Data mining

Refers loosely to the process of semi-automatically analyzing large databases to find useful patterns.

Data mining is mining or discovery of new information in terms of patterns or rules from vast amounts of data. Data mining is not well integrated with database management systems.

Data mining deals with knowledge discovery in the databases.

Some types of knowledge discovered from a database can be represented by a set of rules. For example, "person with monthly income greater than 100000 are most likely to buy a car". Such rules are not universally true, and have degrees of support and confidence.

Other type of knowledge are represented by equations relating different variables to each other, or by other mechanisms for predicting outcomes when the values of some variables are known.

Usually, there is a manual component to data mining, consisting of preprocessing data to a form acceptable to the algorithms and post-processing of discovered patterns to find the novel ones that could be useful.

Overview of data mining technology

- Data Mining versus Data Warehousing
  Data warehousing is creating/maintaining huge amount of data to support decision making with that data. Data Mining is process of discovering information from stored data.
- Data Mining as a part of the knowledge discovery process
  Knowledge Discovery in Databases (KDD) comprises six phases: data selection, data cleansing, enrichment, data transformation or encoding, data mining, and the reporting/display.
- Goals of Data Mining
  Prediction, Identification, Classification, Optimization
- Types of Knowledge discovered during Data Mining
  Association Rules, Classification Trees, Sequential Patterns, Time Series Patterns, Clustering

Association Rules

Association rules refer to the rules discovered with association with another pattern. For example, someone who buys bread is likely also to buy milk.

Bread =>Milk

Rules have an associated support, as well as an associated confidence.

- Support is a measure of what fraction of the population satisfies both the antecedent and the consequent of the rule. For example only 0.001 percent of all purchase include milk and screwdrivers, then the support for the rule milk => screwdriver is very low.
  Businesses are usually not interested in rules that have low support, since they involve few customers.
  On the other hand, if 50 percent of all purchase involve milk and bread, then the support for the rules involving bread and milk is relatively high, and worth attention.
- Confidence is a measure of how often the consequent is true when the antecedent is true. For example, the rule bread => milk has 80% confidence if 80% of the purchases that include bread also include milk. A rule with very low confidence is not meaningful.
  In business applications, rules generally have confidences significantly less than 100%; in scientific applications rules may have high confidences.
- Confidence of bread => milk may be different from the confidence of milk => bread, but they both have same support.

This model of data is called market basket model. The rule generation is done in two steps:

1. Generate all itemsets that have a support exceeding a threshold. These sets of items are called large/frequent itemsets.
2. For each large itemset, all the rules that have a minimum confidence are generated.

Other types of Association rules are:

- Association rules among hierarchies which occur among hierarchies of items.
- Multidimensional Association has more than one value in consequents/antecedents
- Negative associations set of antecedents strictly not followed by consequents.

Classification

Classification is the process of learning a model that describes different predetermined classes of data. For example, in a banking example, customers who apply for a credit card may be classified as a poor risk, fair risk, or good risk.

This process is also called supervised learning; once the model is built, it can be used to classify new data.

The process starts from a sample of data, called training set. For each tuple in the training set, the class to which it belongs is already known. The model produced is usually in the form of a decision tree or a set of rules. The training set for a credit card application may be the existing customers, with their credit-worthiness determined from their payment history. The actual data/population may consist of all people including existing customers as well as new applicants.

Decision Tree Classifiers

Is a widely used technique for classification. Use a tree; each leaf node has an associated class, and each internal node has a predicate. To classify a new instance, we start at the root and traverse the tree to reach a leaf; at an internal node we evaluate the predicate on the data instance, to find which child to go to. This process continues until we reach a leaf node.

