RESEARCHERID
THOMSON REUTERS

[Saurkar* *et al.*, 6(4): April, 2017]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# AN OVERVIEW ON DIFFERENT CLUSTERING METHODS USED IN DATA MINING

**Anand V. Saurkar [*], Shweta A. Gode**

[*] Department of Computer Science & Engineering, Datta Meghe Institute of Engineering, Technology & Research (DMIETR), Wardha (MH), India

[**] Department of Computer Technology, Yashwantrao Chavan College of Engineering (YCCE), Nagpur (MH), India

**DOI**: 10.5281/zenodo.557141

## ABSTRACT

Through data mining, we can able to effectively extract data in the form of knowledge discovery which provides useful helping guide for information processing that can be utilized in varieties of applications. It is the most sought after field in recent scenario and its importance cannot be ignored at all as effective data analysis outputs to extensive information utilization in almost all the fields and a proper data mining provides the appropriate and effective result. In this paper we focus on basics of clustering techniques and different major clustering methods.

**KEYWORDS**: Data Mining, Knowledge Base, clustering.

## INTRODUCTION

Nowadays databases are comprised of terabytes or more data in it. As they are able to accommodate huge mass of heterogeneous data, different variety of strategic information lies hidden inside it. Database technology has evolved from primitive file processing to the development of database management systems with query and transaction processing. Further progress has led to the increasing demand for efficient and effective advanced data analysis tools. This need is a result of the explosive growth in data collected from applications, including business and management, government administration, science and engineering, and environmental control. So, through effective data mining only we can able to draw meaningful conclusions which are the basic purpose of data mining. As data are being accumulated continuously as well as rapidly whether it is a research field, education, market products, medical science, electronic information, media, entertainment etc. it is difficult to get faster and appropriate information by traditional manual analysis which is tedious as well as very cumbersome. So, data mining is used basically i) to reduce costs through proper detection and prevention of waste and fraud, ii) obtaining appropriate and up-to-date information and iii) increase revenues through improved marketing strategy.

## WHAT IS DATA MINING?

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data,". "Knowledge mining," a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a bright term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

Thus, such a misnomer that carries both "data" and "mining" became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in Figure 1.1 and consists of an iterative sequence of the following steps:

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

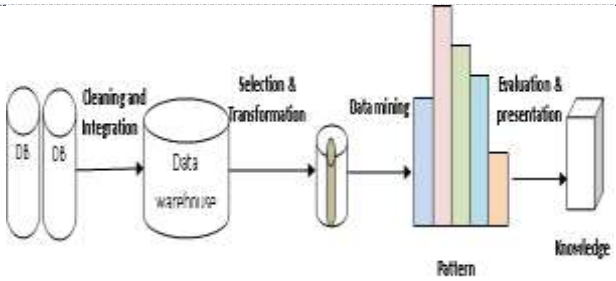


Fig 1: Data mining as a step in the process of knowledge discovery

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)1

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)2

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Based on this view, the architecture of a typical data mining system may have the following major components :


Fig 2: Architecture of a typical data mining system.

Database, data warehouse, World Wide Web, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Imagine that you are given a set of data objects for analysis where, unlike in classification, the class label of each object is not known. This is quite common in large databases, because assigning class labels to a large number of objects can be a very costly process. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning.

## WHAT IS CLUSTER ANALYSIS?

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

As a branch of statistics, cluster analysis has been extensively studied for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS. In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

## A CATEGORIZATION OF MAJOR CLUSTERING METHODS

Many clustering algorithms exist in the literature. It is difficult to provide a crisp categorization of clustering methods because these categories may overlap, so that a method may have features from several categories. Nevertheless, it is useful to present a relatively organized picture of the different clustering methods.

In general, the major clustering methods can be classified into the following categories.

Partitioning methods: Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$. That is, it classifies the data into k groups, which together satisfy the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. Notice that the second requirement can be relaxed in some fuzzy partitioning techniques. References to such techniques are given in the bibliographic notes.

Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects of different clusters are "far apart" or very different. There are various kinds of other criteria for judging the quality of partitions.

To achieve global optimality in partitioning-based clustering would require the exhaustive enumeration of all of the possible partitions. Instead, most applications adopt one of a few popular heuristic methods, such as (1) the k-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, and (2) the k-medoids algorithm, where each cluster is represented by one of the objects located near the center of the cluster. These heuristic clustering methods work well for finding spherical-shaped clusters in small to medium-sized databases. To find clusters with complex shapes and for clustering very large data sets, partitioning-based methods need to be extended.

Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The 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 (the topmost level of the hierarchy), 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.

Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices. However, such techniques cannot correct erroneous decisions. There are two approaches to improving the quality of hierarchical clustering: (1) perform careful analysis of object "linkages" at each hierarchical partitioning, such as in Chameleon, or (2) integrate hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into microclusters, and then performing macroclustering on the microclusters using another clustering method such as iterative relocation, as in BIRCH.

Density-based methods: Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape.

DBSCAN and its extension, OPTICS, are typical density-based methods that grow clusters according to a density-based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions.

Grid-based methods: Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

STING is a typical example of a grid-based method. WaveCluster applies wavelet transformation for clustering analysis and is both grid-based and density-based.

Model-based methods: Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

of clusters based on standard statistics, taking "noise" or outliers into account and thus yielding robust clustering methods.

EM is an algorithm that performs expectation-maximization analysis based on statistical modeling. COBWEB is a conceptual learning algorithm that performs probability analysis and takes concepts as a model for clusters. SOM (or self-organizing feature map) is a neural network-based algorithm that clusters by mapping high dimensional data into a 2-D or 3-D feature map, which is also useful for data visualization.

## CONCLUSION

Data mining is a very demanding and most sought after area now-a days. Data mining enhances understanding by showing which factors most affect specific outcome. For any development purpose/analysis purpose effective study of data provides an effective outcome which is possible through perfect data mining. In today's advanced world, by implementing various advanced data mining techniques, we can able to obtain effective data mining outputs which provide immense knowledge through which a system works perfectly and reasonability according to their own requirement and extreme satisfaction.

## REFERENCES

1. Jiawei Han, Micheline Kamber University of Illinois at Urbana-Champaign, Data Mining: Concepts and Techniques Second Edition.
2. Chin-Ang Wu, Wen-Yang Lin, Chang-Long Jiang, ChuanChun Wu, Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining, Expert Systems with Applications, Volume 38, Issue 9, September 2011, Pages 11011-11023.
3. Ling Chen, Mingqi Lv, Qian Ye, Gencai Chen, John Woodward, A personal route prediction system based on trajectory data mining, Information Sciences, Volume 181, Issue 7, 1 April 2011, Pages 1264-1284
4. Aronis, J.M., Provost, F.J., & Buchanan, B.G. Exploiting background knowledge in automated discovery. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (pp. 355–358).
5. Yunfei Yin, A proximate dynamics model for data mining, Expert Systems with Applications, Volume 36, Issue 6, August 2009, Pages 9819-9833.
6. Haifeng Li, Hong Chen, Mining non-derivable frequent itemsets over data stream, Data & Knowledge Engineering, Volume 68, Issue 5, May 2009, Pages 481-498.
7. Naixue Xiong, Laurence T. Yang, Yingshu Li, ODMCA: An adaptive data mining control algorithm in multicarrier networks, Computer Communications, Volume 32, Issue 3, 25 February 2009, Pages 560-567.
8. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-211.