19.1 Introduction

In today’s information age, the competitive advantage of an organization no longer depends on their information storage and processing capabilities (Carr 2003) but on their ability to analyze information and discover and manage valuable knowledge. Knowledge discovery and data mining, which is often referred to as knowledge discovery from data (KDD), is a multidisciplinary field bringing together several disciplines, including computer science, information systems, mathematics, and statistics (Han et al. 2011). The field emerged in response to the demand for knowledge management and decision support based on large volumes of data in business, medicine, sciences, engineering, and many other domains. Since the first KDD conference in 1989 (Fayyad et al. 1996b), KDD research has made significant progress and had a far-reaching impact on many aspects in our lives.

The primary goal of knowledge discovery and data mining is to identify “valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad et al. 1996c, p. 30). Huge amounts of data are gathered daily and stored in large-scale databases and data warehouses: sales transactions, flight schedules, patient medical records, news articles, financial reports, blogs, and instant messages, among many others. Valuable knowledge can be extracted from these data to generate important intelligence for various purposes. For example, credit card companies can analyze their customers’ credit card transaction records and profile each customer’s purchasing patterns (e.g., the average spending amount, the average monthly balance, and the time and locations of regular purchases). Such knowledge can, for example, be used to detect credit card fraud when some unusual, suspicious transactions occur (e.g., unusually high spending amount) (Bhattacharyya et al. 2011).
Many KDD techniques have been developed and applied to a wide range of domains (e.g., business, finance, and telecommunication). However, KDD still faces several fundamental challenges, such as:

- **Data volume**: Because of the rapid advances in information technology, the volumes of data that can be used for knowledge discovery purposes grow exponentially: companies collect transactional data in their daily business, research institutions and laboratories record and make available results from their scientific experiments, and healthcare providers and hospitals monitor patients’ health conditions and changes. Therefore, the KDD techniques must be sufficiently efficient and scalable to handle large volumes of data in a timely manner. Especially for real-time applications such as computer network intrusion detection, the KDD technology must be quick enough to detect and respond to any intrusion to prevent it from causing devastating damages to computer systems and networks (Ryan et al. 1998).

- **Data types and formats**: Data can be of many different types and formats: numerical, text, multimedia, temporal, spatial, and object-oriented. Numerical data (e.g., product price, order total, blood pressure, and weather temperature) are also called structured data and can be organized easily using relational databases. Other types of data are often called unstructured data because they may contain text and other data types that do not fit in some predefined data models. Because of the variety and complexity in data types and formats, the KDD techniques must be able to handle the specific characteristics of a data type. For example, text mining techniques must be able to recognize terms and phrases in free-text documents, parse sentences into syntactical components, and extract meaningful syntactical and semantic patterns from texts (Feldman and Sanger 2006).

- **Data dynamics**: In many applications, the data are constantly changing. For example, in traffic monitoring and control systems, the data collected using video cameras and sensors installed along streets are in the form of continuous sequences of digitalized signals. This poses a significant challenge on the storage and processing power of the systems.

- **Data quality**: The quality of data can substantially affect the quality and value of the patterns extracted from the data. When the data contain many errors, noise, and missing values, the extracted patterns based on such data may be inaccurate, incorrect, misleading, or useless. Therefore, the KDD techniques must be robust enough to handle low-quality data to ensure the quality of the results.

The KDD community has kept developing innovative techniques and methods to address these challenges. This chapter will provide an overview of the principles, techniques, applications, and practical impacts of KDD research. The next section will introduce the KDD process and the underlying principles for major data mining techniques. We will then review the applications and impacts of KDD technologies. The research issues and trends of future KDD research will be discussed. The last section will summarize this chapter.

## 19.2 KDD Principles and Techniques

As introduced earlier, KDD is a multidisciplinary field, which is related to several different research areas and fields—statistics, machine learning, artificial intelligence, neural networks, databases, information retrieval, and linguistics—which all have contributed to the development and evolution of the KDD field. There have been several views regarding the fundamental principles of KDD: data reduction view, data compression view, probabilistic view, microeconomic view, and inductive database view (Mannila 2000).

The data reduction view treats KDD as a process to reduce the data representation. In the data compression view, KDD is a way to compress given data into certain patterns or rules. The probabilistic view regards KDD as a statistic problem and the extracted knowledge as hypotheses about the data (Mannila 2000). The microeconomic view considers KDD as a problem-solving process intended to find useful
knowledge to support decision making (Kleinberg et al. 1998). The inductive database view assumes that the patterns are already embedded in the data and the task of KDD is to query the database and find the existing patterns (Imielinski and Mannila 1996).

These views and theories have together built the foundation of the KDD field. In reality, knowledge discovery is not a simple, one-step task but a complex process involving multiple stages and activities. With different contexts, goals, applications, and techniques, the underlying principles may not be the same.

19.2.1 KDD Process

Although many people use “knowledge discovery” and “data mining” interchangeably, the two terms actually refer to different things. More specifically, data mining is a core step in the multistage process of knowledge discovery (Han et al. 2011). Figure 19.1 presents an overview of a typical knowledge discovery process:

- **Cleaning and integration**: This preprocessing step includes such operations as removing noise, correcting errors, resolving inconsistencies, handling missing and unknown values, and consolidating data from multiple sources. For example, when mining sales data, abandoned or invalid transactions can be removed from the dataset. This step is critical to ensure that the preprocessed data meet the quality requirements for the later steps in the process.

