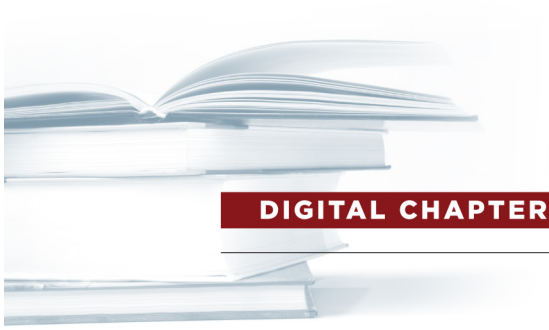

DIGITAL CHAPTERS

HARVARD BUSINESS PUBLISHING

BEP506

CHAPTER 9

Data Mining: Tools and Applications in Predictive Analytics

From *Business Analytics, Volume II: A Data-Driven Decision-Making Approach for Business*

By Amar Sahay
(A Business Expert Press Book)

CHAPTER 9

Data Mining: Tools and Applications in Predictive Analytics

Chapter Highlights

- Introduction to Data Mining
- Data Mining Defined
- Some Application Areas of Data Mining
- Machine Learning and Data Mining
- Data Mining and Its Origins and Areas It Interacts with
- Process of Data Mining and Knowledge Discovery in Databases (KDD)
- Data Mining Methodologies and Data Mining Tasks
  - Data Preparation or Data Preprocessing, and
    - Data cleaning
    - Data integration
    - Data selection
    - Data transformation
  - Data Mining
    - Pattern evaluation
    - Knowledge representation
- Data Mining Tasks
  - Descriptive Data Mining
  - Predictive Data Mining
- Difference between descriptive and predictive data mining?
- Additional Tools and Applications of Predictive Analytics: Data Mining Tasks

  - anomalies (or outlier) detection,
  - association learning,
  - classification,
  - clustering,
  - sequence, and
  - Time Series and forecasting
- Difference between Classification and Clustering
- Data Mining and Machine Learning
- Machine Learning Problems and Tasks
- Supervised and Unsupervised Machine Learning
- Artificial neural networks
- Deep Learning
- Summary

Introduction to Data Mining

Data mining involves exploring new patterns and relationships from the collected data—a part of predictive analytics that involves processing and analyzing huge amounts of data to extract useful information and patterns hidden in the data. The overall goal of data mining is knowledge discovery from the data. Data mining techniques are used to (i) extract previously unknown and potentially useful knowledge or patterns from massive amounts of data collected and stored, (ii) explore and analyze these large quantities of data to discover meaningful pattern, and (iii) transform data into an understandable structure for further use. The field of data mining is rapidly growing, and statistics plays a major role in it. Data mining is also known as knowledge discovery in databases (KDD), pattern analysis, information harvesting, business intelligence, analytics, etc. Besides statistics, data mining uses artificial intelligence (AI), machine learning, database systems, advanced statistical tools, and pattern recognition.

Successful companies use their data as an asset and use them for competitive advantage. These companies use business analytics and data mining tools as an organizational commitment to data-driven decision making. Business data mining combined with machine learning and AI

techniques helps businesses in making informed business decisions. It is also critical in automating and optimizing business processes.

Data Mining Defined

Data mining may be defined in the following ways:

- Data mining is Knowledge Discovery in Data Bases (KDD).
- Data mining is the extraction of interesting (nontrivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amounts of data (big data).
- Data mining can also be seen as the exploration and analysis by automatic or semi-automatic means of large quantities of data (big data) in order to discover meaningful patterns.

Why Data Mining?

In this age of technology, companies collect massive amounts of data automatically using different means. A large quantity of data is also collected using remote sensors and satellites. With the huge quantities of data collected today—usually referred to as big data—traditional techniques of data analysis are infeasible for processing the raw data. The data in its raw form have no meaning unless processed and analyzed. Among several tools and techniques available and currently emerging with the advancement of technology and computers, it is now possible to analyze big data using data mining, machine learning, and AI techniques.

The other reason to mine data is to discover the hidden patterns and relationship in the data. There is often hidden information in the data that is not readily apparent, and it is usually difficult to discover using traditional statistical tools. Sometimes it may take significant amount of time to discover useful information using traditional methods.

Data mining automatically processes massive amounts of data using specially designed software. A number of techniques, for example, classification and clustering, are used to analyze huge quantities of data. These provide useful information to the analysts and are critical in analyzing business, financial, or scientific data.

Some Application Areas of Data Mining

Data mining is one of the major tools of predictive analytics. In business, data mining is used to analyze business data. Business transaction data along with other customer and product-related data are continuously stored in the databases. The data mining software is used to analyze the vast amount of customer data to reveal hidden patterns, trends, and other customer behavior. Businesses use data mining to perform market analysis to identify and develop new products, analyze their supply chain, find the root cause of manufacturing problems, study the customer behavior for product promotion, improve sales by understanding the needs and requirements of their customer, prevent customer attrition, and acquire new customers. For example, Wal-Mart collects and processes over 20 million point-of-sale transactions every day. These data are stored in a centralized database and are analyzed using data mining software to understand and determine customer behavior, needs, and requirements. The data are analyzed to determine sales trends and forecasts, develop marketing strategies, and predict customer-buying habits [http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/].

