DWM Semester 6 27/4/20

Data mining , also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists).

Data-mining techniques

There are many types of data mining, typically divided by the kind of information (attributes) known and the type of knowledge sought from the data-mining model.

Predictive modeling

Predictive modeling is used when the goal is to estimate the value of a particular target attribute and there exist sample training data for which values of that attribute are known. An example is classification, which takes a set of data already divided into predefined groups and searches for patterns in the data that differentiate those groups. These discovered patterns then can be used to classify other data where the right group designation for the target attribute is unknown (though other attributes may be known). For instance, a manufacturer could develop a predictive model that distinguishes parts that fail under extreme heat, extreme cold, or other conditions based on their manufacturing environment, and this model may then be used to determine appropriate applications for each part. Another technique employed in predictive modeling is regression analysis, which can be used when the target attribute is a numeric value and the goal is to predict that value for new data.

Descriptive modeling

Descriptive modeling, or clustering, also divides data into groups. With clustering, however, the proper groups are not known in advance; the patterns discovered by analyzing the data are used to determine the groups. For example, an advertiser could analyze a general population in order to classify potential customers into different clusters and then develop separate advertising campaigns targeted to each group. Fraud detection also makes use of clustering to identify groups of individuals with similar purchasing patterns.

Pattern mining

Pattern mining concentrates on identifying rules that describe specific patterns within the data. Market-basket analysis, which identifies items that typically occur together in purchase transactions, was one of the first applications of data mining. For example, supermarkets used market-basket analysis to identify

items that were often purchased together—for instance, a store featuring a fish sale would also stock up on tartar sauce. Although testing for such associations has long been feasible and is often simple to see in small data sets, data mining has enabled the discovery of less apparent associations in immense data sets. Of most interest is the discovery of unexpected associations, which may open new avenues for marketing or research. Another important use of pattern mining is the discovery of sequential patterns; for example, sequences of errors or warnings that precede an equipment failure may be used to schedule preventative maintenance or may provide insight into a design flaw.

Data mining

Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists).

Data mining

Data

The proliferation of numerous large, and sometimes connected, government and private databases has led to regulations to ensure that individual records are accurate and secure from unauthorized viewing or tampering. Most types of data mining are targeted toward ascertaining general knowledge about a group rather than knowledge about specific individuals—a supermarket is less concerned about selling one more item to one person than about selling many items to many people—though pattern analysis also may be used to discern anomalous individual behaviour such as fraud or other criminal activity.

Origins And Early Applications

As computer storage capacities increased during the 1980s, many companies began to store more transactional data. The resulting record collections, often called data warehouses, were too large to be analyzed with traditional statistical approaches. Several computer science conferences and workshops were held to consider how recent advances in the field of artificial intelligence (AI)—such as discoveries

from expert systems, genetic algorithms, machine learning, and neural networks—could be adapted for knowledge discovery (the preferred term in the computer science community). The process led in 1995 to the First International Conference on Knowledge Discovery and Data Mining, held in Montreal, and the launch in 1997 of the journal Data Mining and Knowledge Discovery. This was also the period when many early data-mining companies were formed and products were introduced.

One of the earliest successful applications of data mining, perhaps second only to marketing research, was credit-card-fraud detection. By studying a consumer's purchasing behaviour, a typical pattern usually becomes apparent; purchases made outside this pattern can then be flagged for later investigation or to deny a transaction. However, the wide variety of normal behaviours makes this challenging; no single distinction between normal and fraudulent behaviour works for everyone or all the time. Every individual is likely to make some purchases that differ from the types he has made before, so relying on what is normal for a single individual is likely to give too many false alarms. One approach to improving reliability is first to group individuals that have similar purchasing patterns, since group models are less sensitive to minor anomalies. For example, a "frequent business travelers" group will likely have a pattern that includes unprecedented purchases in diverse locations, but members of this group might be flagged for other transactions, such as catalog purchases, that do not fit that group's profile.

Modeling And Data-Mining Approaches

Model creation

The complete data-mining process involves multiple steps, from understanding the goals of a project and what data are available to implementing process changes based on the final analysis. The three key computational steps are the model-learning process, model evaluation, and use of the model. This division is clearest with classification of data. Model learning occurs when one algorithm is applied to data about which the group (or class) attribute is known in order to produce a classifier, or an algorithm learned from the data. The classifier is then tested with an independent evaluation set that contains data with known attributes. The extent to which the model's classifications agree with the known class for the target attribute can then be used to determine the expected accuracy of the model. If the model is sufficiently accurate, it can be used to classify data for which the target attribute is unknown.

Data-mining techniques

There are many types of data mining, typically divided by the kind of information (attributes) known and the type of knowledge sought from the data-mining model.

Predictive modeling

Predictive modeling is used when the goal is to estimate the value of a particular target attribute and there exist sample training data for which values of that attribute are known. An example is classification, which takes a set of data already divided into predefined groups and searches for patterns in the data that differentiate those groups. These discovered patterns then can be used to classify other data where the right group designation for the target attribute is unknown (though other attributes may be known). For instance, a manufacturer could develop a predictive model that distinguishes parts that fail under extreme heat, extreme cold, or other conditions based on their manufacturing environment, and this model may then be used to determine appropriate applications for each part. Another technique employed in predictive modeling is regression analysis, which can be used when the target attribute is a numeric value and the goal is to predict that value for new data.

Descriptive modeling

Descriptive modeling, or clustering, also divides data into groups. With clustering, however, the proper groups are not known in advance; the patterns discovered by analyzing the data are used to determine the groups. For example, an advertiser could analyze a general population in order to classify potential customers into different clusters and then develop separate advertising campaigns targeted to each group. Fraud detection also makes use of clustering to identify groups of individuals with similar purchasing patterns.

Pattern mining

Pattern mining concentrates on identifying rules that describe specific patterns within the data. Market-basket analysis, which identifies items that typically occur together in purchase transactions, was one of the first applications of data mining. For example, supermarkets used market-basket analysis to identify items that were often purchased together—for instance, a store featuring a fish sale would also stock up on tartar sauce. Although testing for such associations has long been feasible and is often simple to see in small data sets, data mining has enabled the discovery of less apparent associations in immense data sets. Of most interest is the discovery of unexpected associations, which may open new avenues for marketing or research. Another important use of pattern mining is the discovery of sequential patterns;

for example, sequences of errors or warnings that precede an equipment failure may be used to schedule preventative maintenance or may provide insight into a design flaw.

Anomaly detection

Anomaly detection can be viewed as the flip side of clustering—that is, finding data instances that are unusual and do not fit any established pattern. Fraud detection is an example of anomaly detection. Although fraud detection may be viewed as a problem for predictive modeling, the relative rarity of fraudulent transactions and the speed with which criminals develop new types of fraud mean that any predictive model is likely to be of low accuracy and to quickly become out of date. Thus, anomaly detection instead concentrates on modeling what is normal behaviour in order to identify unusual transactions. Anomaly detection also is used with various monitoring systems, such as for intrusion detection.

Numerous other data-mining techniques have been developed, including pattern discovery in time series data (e.g., stock prices), streaming data (e.g., sensor networks), and relational learning (e.g., social networks).