K. Jain, (1988), *Algorithms for Clustering Data*, Prentice Hall

[10]L. Kaufman and P. J. Rousseeuw, (1990), *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley and Sons.

[11] R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method". *The Computer Journal* (British Computer Society) **16** (1): 30–34.

[12]N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In Proceedings of the thirtyseventh annual ACM Symposium on Theory of Computing, pages 684–693, 2005

[13] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". *ACM SIGMOD international conference on Management of data*. ACM Press. pp. 49–60.

[14]ACM, 1994. ACM CR Classifications. *ACM Computing Surveys 35*, 5–16

[15] BEZDEK, J. C. 1981. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, NY.

[16] COLEMAN, G. B. AND ANDREWS, H. C. 1979. Image segmentation by clustering. *Proc. IEEE 67*, 5, 773–785.\

[17]LEE, R. C. T. 1981. Cluster analysis and its applications. In *Advances in Information Systems Science*, J. T. Tou, Ed. Plenum
Press, New York, NY