The success with data mining and predictive modeling has encouraged many businesses to invest in data mining to achieve a competitive advantage. Data mining has been successfully applied in several areas of business and industry including customer service, banking, credit card fraud detection, risk management, sales and advertising, sales forecast, customer segmentation, and manufacturing.

Data mining is "the process of uncovering hidden trends and patterns that lead to predictive modeling using a combination of explicit knowledge base, sophisticated analytical skills and academic domain knowledge" (Luan, Jing, 2002). Data mining has been used successfully in science, engineering, business, and finance to extract previously unknown patterns in the databases containing massive amounts of data and to make predictions that are critical in decision making and improving the overall system performance.

In recent years, data mining combined with machine learning/AI is finding larger and wider applications in analyzing business data, thereby predicting future business outcomes. The reason for this is the growing interest in knowledge management and in moving from data to information and finally to knowledge discovery.

Machine Learning and Data Mining

Machine learning and data mining are similar in some ways and often overlap in applications. Machine learning is used for prediction, based on *known* properties learned from the training data, whereas data mining algorithms are used for discovery of (previously) *unknown* patterns. Data mining is concerned with KDD.

Data mining and its origins and areas it interacts with: Data mining has multiple objectives including knowledge discovery form large data sets and predicting the future business outcomes. Recently, machine learning, deep learning, and AI are finding more applications in data mining techniques. Figure 9.1 shows different areas that data mining interacts.

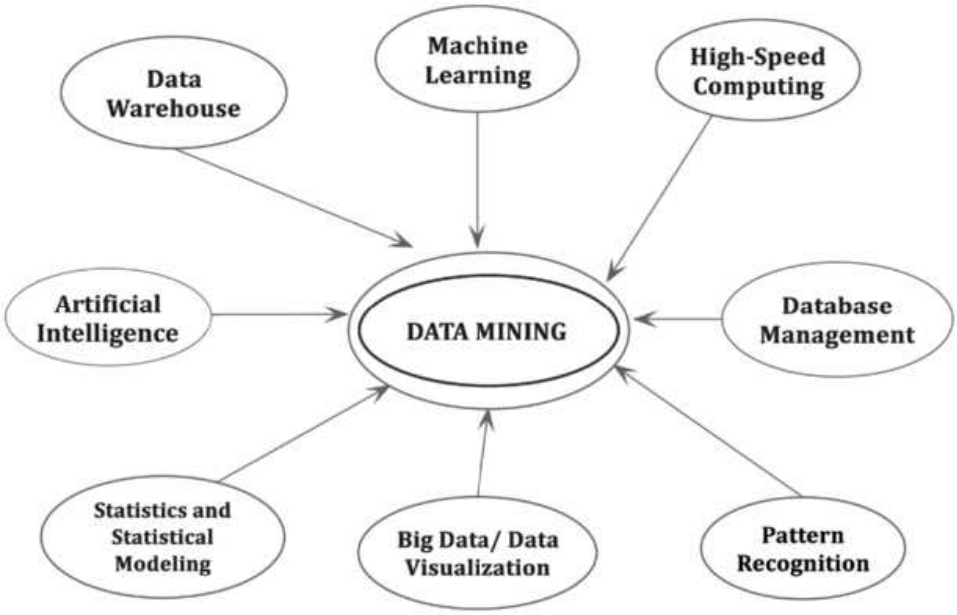

Figure 9.1 Data mining, its origin, and areas of interaction

Process of Data Mining and Knowledge Discovery in Databases

One of the key objectives of data mining is knowledge discovery. A data mining process involves multiple stages. KDD involves the activities leading up to actual data analysis and evaluation and deployment of the results. The activities in KDD are shown in Figure 9.2 and described below.

- *Data Collection:* The goal of this phase is to extract the data relevant to data mining analysis. The data should be stored in a database where data analysis will be applied.