- **Selection and transformation**: Sometimes the original dataset contains features (or fields and attributes in database terms) that are not relevant to a particular KDD application or task. By selecting only relevant features, the number of variables under investigation can be reduced and the efficiency of the KDD process can be improved. In addition, the original data may need to be transformed, summarized, or aggregated before they can be used for pattern extraction. For example, in social network mining applications, the data about social relationships between people may not be directly available but have to be inferred from other types of data such as group membership (Chau and Xu 2007) and paper coauthorship (Newman 2004a).

- **Data mining**: This is the most important step in the entire KDD process. In this step, effective and efficient algorithms, methods, and techniques are applied to the data to extract patterns, rules, or models. We will discuss this step in greater detail in the following sections.

![Figure 19.1 Overview of the KDD process. (Adapted from Han J. et al., Data Mining: Concepts and Techniques, 3 edn., Morgan Kaufmann, Amsterdam, the Netherlands, 2011.)](image-url)
- **Evaluation and presentation**: The patterns extracted from the data must be subject to evaluation (e.g., by domain experts) to determine its quality, validity, interestingness, and value. Patterns that are trivial or not useful for the particular application will not be considered knowledge. Finally, the valid, novel patterns extracted from the data must be presented in a way that they can be interpreted and understood relatively easily by users or decision makers because the ultimate goal of KDD is to provide “actionable” knowledge that supports decision making (Han et al. 2011). Various presentation and visualization techniques may be used in this step to accomplish this goal.

Because data mining is the key step in the KDD process, we will focus in the following sections on data mining techniques, applications, and impacts. Depending on the goals of KDD tasks, traditional data mining techniques can be categorized into four types: association mining, classification and prediction, clustering, and outlier analysis. We will discuss the basic principles for major algorithms and methods of these types in this section. We will also briefly introduce a few relatively new mining types including text mining, web mining, and network mining.

### 19.2.2 Association Mining

Association mining identifies frequently occurring relationships among items in a dataset (Agrawal et al. 1993). A widely used example of association mining is the market basket analysis, which is intended to find items that customers often purchase together. For example, a market basket analysis on customer grocery shopping transactions may reveal that when customers buy hot dogs, very likely they will also buy buns to go with the hot dogs. Such findings may help stores better design the floor layout and plan the shelf space (e.g., placing buns close to hot dogs) so as to encourage sales of both items. Market basket analysis has also been used in online stores. The online retailer, Amazon.com, has long been using customers’ “co-purchasing” information to promote sales of related items.

An association relationship is often represented as a rule, \(X \Rightarrow Y\), where \(X\) and \(Y\) are two itemsets, to indicate that itemset \(X\) and itemset \(Y\) are associated. Moreover, the rule implies that the occurrence of \(X\) will lead to the occurrence of \(Y\). Note that the reverse of this rule may not necessarily be true. In the aforementioned example about hot dogs and buns, the rule can be represented as \({\text{Hot dog}} \Rightarrow {\text{Bun}}\), which means that customers who buy hot dogs will also buy buns. An itemset consisting of \(k\) items is called a \(k\)-itemset. In this example, both \({\text{Hot dog}}\) and \({\text{Bun}}\) are 1-itemsets.

It can be imagined that a huge number of association rules can be mined out of a transactional dataset. However, not all the association rules are necessarily useful. Some associations may occur frequently, while others occur only a few times. To evaluate the “strength” of an association rule, two measures have been used to filter out weak associations: support and confidence. An itemset is frequent if it satisfies a predefined minimum support. Only associations that meet both the predefined minimum support and minimum confidence requirements will be considered strong, useful rules.

Support and confidence are defined in the form of probability:

\[
\text{Support} (X \Rightarrow Y) = P(X \cup Y),
\]

\[
\text{Confidence} (X \Rightarrow Y) = P(Y|X).
\]

For example, suppose 100 customer transactions are evaluated and it is found that 40 customers purchased hot dogs, 50 customers purchased buns, and 30 customers bought both hot dogs and buns. Using the two formulas, we get that Support = 30% and Confidence = 75%. Therefore, the complete rule will be represented as:

\[
{\text{Hot dog}} \Rightarrow {\text{Bun}} \ [\text{Support} = 30\%, \text{Confidence} = 75\%].
\]
This means that this association rule is supported by 30% of the transactions under consideration (i.e., hot dogs and buns co-occur in 30% of the transactions), and 75% of the customers who purchased hot dogs also purchased buns.

Researchers have proposed many methods and algorithms for finding association rules in databases. Among these methods, the Apriori algorithm (Agrawal and Srikant 1994) is the most widely used. The fundamental principle on which this algorithm is based is that there exists the prior knowledge of the property of a frequent itemset: all nonempty subsets of a frequent itemset must also be frequent. This property is used to search from data for all frequent itemsets, which have at least the minimum support. The frequent itemsets are then further examined to find strong association rules that also have minimum confidence.

The search for frequent itemsets is a progressive, iterative process in which the frequent \( k \)-itemsets are found using \((k - 1)\)-itemsets. The algorithm has two steps in each iteration: joining and pruning. The algorithm starts with the initial scan of the dataset and finds the set \( S_1 \), which contains all 1-itemsets that meet the minimum support requirement. \( S_1 \) will be joined with itself to find the candidate set, \( S_2 \), for frequent 2-itemsets. Based on the prior knowledge about the property of frequent itemsets, all nonfrequent subsets of \( S_2 \) (i.e., the subsets are not contained in \( S_1 \)) will be removed from \( S_2 \) in the pruning step. These two steps are repeated for each \( k \) until no frequent itemset can be found for \( k + 1 \).

After all the frequent itemsets are identified, association rules can be generated by finding the subsets of each itemset and calculate their confidences. Rules that do not satisfy the minimum confidence will be discarded. The output is a number of strong association rules.

