values—1 or 0). Because this process may increase the number of variables, one should be cautious about the effect of such representations, especially for the categorical variables that have large numbers of unique values.

Similarly, some data mining methods, such as ID3 (a classic decision tree algorithm) and rough sets (a relatively new rule induction algorithm), need all of the variables represented as categorically valued variables. Early versions of these methods required the user to discretize numeric variables into categorical representations before they could be processed by the algorithm. The good news is that most implementations of these algorithms in widely available software tools accept a mix of numeric and nominal variables and internally make the necessary conversions before processing the data.

APPLICATION CASE 2
Law Enforcement Organizations Use Data Mining to Better Fight Crime

In the midst of these unfavorable economic conditions, police departments all over the world are facing difficult times in fighting crimes with continually shrinking resources along with fewer leads, a larger number of cases, and increasingly more complicated crimes. At a police department in the United Kingdom, investigators find that these challenges limit the cases they can tackle. A high volume of cases without definite leads—such as house burglaries and vehicle thefts that lack clear evidence—are often filed away until new evidence is found. Therefore, the challenge for the police department was to determine a way to quickly and easily find patterns and trends in unsolved criminal cases.

Each electronic case file at the police department contains physical descriptions of the thieves as well as their modus operandi (MO). Whereas many cases lacking evidence were previously filed away, the department is now re-examining them and doing it more quickly than ever before. In PASW Modeler (formerly Clementine), the data modeler uses two Kohonen neural network models to cluster similar physical descriptions and MOs and then combines clusters to see whether groups of similar physical descriptions coincide with groups of similar MOs. If a good match is found and the perpetrators are known for one or more of the offenses, it is possible that the unsolved cases were committed by the same individuals.

The analytical team further investigates the clusters, using statistical methods to verify the similarities’ importance. If clusters indicate that the same criminal may be at work, the department is likely to reopen and investigate the other crimes. Or, if the criminal is unknown but a large cluster indicates the same offender, the leads from these cases can be combined and the case reprioritized. The department is also investigating the behavior of prolific repeat offenders with the goal of identifying crimes that seem to fit their behavioral pattern. The department hopes that the PASW Modeler will enable it to reopen old cases and make connections with known perpetrators.

Another police department in the United States is facing similar challenges: lack of sufficient resources coupled with an increasing number of criminal cases. In order to produce sustainable solutions to a wide range of crime and community disorders, the department is pursuing a community-oriented policing philosophy, which is a holistic approach requiring collaborative partnerships between citizens and community agencies and careful analysis of information surrounding the criminal cases. The underlying process aims to find long-term solutions to crimes by identifying their root causes, educating the community on the extent of the problems, and then working with the community to develop collaborative solutions that effectively address these causes. The main challenge was to convince the community that their involvement is necessary for any solution to be effective.

Using PASW statistical analysis and a data mining software tool, the police department conducted extensive data analysis to discover the variables strongly associated with the criminal cases, as well as assess citizen satisfaction with community policing. The results of this analysis presented compelling
evidence that community involvement coupled with intelligent data analysis are necessary ingredients in developing effective long-term solutions in the midst of economic difficulties.

Police departments around the globe are enhancing their crime-fighting techniques with innovative twenty-first-century approaches of applying data mining technology to prevent criminal activity. Success stories can be found on Web sites of major data mining tool and solution providers (e.g., SPSS, SAS, StatSoft, Salford Systems), as well as the major consultancy companies.


How Data Mining Works

Using existing and relevant data, data mining builds models to identify patterns among the attributes presented in the dataset. Models are the mathematical representations (simple linear relationships and/or complex highly nonlinear relationships) that identify the patterns among the attributes of the objects (e.g., customers) described in the dataset. Some of these patterns are explanatory (explaining the interrelationships and affinities among the attributes), whereas others are predictive (foretelling future values of certain attributes). In general, data mining seeks to identify four major types of patterns:

1. **Associations** find the commonly co-occurring groupings of things, such as beer and diapers going together in market-basket analysis.
2. **Predictions** tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the Super Bowl or forecasting the absolute temperature of a particular day.
3. **Clusters** identify natural groupings of things based on their known characteristics, such as assigning customers in different segments based on their demographics and past purchase behaviors.
4. **Sequential relationships** discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.