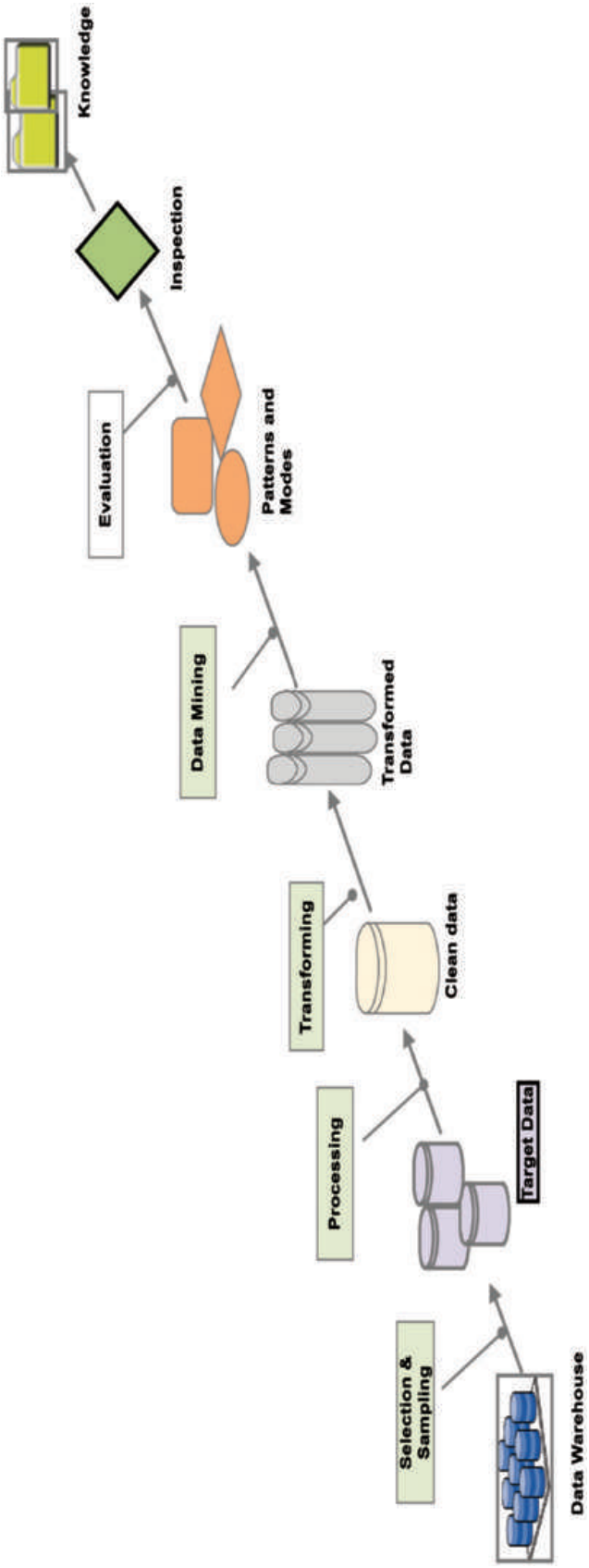

Figure 9.2 The knowledge discovery in data mining (KDD) process

- *Data Cleaning and Preprocessing*: This phase of KDD involves data cleansing and preparation of data to achieve the desired results. The purpose is to remove the noise and irrelevant information from the data set.
- *Data Transformation*: This phase is aimed at converting the data suitable for processing and obtaining valid results. An example would be transforming a Boolean column type to integer.
- *Data Mining*: The purpose of data mining phase is to analyze the data using appropriate algorithm to discover meaningful patterns and rules to produce predictive models. This is the most important phase of the KDD cycle.
- *Interpretation and Evaluation of Results:* This final phase involves selecting the valid models for making useful decisions. This phase also involves pattern evaluation. Not all of the patterns determined from the previous data mining phase may be meaningful. It is important to select the valid and meaningful patterns.

The KDD process depicted in Figure 9.2 involves a number of processes. The entire process can be divided into two broad categories, each involving a number of steps. These are:

1. **Data preparation or data preprocessing**
2. **Data mining**

Data preparation or preprocessing has several steps including:

(a) Data cleaning
(b) Data integration
(c) Data selection
(d) Data transformation

The above steps are necessary to prepare the data for further processing. The steps provide clean or processed data so that data mining tasks can be performed. The data mining tasks involve:

A) Data mining
B) Pattern evaluation
C) Knowledge representation

Figure 9.3 shows the data mining tasks in detail.

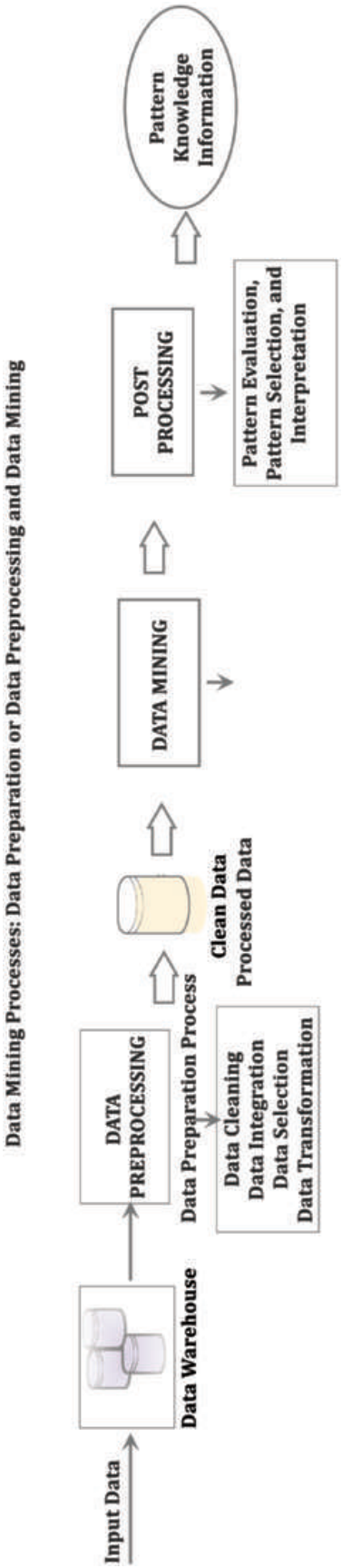

Figure 9.3 Data mining (KDD) process: data preprocessing and data mining tasks

To obtain useful information from the large volume of data is a complex process and requires several tasks. As stated above, it requires data preprocessing—the task that must be performed before data mining techniques can be applied to obtain useful information.

The data preprocessing or preparation involves several steps including data cleaning, data integration, data selection, and data transformation. Once the processed or cleaned data are available, data mining techniques are applied for pattern evaluation and obtaining knowledge and useful information. The data preprocessing steps are briefly discussed below followed by data mining tasks.