The Apriori algorithm is intuitive and easy to implement. However, as the data volume increases, it becomes quite slow because it requires scanning the dataset repeatedly. Many variations of the algorithm have been proposed to increase the efficiency and scalability to handle large volumes of data, including the hash-based methods (Park et al. 1995), transaction reduction, partitioning (Savasere et al. 1995), and sampling (Toivonen 1996). An approach called frequent pattern growth (FP-Growth) that compresses the original dataset to avoid time-consuming candidate itemset generation has also been proposed (Han et al. 2000). All these new techniques have improved the performance of association rule mining in large databases.

### 19.2.3 Classification and Prediction

Classification is a type of data mining technology used to map records into one of several predefined categories based on attribute values of the records. For example, attributes in a patient’s medical records such as his/her blood sugar level, age, weight, and family medical history can be used to predict the patient’s risk level of having diabetes in the future (low vs. high) (Prather et al. 1997).

Classification is closely related to machine learning, pattern recognition, and statistics (Han et al. 2011). It usually consists of a training step and a testing step. During the training step, a classification algorithm reads records with their known category labels and generates a classification model (i.e., the classifier) based on the training data. In the testing step, the known category labels of the testing records are removed and the algorithm will predict the labels based on the learned model. These classifier-assigned labels are then compared against the known labels of the testing records to determine the performance of the classification. If the performance exceeds a predefined threshold, the algorithm, together with the learned model, can then be used to classify new data whose category labels are unknown. Because classification involves the training step in which the category labels are given, classification is also called supervised learning. Examples of classification applications include fraud detection (Bhattacharyya et al. 2011), computer and network intrusion detection (Ryan et al. 1998), corporate failure prediction (Eksi 2011), and image categorization (Zaiane et al. 2001).

Classifiers are used to generate discrete-valued category labels (e.g., low risk vs. high risk, success vs. failure). For applications dealing with continuous variables (e.g., yearly revenue), a traditional statistical method, regression analysis, can be used. Because regression analysis has long been studied and used
in statistics and many textbooks and literature that introduce and explain regression analysis can be found, we focus in this chapter on classifiers.

The model generated by a classifier can be represented as decision rules (if-then statements), decision trees, mathematical formulas, or neural networks. In a decision tree, for example, each internal node represents a test on the value of an attribute, each branch specifies an outcome from the test, and each leaf node corresponds to a category label. Figure 19.2 presents an example of a hypothetical decision tree for predicting the risk level (low vs. high) of diabetes for patients.

There have been many classification techniques including induction decision tree-based methods (Breiman et al. 1984; Quinlan 1986, 1993), naïve Bayesian classifiers (Weiss and Kulikowski 1991), Bayesian belief networks (Heckerman 1996), neural networks (Rumelhart and McClelland 1986), support vector machines (SVM) (Burges 1998; Vapnik 1998), and genetic algorithms (Goldberg 1989), among many others. Different classification methods are based on different principles. We briefly introduce a few methods as follows.

Decision tree induction methods such as ID3 (Quinlan 1986), C4.5 (Quinlan 1993), and classification and regression trees (CART) (Breiman et al. 1984) are based on the principle of information entropy in information theory (Shannon 1951). Information entropy generally measures the uncertainty and randomness in the “information content” of messages. The higher the uncertainty is, the higher the information entropy is. These classification methods take a divide-and-conquer strategy. The process begins with the entire set of attributes, and the attribute set is recursively partitioned into smaller subsets while the decision tree is built. The attribute that can reduce the information entropy the most will be selected as the splitting attribute for a partition. Each partition corresponds to an internal node of the decision tree.

The underlying principle of naïve Bayesian classification is Bayes’ theorem of posterior probability, along with the assumption that the effects of different attributes on the categories are independent of each other (Weiss and Kulikowski 1991). Based on this theorem, the posterior probability, \( P(C_i|R) \), for a record \( R \) belonging to a category \( C_i \), given the attribute values of \( R \), can be calculated using the posterior probability \( P(R|C_i) \) and prior probabilities \( P(R) \) and \( P(C_i) \):

\[
P(C_i|R) = \frac{P(R|C_i)P(C_i)}{P(R)}.
\]

The record \( R \) is assigned to a specific category \( C_i \) if \( P(C_i|R) > P(C_j|R) \), for all \( j \neq i \).

Neural network classifiers learn a classification model by simulating the ways in which neurons process information (Rumelhart and McClelland 1986). A neural network usually consists of several layers of nodes (neurons). The nodes at the first layer (the input layer) correspond to the record attributes and the nodes at the last layer (the output layer) correspond to the categories with associated labels. There may be one or more layers between the first and the last layer. The nodes at a layer are connected with nodes at the adjacent layers and each connection has an associated weight. During the training step,
the network receives the values of the attributes from the input layer. In order to generate the category labels that match the given labels at the output layer, the network dynamically adjusts the weights of the connections between layers of nodes. During the testing step, the learned weights are used to predict the labels of testing records. The biggest disadvantage of neural networks is that the learned classification model is hard to interpret because it is encoded in the weights of the connections.

Genetic algorithms incorporate the principles of natural evolution by imitating the genetic crossover and mutation operations during reproduction processes of living beings (Bäck 1996). The training stage involves multiple iterations. The initial set of decision rules, which are represented by strings of bits, are randomly generated. In each iteration, the rules are updated by swapping segments of the strings (crossover) and inverting randomly selected bits (mutation). The fitness of rules, which is usually measured by accuracy, is assessed in each iteration. The process stops when all rules’ fitness scores exceed a predefined threshold.