These types of patterns have been manually extracted from data by humans for centuries, but the increasing volume of data in modern times has created a need for more automatic approaches. As datasets have grown in size and complexity, direct manual data analysis has increasingly been augmented with indirect, automatic data processing tools that use sophisticated methodologies, methods, and algorithms. The manifestation of such evolution of automated and semiautomated means of processing large datasets is now commonly referred to as **data mining**.

Generally speaking, data mining tasks can be classified into three main categories: prediction, association, and clustering. Based on the way in which the patterns are extracted from the historical data, the learning algorithms of data mining methods can be classified as either supervised or unsupervised. With supervised learning algorithms, the training data includes both the descriptive attributes (i.e., independent variables or decision variables) as well as the class attribute (i.e., output variable or result variable). In contrast, with unsupervised learning the training data includes only the descriptive attributes. Figure 4 shows a simple taxonomy for data mining tasks, along with the learning methods, and popular algorithms for each of the data mining tasks.
**PREDICTION** Prediction is commonly referred to as the act of telling about the future. It differs from simple guessing by taking into account the experiences, opinions, and other relevant information in conducting the task of foretelling. A term that is commonly associated with prediction is forecasting. Even though many believe that these two terms are synonymous, there is a subtle but critical difference between the two. Whereas prediction is largely experience and opinion based, forecasting is data and model based. That is, in order of increasing reliability, one might list the relevant terms as guessing, predicting, and forecasting, respectively. In data mining terminology, prediction and forecasting are used synonymously, and the term prediction is used as the common representation of the act. Depending on the nature of what is being predicted, prediction can be named more specifically as classification (where the predicted thing, such as tomorrow’s forecast, is a class label such as “rainy” or “sunny”) or regression (where the predicted thing, such as tomorrow’s temperature, is a real number, such as “65°F”).

**CLASSIFICATION** Classification, or supervised induction, is perhaps the most common of all data mining tasks. The objective of classification is to analyze the historical data stored in a database and automatically generate a model that can predict future behavior. This induced model consists of generalizations over the records of a training dataset, which help distinguish predefined classes. The hope is that the model can then be used to predict the classes of other unclassified records and, more important, to accurately predict actual future events.

Common classification tools include neural networks and decision trees (from machine learning), logistic regression and discriminant analysis (from traditional statistics), and
emerging tools such as rough sets, support vector machines, and genetic algorithms. Statistics-based classification techniques (e.g., logistic regression and discriminant analysis) have received their share of criticism—that they make unrealistic assumptions about the data, such as independence and normality—which limit their use in classification-type data mining projects.

Neural networks involve the development of mathematical structures (somewhat resembling the biological neural networks in the human brain) that have the capability to learn from past experiences presented in the form of well-structured datasets. They tend to be more effective when the number of variables involved is rather large and the relationships among them are complex and imprecise. Neural networks have disadvantages as well as advantages. For example, it is usually very difficult to provide a good rationale for the predictions made by a neural network. Also, neural networks tend to need considerable training. Unfortunately, the time needed for training tends to increase exponentially as the volume of data increases, and, in general, neural networks cannot be trained on very large databases. These and other factors have limited the applicability of neural networks in data-rich domains.

Decision trees classify data into a finite number of classes based on the values of the input variables. Decision trees are essentially a hierarchy of if-then statements and are thus significantly faster than neural networks. They are most appropriate for categorical and interval data. Therefore, incorporating continuous variables into a decision tree framework requires discretization; that is, converting continuous valued numerical variables to ranges and categories.

A related category of classification tools is rule induction. Unlike with a decision tree, with rule induction the if-then statements are induced from the training data directly, and they need not be hierarchical in nature. Other, more recent techniques such as SVM, rough sets, and genetic algorithms are gradually finding their way into the arsenal of classification algorithms.