Data Cleaning

Data cleaning is the process of preparing and making data ready for further processing. The data collected are raw data and are usually unstructured, incomplete, noisy, have missing values, and are inconsistent. The data may also be missing attributes, for example, a huge number of customer data of a financial company may miss attributes like age and gender. Such data are incomplete with missing values. Data may also have outliers or extreme values. There may be recording errors, for example, a person's age may be wrongly recorded as 350 years.

The data available in data sources might be lacking attribute values. For example, we may have data that do not include attributes for the gender or age of the customers. These data are, of course, incomplete. Sometimes the data might contain errors or outliers. An example is an age attribute with value 200. It is obvious that the age value is wrong in this case. The data could also be inconsistent. For example, the name of an employee might be stored differently in different data tables or documents. Here, the data are inconsistent. If the data are not clean and structured, the data mining results would be neither reliable nor accurate.

Data cleaning involves a number of techniques including filling in the missing values manually, combined computer and human inspection, etc. The output of data cleaning process is adequately cleaned data ready for further processing.

Data Integration

Data integration is the process where data from different data sources are integrated into one. Data lie in different formats in different locations and could be stored in databases, text files, spreadsheets, documents, data cubes, Internet, and so on. Data integration is a really complex and tricky task because data from different sources may not match normally. For example, suppose table A contains an entity, named customer-id, whereas table B contains an entity named "number" instead of customer-id. In such cases, it is difficult to ensure whether both these entities refer to the same value. Metadata can be used effectively to reduce errors in the data integration process. Another issue faced is data redundancy where the same data may be available in different tables in the same database or are available in different data sources. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.

Data Selection

Data mining process uses large volumes of historical data for analysis. Sometimes, the data repository with integrated data may contain much more data than actually required. Before applying any data mining task or algorithm, the data of interest needs to be separated, selected, and stored from the available stored data. Data selection is the process of retrieving the relevant data for analysis from the database.

Data Transformation

Data transformation is the process of transforming and consolidating data into different forms that are suitable for mining. Data transformation normally involves normalization, aggregation, generalization, etc. For example, a data set available as "–7, 57, 200, 99, 68" can be transformed as "–0.07, 0.57, 2.00, 0.99, 0.68." This may be more desirable for data mining. After data integration, the available data are ready for data mining. Data transformation may involve smoothing, aggregation, generalization, or normalization of data.

Data Mining

Data mining is the core process that uses a number of complex methods to extract patterns from data. This purpose of data mining phase is to analyze the data using appropriate algorithms to discover meaningful patterns and rules to produce predictive models. This is the most important phase of KDD cycle.

Data mining process includes a number of tasks such as *association*, *classification*, *prediction*, *clustering*, *time series analysis*, *machine learning*, and *deep learning*. Table 9.1 outlines the data mining tasks.

Data Mining Methodologies: Data Mining Tasks

Data mining tasks can be broadly classified into *descriptive data mining* and *predictive data mining.*

There are a number of **data mining tasks** such as classification, prediction, time series **analysis**, association, clustering, and summarization. All these **tasks** are either *predictive* or *descriptive* **data mining tasks**. Figure 9.4 shows a broad view of data mining tasks.

Difference between Descriptive and Predictive Data Mining

Descriptive data mining tasks make use of collected data and data mining methodologies to look into the past behavior, relationships, and patterns to understand and explain what exactly happened in the past. Predictive analytics employs various predictive data mining and statistical models including regression, forecasting techniques, and other predictive models including simulation, machine learning, and AI to understand what could happen in the future and predict future business outcomes.

Predictive data mining uses models from the available data to predict future values of future business outcomes. An operations manager using simulation and queuing models to predict the future behavior of a call center to improve its performance can be considered as a predictive data mining task. Descriptive data mining tasks use graphical visual and numerical methods to find data describing patterns to learn about the

Table 9.1 Key data mining tasks

Data Mining	Brief Description	Application Areas
Data Mining and Tasks	**Data Mining** involves exploring new patterns and relationships from the collected data—a part of predictive analytics that involves processing and analyzing huge amounts of data to extract useful information and patterns hidden in the data. The overall goal of data mining is knowledge discovery from the data. Data mining techniques are used to (i) extract previously unknown and potential useful knowledge or patterns from massive amount of data collected and stored, (ii) explore and analyze these large quantities of data to discover meaningful pattern, and transform data into an understandable structure for further use. The field of data mining is rapidly growing and statistics plays a major role in it. Data mining is also known as KDD, pattern analysis, information harvesting, business intelligence, analytics, etc. Besides statistics, data mining uses AI, machine learning, database systems, advanced statistical tools, and pattern recognition. In this age of technology, companies collect massive amounts of data automatically using different means. A large quantity of data are also collected using remote sensors and satellites. With the huge quantities of data collected today—usually referred to as big data, traditional techniques of data analysis are infeasible for processing the raw data. The data in its raw form have no meaning unless processed and analyzed. Among several tools and techniques available and currently emerging with the advancement of technology and computers, it is now possible to analyze big data using data mining, machine learning, and AI techniques.	*Data mining* is one of the major tools of predictive analytics. In business, data mining is used to analyze business data. Business transaction data along with other customer and product-related data are continuously stored in the databases. The data mining software are used to analyze the vast amount of customer data to reveal hidden patterns, trends, and other customer behavior. Businesses use data mining to perform market analysis to identify and develop new products, analyze their supply chain, find the root cause of manufacturing problems, study the customer behavior for product promotion, improve sales by understanding the needs and requirements of their customer, prevent customer attrition, and acquire new customers. For example, Wal-Mart collects and processes over 20 million point-of-sale transactions every day. These data are stored in a centralized database and are analyzed using data mining software to understand and determine customer behavior, needs, and requirements. The data are analyzed to determine sales trends and forecasts, develop marketing strategies, and predict customer-buying habits [http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/]. The success with data mining and predictive modeling has encouraged many businesses to invest in data mining to achieve a competitive advantage. Data mining has been successfully applied in several areas of business and industry including customer service, banking, credit card fraud detection, risk management, sales and advertising, sales forecast, customer segmentation, and manufacturing. Data mining is "the process of uncovering hidden trends and patterns that lead to predictive modeling using a combination of explicit knowledge base, sophisticated analytical skills and academic domain knowledge" (Luan, Jing, 2002). Data mining has been used successfully in science, engineering, business, and finance to extract previously unknown patterns in the databases containing massive amounts of data and to make predictions that are critical in decision making and improving the overall system performance. In recent years, data mining combined with machine learning/AI is finding larger and wider applications in analyzing business data, thereby predicting future business outcomes. The reason for this is the growing interest in knowledge management and in moving from data to information and finally to knowledge discovery.