The performance of a classification method can be measured by efficiency and effectiveness (accuracy, precision, and recall) in general. Studies have shown that decision tree induction methods are quite accurate but slow, limiting their applicability to large datasets. The naïve Bayesian classifiers are comparable to decision tree induction methods in effectiveness. Both neural networks and genetic algorithms can achieve high effectiveness in some domains but often require long training time (Han et al. 2011).

19.2.4 Clustering

Clustering is used to group similar data items into clusters without the prior knowledge of their category labels (Jain et al. 1999). In this sense, clustering is a type of exploratory, unsupervised learning. The basic principle of clustering is to maximize within-group similarity while minimizing between-group similarity (Jain et al. 1999). Figure 19.3a presents an example of clustering analysis. Except for the two data points \( p \) and \( q \), all points fall into either one of the clusters represented by the dashed circles. Points within a circle are closer to each other than to points in the other circle. Clustering has been used in a variety of applications including image segmentation (Jain and Flynn 1996), gene clustering (Getz et al. 2000), and document categorization (Roussinov and Chen 1999).

Two types of clustering methods have been widely used: hierarchical and partitional. Hierarchical methods group data items into a series of nested clusters, forming a hierarchy of partitions. Partitional methods, in contrast, generate only one partition of the entire dataset.

Hierarchical methods are further categorized into agglomerative and divisive methods. Agglomerative methods such as variations of the single-link algorithm (Sibson 1973) and complete-link algorithm (Defays 1977) take a bottom-up approach and grow the clusters progressively. At the beginning, each data item is treated as a cluster. Using distance as the measure for the similarity (or dissimilarity) between data items, the two closest data items are merged into a new cluster and the distances between this new cluster and other clusters are updated. The algorithm then searches for another pair of clusters.

![FIGURE 19.3](image-url)
whose distance is the smallest among all between-cluster distances and merges them together. This process is repeated until all clusters are merged into a single cluster. Divisive methods reverse the clustering process of agglomerative methods by taking a top-down approach. The initial cluster includes all data items and the algorithms progressively split the clusters into smaller ones.

The pattern (i.e., nested partitions) generated by hierarchical methods is represented as a dendrogram in which each distance level corresponds to a merger of two clusters. The dendrogram can be cut at any level to generate a particular partition of the dataset. Figure 19.3b presents an example of a dendrogram.

Partitional clustering methods require the prior knowledge of the number of clusters, $k$. A partitional algorithm groups the data items into $k$ clusters by maximizing an objective function. The most commonly known partitional method is the $k$-means algorithm (Lloyd 1957). At the beginning, the algorithm randomly selects $k$ data items and treats them as the centers of the $k$ clusters. The remaining items are then assigned to their closest clusters. The new center of each cluster is recalculated after receiving new items. All the data items are then reassigned to clusters based on the updated centers. The process terminates if the cluster centers do not change. Because $k$-means algorithms often terminate at the local optimum, they may need to be run several times to find the global optimum.

Both hierarchical and partitional methods have their advantages and disadvantages. The number of clusters does not need to be prespecified for hierarchical methods. However, it is often difficult to determine the level at which the dendrogram should be cut to yield a meaningful partition. Partitional methods require the prior knowledge regarding the number of clusters but can generate the desired partition of the data (Jain and Dubes 1988). In terms of efficiency and scalability, partitional methods are faster and more scalable for large datasets than hierarchical methods are.

Other types of clustering techniques have also been developed. For example, density-based methods can be used to cluster spatial data (Ester et al. 1996). Self-organizing map (SOM) (Kohonen 1995), which is a neural network-based clustering technique, can be used to directly distribute multidimensional data items into regions (clusters) in a two-dimensional space (Chen et al. 1996). In network data mining applications, various clustering methods have been proposed for detecting communities, which can be viewed as densely knit clusters of nodes in networks. For instance, Girvan and Newman (2002) proposed a divisive algorithm that progressively removes links to break a connected network into communities. However, the algorithm is rather slow. A few alternative methods, such as the modularity-based algorithm (Clauset et al. 2004), have been proposed to provide better efficiency.

19.2.5 Outlier Analysis

Most data mining tasks, such as association mining, classification, and clustering, are intended to search for commonality among data and to seek patterns that occur frequently, regularly, or repeatedly. Outliers, which deviate substantially from the rest of the data items, are often treated as noise and removed from the data during the cleaning stage in a KDD process. For example, in Figure 19.3a, points $p$ and $q$ are outliers. They clearly do not belong to either cluster and may be ignored or removed in clustering applications. However, in some situations, outliers that represent unusual or abnormal behaviors must be identified. Credit card fraud detection, for instance, is looking for abnormal transactions that do not fall into the range of normal transactions or are significantly different from the regular purchasing patterns of credit card owners.

Unlike other data mining types such as classification and clustering, in which the development of the mining algorithms may not have to depend on the characteristics of specific applications, outlier detection is largely application dependent. For different applications in different domains, the definitions for outliers and the detection methods may be drastically different. As a result, no dominating methods have emerged to be used to detect outliers in most applications. Basically, depending on the goals and requirements of a specific application, classification and clustering methods may be used to detect outliers (Han et al. 2011).
For example, if the data sample contains records that are labeled by domain experts as outliers, classification methods can be used to detect future outliers that are similar to the known ones. For data without category labels, unsupervised methods such as clustering algorithms can be used to identify data items that are exceptionally dissimilar with the rest of the data.

In addition, statistical approaches have also been proposed for outlier detection in situations where the “normal behavior” of the data is known (Abraham and Box 1979; Agarwal 2006). Such approach is under the assumption that the normally behaved data are all generated by the same stochastic mechanism and outliers are those generated by different mechanisms form the rest of the data.