**CLUSTERING**  
Clustering partitions a collection of things (e.g., objects, events, etc. presented in a structured dataset) into segments (or natural groupings) whose members share similar characteristics. Unlike in classification, in clustering the class labels are unknown. As the selected algorithm goes through the dataset, identifying the commonalities of things based on their characteristics, the clusters are established. Because the clusters are determined using a heuristic-type algorithm, and because different algorithms may end up with different sets of clusters for the same dataset, before the results of clustering techniques are put to actual use it may be necessary for an expert to interpret, and potentially modify, the suggested clusters. After reasonable clusters have been identified, they can be used to classify and interpret new data.

Not surprisingly, clustering techniques include optimization. The goal of clustering is to create groups so that the members within each group have maximum similarity and the members across groups have minimum similarity. The most commonly used clustering techniques include k-means (from statistics) and self-organizing maps (from machine learning), which is a unique neural network architecture developed by Kohonen (1982).

Firms often effectively use their data mining systems to perform market segmentation with cluster analysis. Cluster analysis is a means of identifying classes of items so that items in a cluster have more in common with each other than with items in other clusters. It can be used in segmenting customers and directing appropriate marketing products to the segments at the right time in the right format at the right price. Cluster analysis is also used to identify natural groupings of events or objects so that a common set of characteristics of these groups can be identified to describe them. Application Case 3 describes how cluster analysis was combined with other data mining techniques to identify the causes of accidents.
APPLICATION CASE 3
Motor Vehicle Accidents and Driver Distractions

Driver distraction is at center stage in highway safety. A study published in 1996 by the National Highway Traffic Safety Administration (NHTSA) concluded that roughly 25 to 30 percent of the injuries caused by car crashes were due to driver distraction. In 1999, according to the Fatality Analysis Reporting System (FARS) developed by the National Center for Statistics and Analysis (NCSA), 11 percent of fatal crashes (i.e., 4,462 fatalities) were due to driver inattention.

A study was conducted to extract the patterns of distraction factors at traffic accidents. Data mining was used to draw the correlations and associations of factors from the crash datasets provided by FARS. Three data mining techniques (Kohonen-type neural networks, decision trees, and multilayer perceptron-type neural networks) were used to find different combinations of distraction factors that correlated with and potentially explained the high accident rates. The Kohonen-type neural network identified natural clusters and revealed patterns of input variables in the collection of data. Decision trees explored and classified the effect of each incident on successive events and also suggested the relationship between inattentive drivers and physical/mental conditions. Finally, a multilayer perceptron-type neural network model was trained and tested to discover the relationships between inattention and other driver-related factors in these traffic crashes. Clementine from SPSS was used to mine the data obtained from the FARS database for all three model types.

The prediction and exploration model identified 1,255 drivers who were involved in accidents in which inattention was one of the leading driver factors that led to a crash. Rear, head-on, and angled collisions, among other various output variables, were among the factors that had significant impact on the occurrence of crashes and their severity.


ASSOCIATIONS

Associations, or association rule learning in data mining, is a popular and well-researched technique for discovering interesting relationships among variables in large databases. Thanks to automated data-gathering technologies such as bar code scanners, the use of association rules for discovering regularities among products in large-scale transactions recorded by point-of-sale systems in supermarkets has become a common knowledge-discovery task in the retail industry. In the context of the retail industry, association rule mining is often called market-basket analysis.

Two commonly used derivatives of association rule mining are link analysis and sequence mining. With link analysis, the linkage among many objects of interest is discovered automatically, such as the link between Web pages and referential relationships among groups of academic publication authors. With sequence mining, relationships are examined in terms of their order of occurrence to identify associations over time. Algorithms used in association rule mining include the popular Apriori (where frequent itemsets are identified) and FP-Growth, OneR, ZeroR, and Eclat.

VISUALIZATION AND TIME-SERIES FORECASTING

Two techniques often associated with data mining are visualization and time-series forecasting. Visualization can be used in conjunction with other data mining techniques to gain a clearer understanding of underlying relationships. With time-series forecasting, the data are a series of values of the same variable that is captured and stored over time. These data are then used to develop models to extrapolate the future values of the same phenomenon.