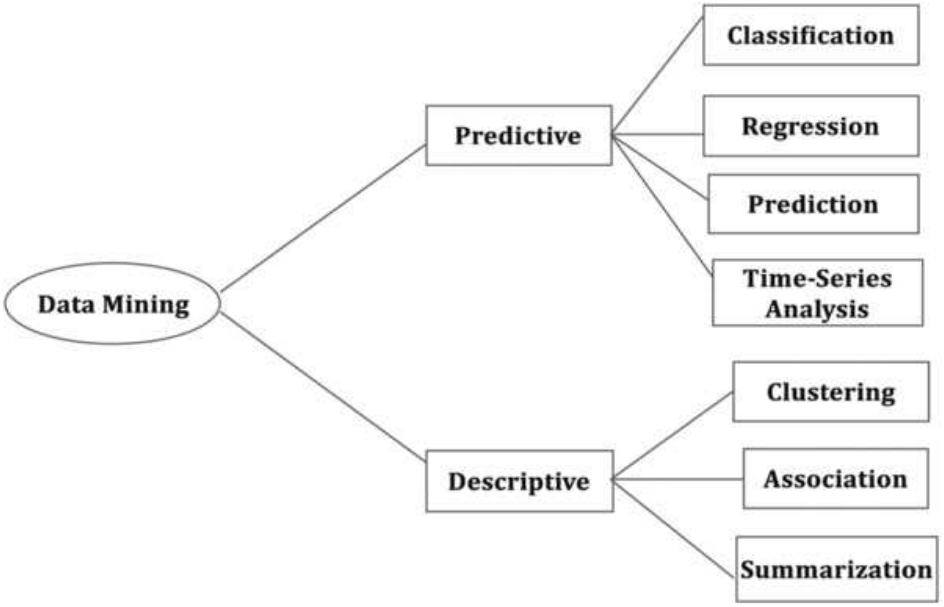

Figure 9.4 Data mining tasks

new information from the available data set that is not apparent otherwise. Businesses use a number of data visualization techniques including dashboards, heat maps, and a number of other graphical tools to study the current behavior of their businesses. These visual tools are simple but rather powerful tools in studying the current business behaviors and are used in building predictive analytics models.

Additional Tools and Applications of Predictive Analytics: Data Mining Tasks

Data mining is a software-based predictive analytics tool used to gain insights from massive amounts of data by extracting hidden patterns and relationships and using them to predict future business behaviors or outcomes. Data mining uses a number of methodologies including *anomalies (or outlier) detection, patterns*, *association learning, classification*, *clustering, sequence, and forecasting* to predict the probabilities and future business outcomes. We briefly describe them here. Figure 9.5 shows the broad categories of data mining methodologies.

Association learning is used to identify the items that may co-occur and the possible reasons for their co-occurrence. Classification and clustering techniques are used for association learning.

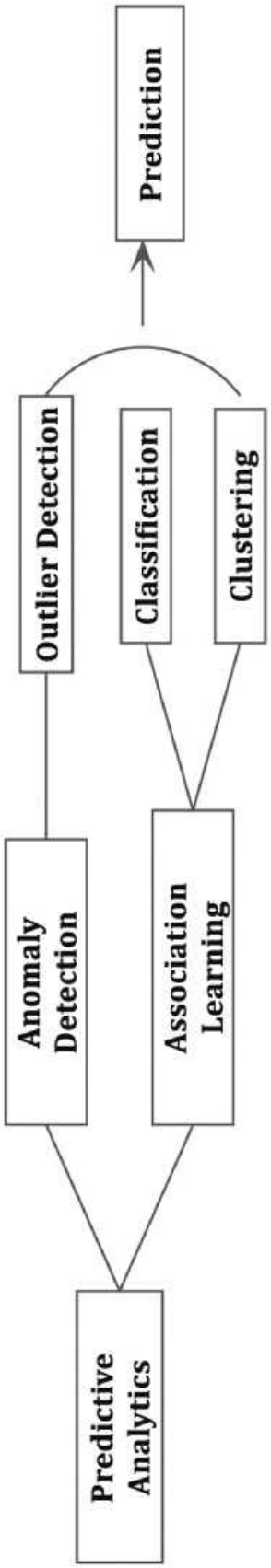

Figure 9.5 Data mining methodologies

Anomaly detection is also known as outlier detection and is used to identify specific events, or items, that do not conform to usual or expected pattern in the data. Typical example would be the detection of bank fraud.