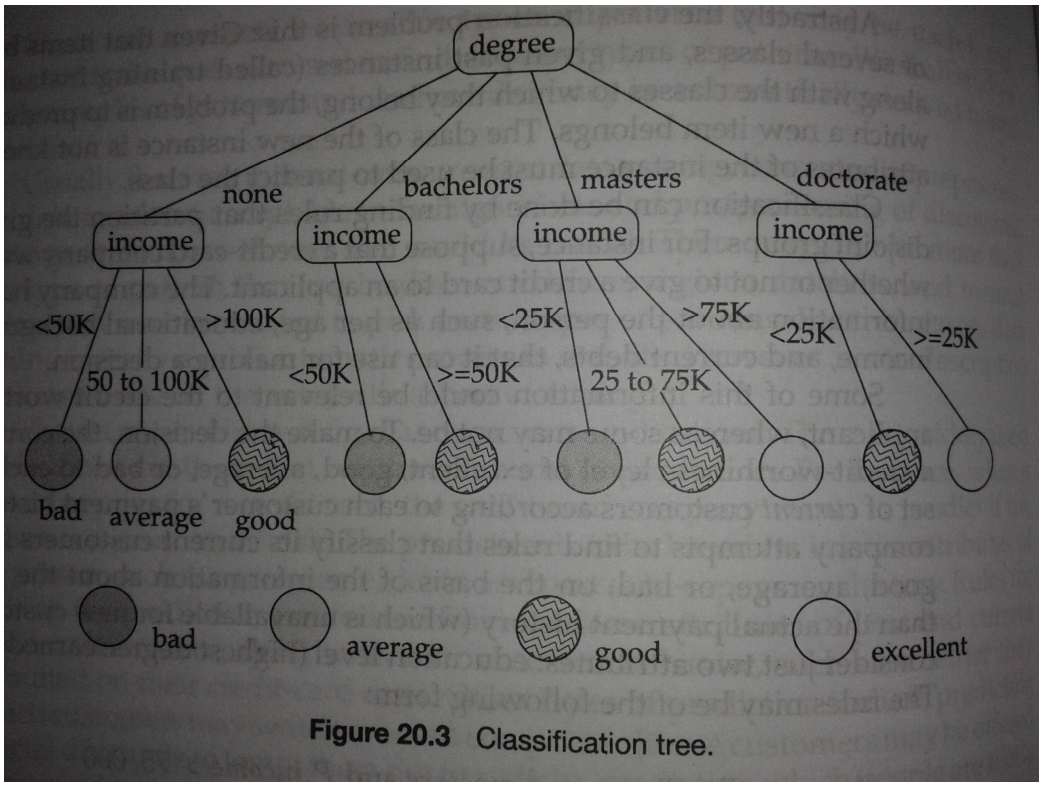

Figure 20.3 Classification tree.

The question is how to build a decision tree classifier, given a set of training instances. The most common way is to use greedy algorithm, which works recursively, starting at the root and building the tree downward. Initially, there is only one node, the root, and all training instances are associated with that node. At each node, if all, or almost all training instances associated with the node belong to the same class, then the node becomes a leaf node associated with that class. Otherwise, a partitioning attribute and partitioning conditions must be selected to create child nodes. The data associated with each child node is the set of training instances that satisfy the partitioning condition of that child node.

Clustering

Refers to the problem of finding clusters of points in given data without having a training sample; also known as unsupervised learning. For example in business, it may be important to determine groups of customers who have similar buying patterns.

The goal of clustering is to place records into groups, such that records in a group are similar to each other and dissimilar to records in other groups. Such groups are usually disjoint.

A classic clustering algorithm is the k-means algorithm.

Input: a database D of m records, r1, r2,…rm and a desired number of clusters k

Output: set of k clusters that minimizes the squared error criterion

Begin

Randomly choose k records as the centroids for k clusters;

Repeat

Assign each record, ri, to a cluster such that the distance between ri and cluster centroid is smallest among the k clusters

Recalculate the centroid for each cluster based on the records assigned to the cluster;

Until no change;

End;

Another type of clustering is hierarchical clustering, as used in biology. Hierarchical clustering can be:

1. Agglomerative clustering: begin from small clusters and create higher levels.
2. Divisive clustering: start from higher level clusters and refine each into lower level clusters.

Applications of data mining

- Marketing
  - Analysis of consumer behavior based on buying patterns
  - Determination of marketing strategies including advertising, store location and targeted mailings
  - Segmentation of customers, stores or products
  - Design of catalogs, store layouts, and advertising campaigns
- Finance
  - Analysis of creditworthiness of clients
  - Segmentation of account receivables
  - Performance analysis of finance investments like stocks, bonds and mutual funds;
  - Evaluation of financing options
  - Fraud detection
- Manufacturing
  - Optimization of resources like machines, manpower, and materials
  - Optimal design of manufacturing processes, shop floor layouts, and product design such as for automobiles based on customer requirements.
- Health care
  - Discovery of patterns in radiological images
  - Analysis of microarray experimental data to cluster genes and relate to symptoms or diseases
  - Analysis of side effects of drugs and effectiveness of certain treatments
  - Optimization of processes within a hospital
  - Analysis relationship of patient wellness data with doctor qualifications.

Data warehouse

A data warehouse is a repository or archive of information gathered from multiple sources, stored under a unified schema, at a single site. Once gathered, the data are stored for a long time, permitting access to historical data.

It is a single, complete and consistent store of data obtained from a variety of different sources made available to end users in a way what they can understand and use in a business context.

A decision support database maintained separately from the organization’s operational database.