### 19.2.6 Other Data Mining Types

In addition to the four types of traditional data mining, there have been several new types of data mining including text mining, web mining, network mining, spatiotemporal data mining, stream data mining, and visual data mining, among many others. We briefly introduce the first three types here:

1. **Text mining**: Text mining has been employed in a wide range of applications such as text summarization (Fattah and Ren 2009), text categorization (Xue and Zhou 2009), named entity extraction (Schone et al. 2011), and opinion and sentiment analysis (Pang and Lee 2007). Text mining is closely related to computational linguistics, natural language processing, information retrieval, machine learning, and statistics (Feldman and Sanger 2006). Text mining requires a great deal of preprocessing in which the text (e.g., news articles) must be decomposed (parsed) into smaller syntactical units (e.g., terms and phrases). Sometimes, the text data may also need to be transformed into other types. For example, in some text mining applications, terms extracted from the documents in the entire corpus are treated as features and documents are treated as records. Thus, each document can be represented as a Boolean vector in which a true (or false) value for a feature indicates the presence (or absence) of the corresponding term in the document. During the mining stage, depending on the requirements of the specific applications, various data mining methods such as association mining, classification, and clustering may be used to find patterns in the text. For example, classification and clustering are frequently used in text categorization applications (Feldman and Sanger 2006).

2. **Web mining**: The web has provided a vast amount of publicly accessible information that could be useful for knowledge discovery. Web mining techniques can be categorized into three types (Kosala and Blockeel 2000): content mining, structure mining, and usage mining. Web content mining extracts useful information from the text, images, audios, and videos contained in web pages (Zamir and Etzioni 1998; Pollach et al. 2006). Web structure mining examines hyperlink structures of the web. It usually involves the analysis of in-links and out-links of a web page and has been used for search engine result ranking and other web applications (Brin and Page 1998; Kleinberg 1999). Web usage mining analyzes search logs or other activity logs to find patterns of users’ navigation behavior or to learn user profiles (Srivastava et al. 2000; Nasraoui et al. 2008; Malik and Rizvi 2011).

3. **Network mining**: Unlike traditional data mining that extracts patterns based on individual data items, network mining is used to mine patterns based on the relationships between data items. Network mining is grounded on three theoretical foundations: graph theory from mathematics and computer science (Bollobás 1998), social network analysis from sociology (Wasserman and Faust 1994), and topological analysis from statistical physics (Albert and Barabási 2002). Network mining is intended to find various structural patterns, such as structural positions that have special roles (e.g., leaders, gatekeepers) (Freeman 1979), communities (Clauset et al. 2004; Newman 2004b; Bagrow 2008), and future links (Liben-Nowell and Kleinberg 2007). The techniques include descriptive measures (e.g., centrality measures, network diameters), statistical approaches (e.g., degree distributions), and clustering analysis (e.g., modularity-based algorithms), among many others. Network mining is a young, fast-growing area and has great potential for various applications.
19.3 KDD Applications and Impacts

KDD technologies have made huge impact on every aspect of our lives whether we realize it or not. When we go to a retail store to shop, for example, we may be able to find related items easily on the shelves due to the store’s effective use of the results from market basket analyses. When we search for information using a search engine (e.g., Google), by entering just a few query terms, we can find the information we need on the web pages returned by the search engine based on web content and structure mining results. Many credit card companies and banks constantly watch for suspicious, fraudulent transactions to protect us from identity theft crimes. These benefits and the associated convenience have been provided by advanced technologies developed in the KDD research.

The KDD research, to a large extent, is application-driven. Many new methods and techniques are developed in response to the demand for knowledge discovery applications in various domains. In this section, we review a few examples of the domains in which KDD technologies have made significant impacts.

19.3.1 Finance

KDD technologies have long been used in the financial industry to support decision making in various applications including fraud detection, stock market forecasting, corporate distress and bankruptcy prediction, portfolio management, and financial crime (e.g., money laundering) investigation (Zhang and Zhou 2004; Kovalerchuk and Vityaev 2010). In these applications, multiple types of data mining methods often are combined and integrated to achieve high performance, quality, and interpretability of the results:

- **Fraud detection**: Fraud detection is used to identify fraudulent transactions, behaviors, and activities that deviate from the regular patterns of behavior. Phua et al. (2010) categorized frauds into four types: internal frauds (e.g., fraudulent financial reporting) (Kirkos et al. 2007; Glancy and Yadav 2011; Ravisankar et al. 2011), insurance frauds (Bentley 2000; Viaene et al. 2004), credit transaction frauds, and telecommunication frauds (Alves et al. 2006). In the financial context, the most frequently studied type is credit card fraud detection. Because fraudulent transactions are usually outliers that occur infrequently and irregularly, classification-based outlier analysis is often employed in such applications. For example, Paasch (2007) used neural networks together with genetic algorithms to detect fraudulent credit card transactions in a real dataset that contained 13 month’s worth of 50 million credit card transactions. Using the same dataset, Bhattacharyya et al. (2011) combined SVM and logistic regression models for fraud detection. Many other data mining methods have also been used in credit card fraud detection such as Bayesian classifiers (Panigrahi et al. 2009), hidden Markov models (Srivastava et al. 2008), and association rules (Sánchez et al. 2009).