Classification and clustering algorithms are used to divide the data into categories or classes. The purpose is to predict the probabilities of future outcomes based on the classification. Clustering and classification both divide the data into classes and, therefore, seem to be similar, but they are two different techniques. They are learning techniques used widely to obtain reliable information from a collection of raw data. Classification and clustering are widely used in data mining.

Classification

Classification is a process of assigning items to prespecified classes or categories. For example, a financial institution may study the potential borrowers to predict whether a group of new borrowers may be classified as having a high degree of risk. Spam filtering is another example of classification where the inputs are e-mail messages that are classified into classes as "spam" and "no spam."

Classification uses the algorithms to categorize the new data according to the observations of the training set. *Classification is a supervised learning technique* where a training set is used to find similarities in classes. This means that the input data are divided into two or more classes or categories and the learner creates a model that assigns inputs to one or more of these classes. This is typically done in a supervised way. The objects are classified on the basis of the training set of data.

The algorithm that implements the classification is known as the classifier. Some of the most commonly used classification algorithms are K-Nearest Neighbor algorithm and decision tree algorithms. These are widely used in data mining. An example of classification would be credit card processing. A credit card company may want to segment customer database based on similar buying patterns.

Clustering

Clustering technique is used to find natural groupings or clusters in a set of data without prespecifying a set of categories. It is unlike classification

where the objects are classified based on prespecified classes or categories. Thus, clustering is an *unsupervised learning technique* where a training set is not used. It uses statistical tools and concepts to create clusters with similar features within the data. Some examples of clustering are:

- Cluster houses in a town into neighborhoods based on similar features like houses with overall value of over million dollars.
- Marketing analyst may define distinct groups in their customer bases to develop targeted marketing programs.
- City planning may be interested in identifying groups of houses according to their house value, type, and location.
- In cellular manufacturing, the clustering algorithms are used to form clusters of similar machines and processes to form machine component cells.
- Scientists and geologists may study earthquake epicenters to identify clusters of fault lines with high probability of possible earthquake occurrences.

Cluster Analysis

Cluster analysis is the assignment of a set of observations into subsets (called *clusters*) so that observations within the same cluster are similar according to some prespecified criterion or criteria, while observations drawn from different clusters are dissimilar. Clustering techniques differ in application and make different assumptions on the structure of the data. In clustering, the clusters are commonly defined by some *similarity metric or similarity coefficient* and may be evaluated by *internal compactness* (similarity between members of the same cluster) and *separation* between different clusters. Other clustering methods are based on *estimated density* and *graph connectivity*. It is important to note that clustering is *unsupervised learning* and the commonly used method in statistical data analysis.

Difference between Clustering and Classification

Clustering is an unsupervised learning technique used to find groups or clusters of similar instances on the basis of features. The purpose of

clustering is a process of grouping similar objects to determine whether there is any relationship between them. **Classification** is a supervised learning technique used to find similarities in classification based on a training set. It uses algorithms to categorize the new data according to the observations in the training set. Figure 9.6 distinguishes between supervised and unsupervised learning techniques. Table 9.2 outlines the differences between classification and clustering.

Other Applications of Data Mining

Prediction

Predictive modeling is about predicting future business outcomes. Prediction techniques use a number of models based on the available data. These models can be simple to complex and may include simple and other regression models, simulation techniques, analysis of variance and design of experiments, and many others. For example, businesses are often interested in creating models to predict the success of their new products, or predict the sales and revenue from the advertisement expenditures. A number of examples and cases were cited in the previous chapters. When it is difficult to create a single model involving a number of variables, companies create simulation models to study and predict future business outcomes. Simulation models have been used successfully in studying the behavior of call centers, hospital emergency services, fast food drive-through, and many others. Heath care systems have used prediction techniques in the diagnosis and prediction of patient health. Prediction techniques also have applications in fraud detection.

Time Series Analysis

Time series analysis involves data collected over time. Time series is a sequence of historical events over time and studies the past performance to forecast or determine the future events where the next event is determined by one or more of the preceding events.

A number of models are used to analyze time series data. The forecasting chapter in this book discussed a number of time series patterns and