Data warehousing is a technique for assembling and managing data from various sources for the purpose of answering business questions; thus making decisions that were not previous possible

Data warehouses provide the user a single consolidated interface to data, making decision support queries easier to write.

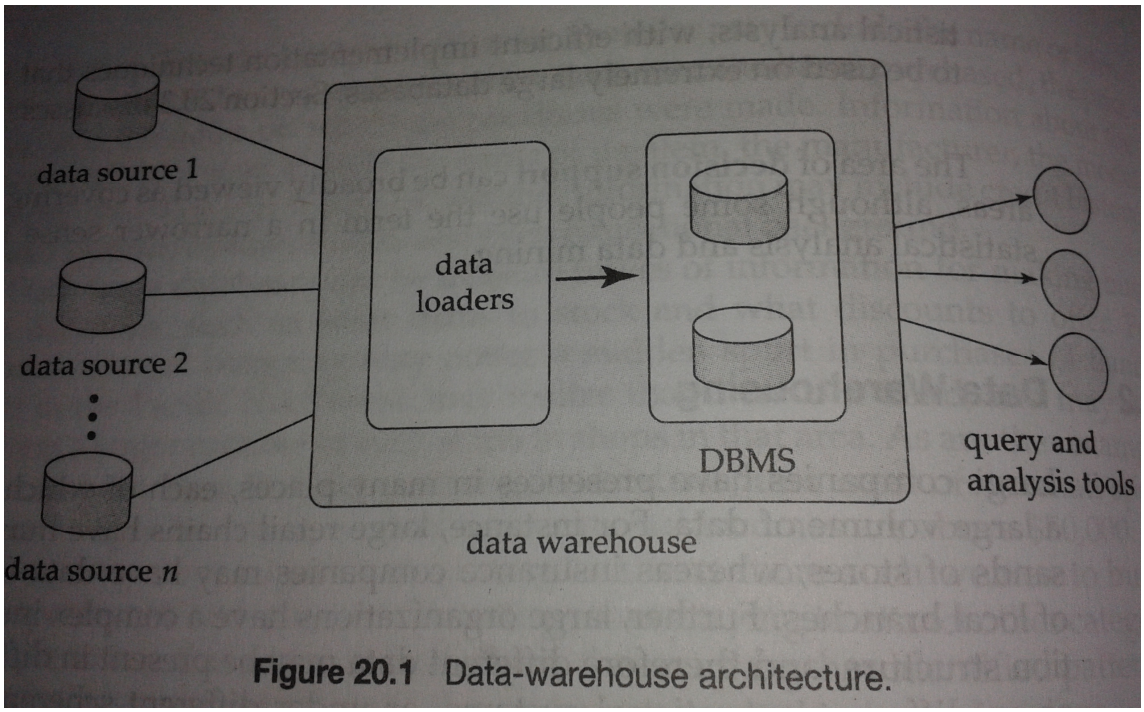

Figure 20.1 Data-warehouse architecture.

Moreover, by accessing information for decision support from a data warehouse, the decision maker ensures that online transaction-processing systems are not affected by the decision-support workload.

Issues to be addressed in building a warehouse are:

- When and how to gather data
- What schema to use
- Data transformation and cleansing
- How to propagate updates
- What data to summarize

For efficiency, data warehouse may use column oriented storage; each attribute of a relation stored in a separate file to have following benefits:

- Less memory consumption when we need only a few attributes of a relation with a large number of attributes.
- Storing values of same type together increases the effectiveness of compression.

Drawback of column oriented storage is that storing or fetching a single tuple requires multiple I/O operations.

Typical Functionality of data Warehouses

- Roll up
- Drill down
- Pivot
- Sorting
- Selection
- Derived attributes

XML- Enabled Database

XML enabled database is nothing but the extension provided for the conversion of XML document. This is relational database, where data are stored in tables consisting of rows and columns. The tables contain set of records, which in turn consist of fields.

Native XML Database

Native XML database is based on the container rather than table format. It can store large amount of XML document and data. Native XML database is queried by the XPath-expressions.

Native XML database has advantage over the XML-enabled database. It is highly capable to store, query and maintain the XML document than XML-enabled database.

NXD defines a (logical) model for an XML document -- as opposed to the data in that document -- and stores and retrieves documents according to that model. At a minimum, the model must include elements, attributes, PCDATA, and document order. Examples of such models are the XPath data model, the XML Infoset, and the models implied by the DOM and the events in SAX 1.0.

It has an XML document as its fundamental unit of (logical) storage, just as a relational database has a row in a table as its fundamental unit of (logical) storage.

Is not required to have any particular underlying physical storage model. For example, it can be built on a relational, hierarchical, or object-oriented database, or use a proprietary storage format such as indexed, compressed files.