- **Stock market forecasting**: Stock market forecasting attempts to predict the future prices or returns of stocks or other securities such as bonds. The input data usually are time-series data (e.g., stock prices) that are measured at successive time points at equal time intervals. There have been a large number of studies on stock market forecasting using neural network-based methods (Trippi and Desjeno 1992; Grudnitski and Osburn 1993; Wood and Dasgupta 1996; Wong and Selvi 1998). Neural networks have been also integrated with other methods to generate better prediction accuracy. For example, Hadavandi et al. (2010) proposed to integrate genetic fuzzy systems and artificial neural networks to build a stock price forecasting expert system. This approach has been tested and used in the prediction of the stock prices of companies in the IT and airline sectors.

- **Corporate distress and bankruptcy prediction**: Firms and corporations may face difficult economic and financial conditions. Financial distress sometimes may lead to bankruptcy. Accurate and timely prediction of financial distress and bankruptcy can help the important stakeholders of...
the firms take appropriate strategies to reduce or avoid possible financial losses. Chen and Du (2009) proposed to use neural networks to construct a distress model based on several financial ratios. Similarly, neural network-based techniques are used in other studies to predict financial distresses facing firms and corporations (Coats and Fant 1991–1992, 1993; Altman et al. 1994). The prediction of bankruptcy is also formulated as a classification problem using various features. In addition to the classic bankruptcy prediction model, the Z-score model, which is based on discriminant analysis* (Altman 1968), a variety of classification methods have been used including genetic algorithms (Shin 2002), neural networks (Fletcher and Goss 1993; Zhang et al. 1999), logistic regression and discriminant analysis (Back et al. 1996), and the hybrid approach combining several classifiers (Lee et al. 1996). Sung et al. (1999) distinguished crisis economic conditions from normal conditions under which a firm is facing the possibility of bankruptcy. The interpretive classification model they used identified different factors that should be used to predict bankruptcy under different conditions.

19.3.2 Business and Marketing

Business is a well-fit domain for knowledge discovery and data mining. Companies around the world are capturing large volumes of data about their customers, sales, transactions, goods transportation and delivery, and customer reviews of products and services. Using these data, KDD technologies have been playing a critical role in supporting various business functions such as customer relationship management (CRM), customer profiling, marketing, supply chain management, inventory control, demand forecasting, and product and service recommendation (Ghani and Soares 2006):

- **CRM and customer profiling**: CRM is aimed at helping businesses understand and profile the needs and preferences of individual customers and manage the interaction and relationships with their customers. As a business expands and the size of its customer base increases, it becomes increasingly difficult to learn customer profiles using manual approaches. Data mining technologies provide a great opportunity to serve the purposes of CRM in terms of customer identification, attraction, retention, and development (Ngai et al. 2009; Chopra et al. 2011). Customer profiles may include not only the demographics of the customers but also their purchasing history and patterns such as frequency and size of purchases and customer lifetime values (Shaw et al. 2001). Association mining-based market basket analysis, as reviewed in the previous section, can provide a collection of frequent itemsets and association rules that represent such patterns. In addition, neural networks, decision trees, and clustering are also widely used in CRM (Ngai et al. 2009). For example, neural networks and genetic algorithms were combined for customer targeting (Kim and Street 2004), a Bayesian network classifier was used to model the changes in the customer lifecycles (Baesens et al. 2004), multiple classifiers were integrated to predict customer purchasing patterns (Kim et al. 2003), and k-means algorithms were used to identify groups of customers motivated by the importance of stores in shopping centers (Dennis et al. 2001).

- **Direct marketing and viral marketing**: As today’s markets become more competitive and fast changing, mass marketing, which relies on mass media such as newspaper, television, and radio to advertise products and services to the general public, becomes less effective. In contrast, direct marketing selects and targets individual customers that are predicted to be more likely to respond to promotions and marketing campaigns (Ling and Li 1998). Direct marketing is closely related to CRM as it is also based on the concept of customer differentiation and profiling. In addition, due to advanced web technologies, online social networks have become important channels for information disseminations and diffusion. As a result, network mining techniques have been used in

---

* Discriminant analysis is a type of statistical analysis used to express a categorical-dependent variable using a linear combination of a set of independent variables.

© 2014 by Taylor & Francis Group, LLC
viral marketing programs (Domingos and Richardson 2001; Watts and Dodds 2007). By identifying and targeting the most influential customers in a company’s online customer networks, the company may be able to spread the information about promotions of new products and services quickly through these customers, achieving high effectiveness with low costs. In addition, by analyzing the community structure in customer networks, companies may also obtain insights into customers’ opinions and attitudes toward their products and services (Chau et al. 2009).

- **E-commerce:** KDD technologies have brought great opportunities for e-commerce. In addition to the patterns that can be mined using traditional data mining techniques such as association rules, classification, and clustering analysis, new patterns that are unique to the ways that customer use the websites can be extracted using web usage mining (Srivastava et al. 2000). For example, the paths along which customers navigate among web pages can help design personalized websites so as to provide better online shopping convenience and experience and to increase the chances of customer retention. In addition, web usage patterns can also help generate better recommendations of products using content-based approaches (i.e., recommending items similar to the items purchased in the past) or collaborative approaches (i.e., recommending items that are purchased by other customers).