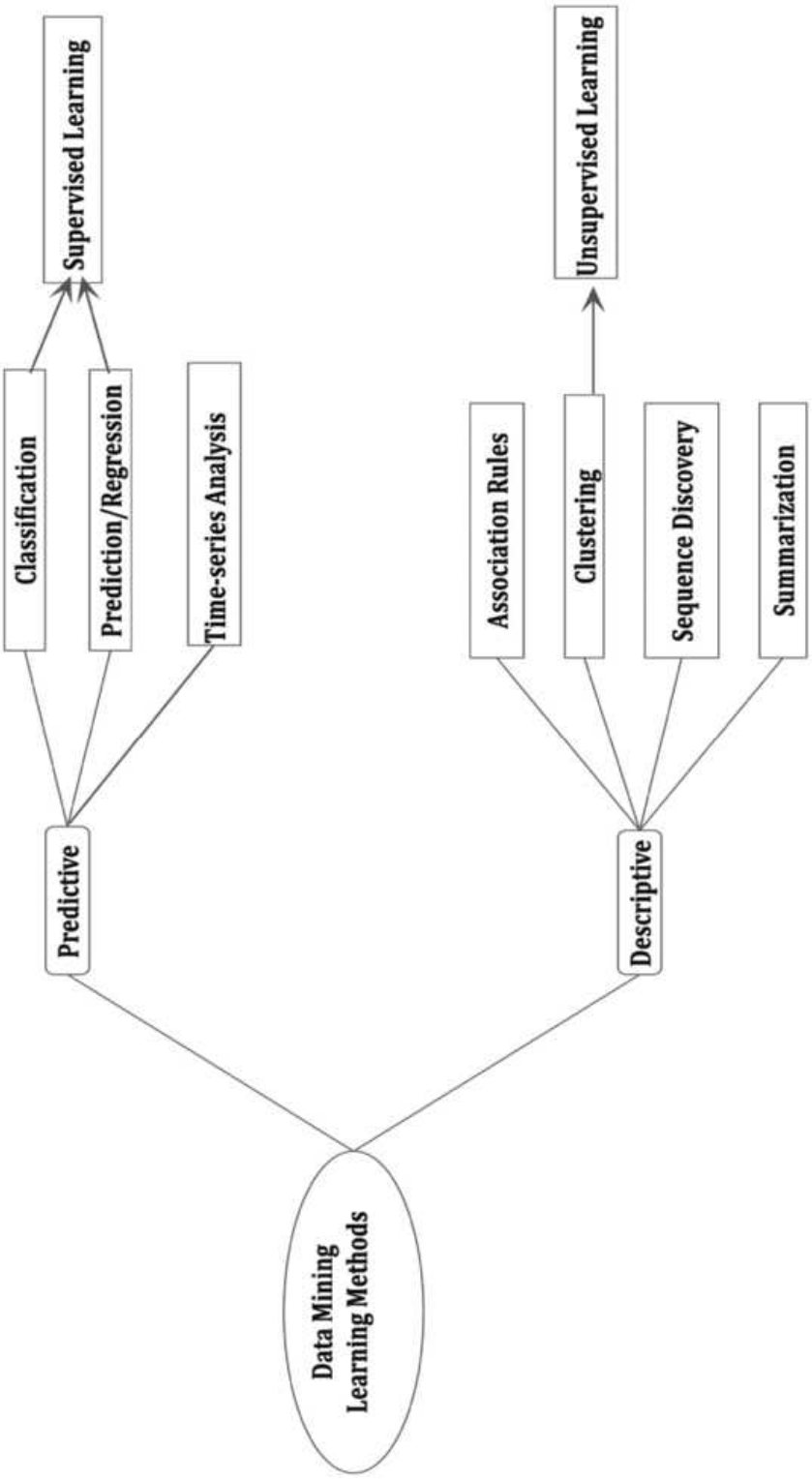

Figure 9.6 Supervised and Unsupervised Learning Techniques

Table 9.2 Difference between classification and clustering

Classification	Clustering
Classification is supervised learning technique where items are assigned to prespecified classes or categories. For example, a bank may study the potential borrowers to predict whether a group of new borrowers may be classified as having a high degree of risk.	Clustering is unsupervised technique used to find natural groupings or clusters in a set of data without prespecifying a set of categories. It is unlike classification where the objects are classified based on prespecified classes or categories.
Classification algorithm requires training data.	Clustering does not require training data.
Classification uses predefined instances.	Clustering does not assign predefined label to each and every group.
In classification the groups (or classes) are prespecified with each training data instance that belongs to a particular class.	In clustering the groups (or clusters) are based on the similarities of data instances to each other.
Classification algorithms are supposed to learn the association between the features of the instance and the class they belong to.	Unlike in classification, the groups are not known beforehand, making this an unsupervised task.
Example: An insurance company trying to assign customers into high-risk and low-risk categories.	Example: Dropbox, a movie rental company, may recommend a movie to customers because others who had made similar movie choices in the past have favorably rated that movie.

their analysis that includes a number of forecasting techniques. These forecasting techniques are simple to complex and use a number of techniques. The future product, process, and workforce requirement planning start with forecasts. Forecasting is a critical component of produce-to-stock firms and impacts future sales, revenue, demand, inventory, and workforce requirements. Time series analysis and forecasting techniques look into the data collected over time to extract useful patterns, trends, rules, and statistical models. The chapter on forecasting in this book outlines a number of time series analysis and forecasting models. Stock market prediction is an important application of time series analysis. These models have a number of applications in businesses including the stock market and predicting power need requirements during peak summer hours when the electricity requirement is highly variable and fluctuates rapidly in a short span of time.

Summarization

Data summarization is the process of reducing the vast amount of data into smaller sets that can provide a compact description of the data. Several techniques are used to summarize the data that help to facilitate the analysis of large amounts of data. One of the major objectives of data mining is the knowledge discovery from the vast amount of data (KDD). Data summarization techniques are applied to both structured and unstructured data. These techniques expedite the KDD tasks by reducing the size of the data. In this digital age, the collection and transfer of data is a fast process. Businesses now work with big data collected from various sources including media, networks, cloud storage, etc. The data may be collected from audio and video sources and are both structured and unstructured. Because of the volume and nature of data, it becomes necessary to summarize the data for further analysis. The summarization techniques result in smaller data set that facilitate further analysis. A number of tools and techniques are discussed in the literature for data summarization.

