CHAPTER 3

DATA MINING AND CLUSTERING

3.1 Introduction

Nowadays, large quantities of data are being accumulated. The amount of data collected is said to be almost doubled every 9 months. Seeking knowledge from massive data is one of the most desired attributes of Data Mining [40]. Data could be large in two senses. In terms of size, example is Image Data or in terms of dimensionality, example for Gene expression data.

Usually there is a huge gap from the stored data to the knowledge that could be constructed from the data. This transition won't occur automatically, that's where Data Mining comes into picture. In Exploratory Data Analysis, some initial knowledge is known about the data, whereas in Data Mining in-depth knowledge is known about the data [87].

Clustering is useful for exploring data. If there are many cases and no obvious natural groupings, clustering algorithms can be used to find natural groupings. Clustering analysis identifies clusters embedded in the data. A cluster is a collection of data objects that are similar in some sense to one another [20]. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words,
members of a cluster are more like each other than they are like members of a different cluster [30].

### 3.2 Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information, which can be used to increase revenue, cut costs, or both [50]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified [82]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

#### 3.2.1 Data Mining and Knowledge Discovery

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Data mining and KDD are frequently treated as synonym; data mining is actually part of the
knowledge discovery process. The following Figure 3.1 shows data mining as a step in an iterative knowledge discovery process.

Figure:3.1 Data Mining and Knowledge Discovery Process

The KDD process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:
**Data cleaning**: Also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

- **Data integration**: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: In this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: This is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse [89]. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or as for the case of data warehouses, the selection is done on transformed data.
The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either shifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "Gold Mining" and not "Rock Mining", thus by analogy, data mining should have been called "Knowledge Mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

3.2.2 Functions and Elements of Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical
software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities [107].

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.

- Store and manage the data in a multidimensional database system.
• Provide data access to business analysts and information technology professionals.

• Analyze the data by application software.

• Present the data in a useful format, such as a graph or table.

3.2.3 Knowledge Discovered in Data Mining

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data.

The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

• **Characterization**: Data characterization is a summarization of general features of objects in a target class, and produces what is called character rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

• **Discrimination**: Data discrimination produces what are called discriminate rules and is basically the comparison of the general features
of objects between two classes referred to as the target class and the contrasting class.

- **Association analysis**: Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

- **Classification**: Classification analysis is the organization of data in given classes [69]. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

- **Prediction**: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the
classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/decrease trends in time related data [35]. The major idea is to use a large number of past values to consider probable future values.

- **Clustering**: Similar to classification, clustering is the organization of data in classes [88]. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels [32]. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (*intra-class similarity*) and minimizing the similarity between objects of different classes (*inter-class similarity*).

- **Outlier analysis**: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

- **Evolution and deviation analysis**: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data.
Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity of an important attribute of a data mining system.

### 3.3 Cluster Analysis in Data Mining

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models. Clustering models are different from supervised models in that the outcome of the process is not guided by a known result, that is, there is no target attribute. Supervised models predict values for a target attribute, and an error rate between the target and predicted
values can be calculated to guide model building. Clustering models, on the other hand, are built using optimization criteria that favor high intra-cluster and low inter-cluster similarity. The model can then be used to assign cluster identifiers to data points.

### 3.3.1 Clustering Methods

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi objective optimization that involves trial and failure, but this research work is stepping stone of optimization. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.
The clustering methods can be classified into five approaches as categories below:

- Partitioning method
- Hierarchical method
- Density Based method
- Grid Based method
- Model Based method

All the above mentioned approaches has its own specialization and traditional methods which are commonly used in the data mining process of real application. The following Figure: 3.2 shows the taxonomy of Clustering Analysis.

Figure: 3.2 Taxonomy of Cluster Analysis
Hierarchical methods obtain a nested partition of the objects resulting in a tree of clusters. These methods either start with own cluster or then split into smaller and smaller clusters (called divisive or top down) or start with each object in an individual cluster and then try to merge similar clusters into larger and larger clusters (called agglomerative or bottom up). In this approach, in contrast to partitioning, tentative clusters may be merged or split based on some criteria.

Partitioning methods obtain a single level partition of objects. These methods usually are based on greedy heuristics that are used iteratively to obtain a local optimum solution. Given n objects, these methods make k<=n clusters of data and use an iterative relocation method. It is assumed that a each cluster has at least one object and each object belongs to only one cluster. Objects may be relocated between clusters as the clusters are refined. Often these methods require that the number of clusters be specified apriori and this number usually does not change during the processing. K-Means and Expectation Maximization are commonly used methods in partitioning method.

Density-based methods are typically for each data point in a cluster, at least a minimum number of points must exist within a given radius. Density-based methods can deal with arbitrary shape clusters since the major requirement of such methods is that each cluster be a dense region of points surrounded by regions of low density. The most popular density based clustering method is DBSCAN, it is based on connecting points within certain distance threshold and its complexity is fairly low.
Grid-based methods works in the object space rather than the data is divided into a grid. Grid partitioning is based on characteristics of the data and such methods can deal with non-numeric data more easily. Grid-based methods are not affected by data ordering.

Model-based methods use an assumed model, perhaps based on a probability distribution. Essentially the algorithm tries to build clusters with a high level of similarity within them and a low level of similarity between them. Similarity measurement is based on the mean values and the algorithm tries to minimize the squared error function.

3.4 Discussion

In cluster analysis, one does not know what classes or clusters exist and the problem to be solved is to group the given data into meaning cluster. Work on clustering has been carried out in statistics for many years. Most of the algorithms developed are based on some concept of similarity or distance so that groups of objects that are similar or near to each other and dissimilar or far away from other objects may be put together in a cluster. The technique may appear simple but it can be rather challenging because to find objects that are dissimilar amongst a large collection of objects requires comparing each objects with every other object which may be prohibitively expensive for large sets. Also, the cost of computing distance between a group of objects grows as the number of attributes grows. Finally, it is easier to compute distances between objects with numeric
attributes but computing distance between categorical attributes are more difficult.

3.5 Summary

Data mining is primarily used today by companies with a strong consumer focus on retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.