### 19.3.3 Healthcare and Biomedicine

KDD has become increasingly popular in the domain of healthcare and biomedicine because of the availability of large-size databases and data warehouses for clinical records, gene sequences, and medical literature. However, the “information overload” problem also poses challenges to healthcare professionals (e.g., doctors, nurses, hospital officials, insurers, and pharmaceutical companies) and medical researchers. It is nearly impossible for anyone to process, analyze, and digest such large volumes of healthcare data to make correct, timely decision using manual approaches. Data mining technologies enable healthcare professionals and researchers to leverage healthcare data for the purposes of diagnosing diseases, developing effective treatments, drugs, nurse care plans, and novel hypotheses (Koh and Tan 2005), and detecting infectious disease outbreaks (Chapman et al. 2004; Zeng et al. 2005):

- **Disease diagnosis:** Data mining techniques, especially classification methods, have been used to help with the diagnosis and treatments of various diseases. For example, Breault et al. (2002) used the decision tree algorithm, CART, to classify over 30,000 records of diabetic patients in a large medical data warehouse and found that younger age was the most important variable associated with bad glycemic control. Similarly, Kaur and Wasan (2006) applied classification methods to the early diagnosis of type-I diabetes in children. Classification algorithms have also been used in the diagnosis of cancers and tumors (Ball et al. 2002; Delen et al. 2005), skin lesions (Dreiseitl et al. 2001), preterm birth (Prather et al. 1997), and neurological disorders (Xu et al. 2011).

- **Gene expression and microarray data analysis:** The advances in new biomedical technologies for genome sequencing and protein identification have brought tremendous opportunities for mining gene expression* and microarray data† (Getz et al. 2000; Sturn et al. 2002) to help develop personalized drugs and treatment for patients (Debouck and Goodfellow 1999), or find new biological solutions to medical problems. Clustering analysis is widely used on microarray data to identify tissue similarity or groups of genes sharing similar expression patterns (Matsumoto et al. 2003). Examples of these applications include using hierarchical clustering for identifying genome-wide expression patterns (Eisen et al. 1998), k-means for clustering complementary CDNA fingerprinting data (Herwig et al. 1999), hierarchical self-organizing methods for discovering cancer classes

---

* Gene expression is a process in which the genetic information encoded in a DNA stretch is used to synthesize a functional protein.
† A microarray is a two-dimensional array on a glass slide or silicon thin film that records the activities and interactions of a large amount of biological entities (e.g., genes, proteins, and antibodies) using high-throughput screening methods.
and identifying marker genes (Hsu et al. 2003), and neural network-based methods for finding gene groups that are consistent with those defined by functional categories and common regulatory motifs (Sawa and Ohno-Machado 2003).

- **Biomedical text analysis**: Text mining techniques have been used in recent years to analyze biomedical texts in the form of research articles, laboratory reports, clinical documents, and patient records to discover relationships between biological entities (e.g., genes, proteins, drugs, and diseases) that are potentially useful but previously unnoticed. For example, Kazama et al. (2002) used SVM to extract biological entities in the GENIA corpus. Tanabe and Wilbur (2002) used the part-of-speech tagging and a Bayesian model to extract genes and proteins in biomedical text. Biological relationships (e.g., gene regulations, metabolic pathways, and protein interactions) are identified using various text mining and data mining methods (Song and Chen 2009; Steele et al. 2009).

### 19.3.4 Security and Intelligence

In response to the tragic events of 9/11 and the following series of terrorist attacks around the world, there has been a pressing demand for advanced technologies for helping intelligence communities and law enforcement agencies combat terrorism and other crimes. A new interdisciplinary field called *Intelligence and Security Informatics (ISI)* (Chen et al. 2003) has emerged to leverage technologies and knowledge from different disciplines to assist crime investigation and help detect and prevent terrorist attacks. KDD is a core component in the ISI technology collection (Chen et al. 2003):

- **Crime investigation**: A number of data mining techniques have been employed in crime investigation applications. For example, classification and clustering techniques have been applied to detect various types of crimes including computer crimes. Adderley and Musgrove (2001) and Kangas et al. (2003) employed the SOM approach to cluster crime incidents based on a number of offender attributes (e.g., offender motives and racial preferences) to identify serial murders and sexual offenders. Ryan et al. (1998) developed a neural network-based intrusion detection method to identify unusual user activities based on the patterns of users’ past system command usage. In addition, spatial pattern analysis and geographical profiling of crimes play important roles in solving crimes (Rossmo 1995). Brown (1998) proposed a $k$-means and nearest neighbor approach to cluster spatial data of crimes to find “hot spot” areas in a city (Murray and Estivill-Castro 1998). Koperski and Han (1995) used spatial association-rule mining to extract cause–effect relations among geographically referenced crime data to identify environmental factors that attract crimes (Estivill-Castro and Lee 2001).

- **Counterterrorism**: Terrorism and terrorist activities substantially threaten national security and have considerable negative impact on the society. However, because terrorist groups are covert organizations, the data about their members and activities are extremely difficult to gather. This has made knowledge discovery for counterterrorism a rather challenging task. Nonetheless, network mining techniques have been used to analyze and extract patterns from terrorist networks such as the Global Salafi Jihad networks (Sageman 2004) that were already destroyed by intelligence agencies. The structural characteristics and evolitional dynamics have been discovered using techniques such as community detection and topological analysis (Krebs 2001; Xu and Chen 2008; Xu et al. 2009). Although these findings were based on historical data, they enhanced our understanding of the organization, behavior, and characteristics of terrorist groups, and such understanding will help develop better strategies to prevent and fight future terrorist attacks (Krebs 2001; van Meter 2001; Dombroski and Carley 2002). In addition, because the web has been widely used by terrorist groups as a communication medium (Burris et al. 2000; Gerstenfeld et al. 2003), web mining techniques have been employed to discover patterns in the contents and structures of terrorist groups’ websites (Chen 2012).
In addition to the application domains we have reviewed, KDD technologies have also made great impacts in a number of other domains such as sciences (Fayyad et al. 1996a), engineering (Grossman et al. 2001), and manufacturing (Harding et al. 2006), among many others.