Data Mining and Machine Learning

A number of data mining and machine learning methods overlap in applications. As indicated earlier, data mining is concerned with knowledge discovery from the data (KDD) where the key task is the discovery of previously *unknown* knowledge. Machine learning, on the other hand, is evaluated with respect to known knowledge. According to Arthur Samuel, machine learning gives "computers the ability to learn without being explicitly programmed" [2, 3].

Machine learning methods use complex models and algorithms that are used to make predictions. The machine learning models allow the analysts to make predictions by learning from the trends, patterns, and relationships in the historical data. The algorithms are designed to learn iteratively from data without being programmed. In a way, machine learning automates model building.

Recently, machine learning algorithms are finding extensive applications in data-driven predictions and are a major decision-making tool. Some applications where machine learning has been used are e-mail filtering, cyber security, signal processing, and fraud detection. Machine

learning is employed in a range of computing tasks. Although machine learning models are being used in a number of applications, it has limitations in designing and programming explicit algorithms that are reproducible and have repeatability with good performance. With current research and the use of newer technology, the fields of machine learning and AI are becoming more promising.

It is important to note that data mining uses unsupervised methods that usually outperform the supervised methods used in machine learning. Data mining is the application of knowledge discovery from the data (KDD) where supervised methods cannot be used due to the unavailability of training data. Machine learning may also employ data mining methods as "unsupervised learning" to improve learner accuracy. The performance of the machine learning algorithms depends on its ability to *reproduce known knowledge*.

Machine Learning Problems and Tasks

Machine learning tasks have the following broad categories: **supervised learning**, **unsupervised learning**, and **reinforced learning**. The difference between supervised and unsupervised learning is explained in Table 9.3.

Other Applications of Machine Learning

Another application of machine learning is in the area of deep learning that is based on artificial neural networks. In this application, the learning tasks may contain more than one hidden layer or tasks, with a single hidden layer known as shallow learning.

Artificial Neural Networks

An artificial neural network learning algorithm, usually called "neural network," is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks. Computations are structured in terms of an interconnected group of artificial neurons, processing information using a connectionist approach to computation.

Table 9.3 Difference between supervised and unsupervised learning

Supervised Learning	Unsupervised Learning
Supervised learning uses a set of input variables ($x_1, x_2, \ldots, x_n$) and an output variable, $y(x)$. An algorithm of the form $y=f(x)$ is used to learn the mapping function relating the input to output. This mapping function or the model relating the input and output variable is used to predict the output variable. The goal is to obtain the mapping function that is so accurate that it can use even the new set of data That is, the model can be used to predict the output variable as the new data become available. The name supervised learning means that in this process the algorithm is trained to learn from the training data set where the learning process is supervised. In supervised learning process, the expected output or the answer is known. The algorithm is designed to make predictions iteratively from the training data and is corrected by the analyst as needed. The learning process stops when the algorithm provides the desired level of performance and accuracy. The most commonly used supervised problems are regression and classification problems. We discussed regression problems earlier. Time series predictions problems, random forest for classification and regression problems, and support vector machines for classification problems also fall in this category.	Unsupervised learning uses a set of input variables but no output variable. No labels are given to the learning algorithm. The algorithm is expected to find the structure in its input. The goals of unsupervised learning may be finding hidden pattern in the large data or feature learning. Thus, unsupervised learning can be a goal in itself or a means toward an end that is not based on general rule of teaching and training the algorithms. Unlike supervised learning algorithms, unsupervised algorithms are designed to devise and find the interesting structure in the data. The most commonly used unsupervised learning problems are clustering and association problems. In clustering, a set of inputs is to be divided into groups. Unlike classification, the groups are not known beforehand, making this typically an unsupervised task. **Association**: Association problems are used to discover rules that describe association such as people that buy X also tend to buy Y.

Modern neural networks are nonlinear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

Deep Learning

Falling hardware prices and the development of graphics processing units for personal use in the last few years have contributed to the development of the concept of deep learning, which consists of multiple hidden layers in an artificial neural network. This approach tries to model the way the

human brain processes light and sound into vision and hearing. Some successful applications of deep learning are computer vision and speech recognition.

Summary

This chapter introduced and provided an overview of the field of data mining. Today, vast amounts of data are collected by businesses. Data mining is an essential tool for extracting knowledge from massive amounts of data. The tools of data mining are used in extracting knowledge from the data—the process is known as KDD. The extracted information and knowledge are used in different models to predict future business outcomes. Besides the process of data mining and KDD, the chapter explained a number of data mining methodologies and tasks. We outlined and discussed several areas where data mining finds application. The essential tasks of data mining including data preparation or data preprocessing, knowledge representation, pattern evaluation, and descriptive and predictive data mining were discussed. The two broad areas of data mining are descriptive and predictive data mining. We discussed both of these areas and outlined the tools in each case with their objectives.

Data mining techniques are also classified as supervised and unsupervised learning. We discussed the tasks of data mining that fall under supervised and unsupervised learning. The key methodologies of data mining including anomalies (or outlier) detection, association learning, classification, clustering, sequence, prediction, and time series and forecasting along with their objectives were discussed. We also introduced the current and growing application areas of data mining. Data mining has wide applications in machine learning. The chapter introduced the relationship between data mining and machine learning. Different types of machine learning problems and tasks—supervised and unsupervised machine learning, applications of data mining in using artificial neural networks, and deep learning—were introduced.