Clustering Analysis in Data Mining

P.Geetha[1] P.Panneerselvan[2]
[1,2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Kaamadhenu Arts and Science College, Sathyamangalam

Abstract— Clustering analysis is a key point used by data processing algorithms in Data Mining. The primary aim of Clustering is to segment the data into more diminutive subsets called clusters, such that the data belonging to the same cluster are similar with some similarity metric. Clustering is imperative idea in data investigation and data mining applications. Over years, K-means has been popular clustering algorithm because of its ease of use and simplicity. This paper presents the introduction to cluster analysis in the field of data mining, where to be the discovery of useful, but non-obvious, information or patterns in large collections of data. Much of this paper is necessarily consumed with providing a general background for cluster analysis, and also discusses a number of clustering techniques that have recently been developed specifically for data mining.

Key words: Data Mining, Cluster Analysis, Clustering Applications, Requirements, Various Clustering Methods

I. INTRODUCTION

Data mining is the useful tool to discovering the knowledge from large data. It is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. A wide variation exists in terms of the problem domains, applications, formulations, and data representations that are encountered in real applications. Therefore, "data mining" is a broad umbrella term that is used to describe these different aspects of data processing.

Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning. The goal of clustering is descriptive, that of classification is predictive.

II. DATA MINING

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

III. CLUSTERING ANALYSIS

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning.

Cluster analysis groups objects based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar to one other and different from the objects in other groups. To better understand the difficulty of deciding what constitutes a cluster, consider figures 1 through 4, which show twenty points and three different ways that they can be divided into clusters. If allow clusters to be nested, then the most reasonable interpretation of the structure of these points is that there are two clusters, each of which has three subclusters. However, the apparent division of the two larger clusters into three subclusters may simply be an artifact of the human visual system. Finally, it may not be unreasonable to say that the points form four clusters.

Fig. 1: Initial Points

Fig. 2: Six Clusters

Fig. 3: Two Clusters

Fig. 4: Four Clusters

IV. REQUIREMENTS OF CLUSTERING IN DATA MINING

The following points throw light on why clustering is required in data mining −

A. Scalability*:*

Need highly scalable clustering algorithms to deal with large databases.

B. Ability to deal with different kinds of attributes:

Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.

C. Discovery of clusters with attribute shape:

The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to

only distance measures that tend to find spherical cluster of small sizes.

D. High dimensionality:

The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.

E. Ability to deal with noisy data:

Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

F. Interpretability:

The clustering results should be interpretable, comprehensible, and usable.

V. CLUSTERING METHODS

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce.

The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

VI. PARTITIONING METHOD

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster.

Figure-5 and Figure-6 presents the original points and partitional clustering respectively.

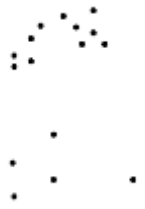

Fig. 5: Original Points

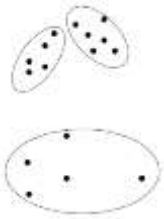

Fig. 6: Partitional Clustering

Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.
- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

VII. HIERARCHICAL METHODS

This method creates a hierarchical decomposition of the given set of data objects. It can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. Figure-7 express about hierarchical clustering with the agglomerative and divisive approach.

There are two approaches here −

- Agglomerative Approach
- Divisive Approach

A. 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. Figure-7 express about hierarchical clustering with the agglomerative and divisive approach.

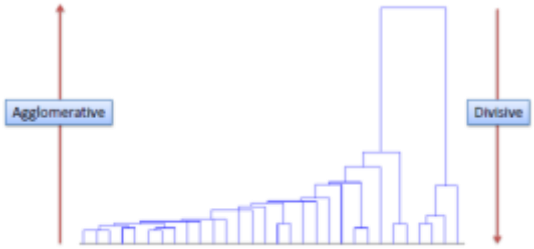

Fig. 7: Hierarchical Clustering

B. Divisive:

This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

1) Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

2) Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure. The major advantage of this method is fast processing time. It is dependent only on the number of cells in each dimension in the quantized space.

3) Model-based Method

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

C. Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering

process. Constraints can be specified by the user or the application requirement.

D. Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Identification of areas of similar land use in an earth observation database 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.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

VIII. Conclusion

Clustering lies at the heart of data analysis and data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change.

In this paper to described the process of clustering from the data mining point of view. Try to give the basic concept of clustering by first providing the definition and clustering and then the definition of some related terms. Give some examples to elaborate the entire concept. Then give different approaches to data clustering and also discussed some algorithms to implement that approaches. The partitioning method and hierarchical method of clustering were explained. The applications of clustering are also discussed.

References

[1] Jim Gray "QqJim Gray's NT Clusters Research Agenda" Microsoft Online Research Papers, Microsoft Corporation.
[2] Kevin E. Voges, Nigel K. Ll. Pope and Mark R. Brown, "Cluster Analysis of Marketing Data Examining", Griffith University, Australia, 2002.
[3] Lior Rokach, Oded Maimon "Clustering Methods", Tel-Aviv University.
[4] Mirek Riedewald," Data Mining Techniques: Cluster Analysis", 2012
[5] Periklis Andritsos, "Data Clustering Techniques", University of Toronto Department of Computer Science, 2002.
[6] SAS, Global forum handout, "Introduction to Data mining", 2015.
[7] S.Parthasarathy, V.Shakila, "Knowledge Clustering On Big Data With K_Means Algorithm", Department of MCA, Valliammai Engineering College, Tamilnadu, India, 2015
[8] Tan, Steinbach, Kumar, "Data Mining Cluster Analysis: Advanced Concepts and Algorithms" 2004.