### 19.4 Research Issues

Research on knowledge discovery and data mining has made considerable progress in recent years (Han et al. 2011). However, many issues still remain challenging. In addition, with the rapid innovation and development of technologies in database management, data collection, computer memory, and computational power, new trends and directions keep emerging in the field of KDD:

- **Diversity of data types**: The diversity of data types will continue to be both a challenge and an opportunity for KDD research. On one hand, more new data types have become available for mining purposes including web data (e.g., hyperlinks, web pages, and web usage data), multimedia data (e.g., videos, autos, and images), sequence data (e.g., time series, symbolic sequences, DNA sequences, and protein sequences), network data (e.g., social networks, computer networks, and information networks), spatiotemporal data, and data streams (e.g., sensor data). These new data types may pose unique challenges on traditional data management and mining technologies. For example, sensor data usually arrive rapidly and continuously, and traditional database management systems are not designed for loading such continuous streams. In addition, many mining algorithms can only access data via a few linear scans and generally cannot perform repeated random access on data streams (Yates and Xu 2010). On the other hand, these new data types also expose the KDD research to new areas and application domains. For example, image mining and spatiotemporal data mining can be used in mobile phones, global positioning system (GPS) devices, weather services, and satellites (Han et al. 2011). Mining on DNA sequence and microarray data may provide new insights and directions to the development of new drugs and medicine (Debouck and Goodfellow 1999).

- **High-performance mining techniques**: As the number of KDD applications continues to grow, the demand for developing new mining techniques with high performance remains strong. In addition to the improvement of general-purpose techniques such as classification and clustering, the development of application-specific algorithms is becoming a trend. Many applications have their unique goals, constraints, and characteristics, which require that the mining techniques be tailored to address the unique problems in the specific applications. For example, in network mining applications, traditional clustering algorithms based on similarity or distance measures could not be used to uncover community structure in unweighted networks since all links are of the same length (distance). As a result, non-distance-based methods need to be developed to address this problem (Clauset et al. 2004). Moreover, as the volumes of data continue to grow, the mining algorithms and methods must be scalable enough to mine large-scale datasets effectively and efficiently.

- **User interaction and domain knowledge**: The process of KDD should be made interactive so that users and analysts can set parameters, select techniques, specify constraints and conditions, and incorporate their domain knowledge to guide the mining process (Fayyad et al. 1996c). However, most existing data mining technologies are not interactive. To many users, the mining process is a black box over which they do not have any control. Another issue is concerning the presentation of KDD results. That is, the patterns generated should be presented in a way that users and analysts, who may not understand the technology, could easily interpret and comprehend them. This requires the development of effective knowledge presentation and visualization techniques to help users and analysts understand the discovered patterns and make correct, timely decisions based on the knowledge.

- **Evaluation and validation of mining results**: Since knowledge discovery is a process to extract novel patterns that are previously unknown, many data mining technologies are exploratory in nature. Unlike statistical results, which can be evaluated using standard tests, results generated...
by data mining algorithms may not have standard tests available for assessing their “significance.” The patterns yielded by hierarchical clustering algorithms, for example, are dendrograms, which can be cut at any similarity (or distance) level to produce a partition of the data. Except for domain experts’ subjective evaluation, there is no “gold” standard for evaluating the quality or utility of a particular partition (Jain et al. 1999). Therefore, while developing new techniques and methods is important, it is also critical to develop appropriate evaluation and validation methods, in addition to existing methods and metrics (e.g., multiple-fold cross validation, accuracy, precision and recall), to ensure the quality of the extracted patterns.

- **Privacy concerns:** A recent trend in KDD research is on the privacy-preserving data mining due to the increasing concerns with data privacy and security. Especially in domains such as healthcare, medicine, and security and intelligence, confidential data (e.g., patients’ medical history and criminals’ identities) must be handled with great caution. There are established rules, regulations, and laws to protect privacy and confidentiality of data. The 1996 Health Insurance Portability and Accountability Act (HIPAA), for example, set the rules for handling patient records in electronic forms. Violations of such regulations may lead to serious social, ethical, or even legal consequences. Privacy-preserving data mining, therefore, is designed to extract knowledge from data without threatening the privacy and confidentiality of records by using various transformation approaches such as data de-identification (Cios and Moore 2002), perturbation (Li and Sarkar 2006), or reconstruction (Zhu et al. 2009).

### 19.5 Summary

KDD is a fast-growing, interdisciplinary field. More than two decades of research and practice have resulted in significant progress in this field. This chapter introduces the principles and major techniques of data mining techniques including association mining, classification, clustering, and outlier analysis. Relatively new data mining topics such as text mining, web mining, and network mining are also discussed. Examples of application domains in which KDD has made tremendous impact include finance, business, marketing, healthcare, biomedicine, security, and intelligence. Although various issues remain to be challenging, the KDD field will keep growing and thriving with emerging new trends and directions.

### Further Information


### References


Domingos, P. and Richardson, M. 2001. Mining the network value of customers. Paper read at the *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 26–29, San Francisco, CA.


Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. Paper read at 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD’00), May 2000, Dallas, TX.


Koperski, K. and Han, J. 1995. Discovery of spatial association rules in geographic information databases. Paper read at the *4th International Symposium on Large Spatial Databases*, August 6–9, Portland, ME.


Paasch, C. 2007. *Credit Card Fraud Detection Using Artificial Neural Networks Tuned by Genetic Algorithms*. Hong Kong, China: Hong Kong University of Science and Technology.


Knowledge Discovery and Data Mining 19-21


Toivonen, H. 1996. Sampling large databases for association rules. Paper read at International Conference on Very Large Data Bases (VLDB’96), September 1996, Bombay, India.


