

Data Mining Presentation

Robert James

Eastern Michigan University

Data Mining is a multidisciplinary field, drawing work from areas including database management systems, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing, and data visualization.

There are a number of definitions of Data Mining in the literature. However, they all have in common the following; “extraction”, “knowledge”, and “large data.”(Abbass, 2005).

Data mining refers to the extracting or “mining” knowledge from large amount of data. The process of performing data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. The exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

The rules include the iterative process of detecting and extracting patterns from large databases. It lets us identify “signatures” hidden in large databases, as well as learn from repeated examples. The extraction of implicit, previously unknown, and potentially useful information from data is the ultimate goal of any statically viable approach. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data.

Data Mining automates the detection of relevant patterns in databases. For example, a pattern might indicate that married males with children are twice as likely to drive a particular sports car than married males with no children. It uses well-established statistical and machine learning techniques to build models that predict customer behavior.

Data Mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases. The objective of this process is to sort through large quantities of data and discover new information. The benefit of data mining is to turn this newfound knowledge into actionable results, such as increasing a customer’s likelihood to buy, or decreasing the number of fraudulent claims.

Data Mining is also the search for valuable information in large volumes of data. It is a cooperative effort of humans and computers. Humans design databases, describe problems and set goals. Computers sift through data, looking for patterns that match these goals. Data Mining

Running Head: DATA MINING

is a search for very strong patterns in big data that can generalize to accurate future decisions. It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. It is currently regarded as key element of much more elaborate process called Knowledge Discovery in Databases (KDD).

The term Data Mining can be used to describe a wide range of activities. A marketing company using historical response data to build models to predict who will respond to a direct mail or telephone solicitation is using data mining. A manufacturer analyzing sensor data to isolate conditions that lead to unplanned production stoppages is also using data mining. The government agency that sifts through the records of financial transaction looking for patterns that indicate money laundering or drug smuggling is mining data for evidence of criminal activity. Other terms in the literature are sometimes used to describe this process in addition to Data Mining. Among these are, knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Data mining models produces one or more output values for a given set of inputs. Analyzing data is often the process of building an appropriate model for the data. It is an abstract representation of reality. Models in Data Mining are either Predictive or Descriptive and include the following (Han, 2003):

- **Classification:** This model is used to classify database records into a number of predefined classes based on certain criteria. For example, a credit card company may classify customer records as a good, medium, or poor risk. A classification system may then generate a rule stating that “If a customer earns more than \$40,000, is between 45 to 55 in age, and lives within a particular ZIP code, he or she is a good credit risk.”
- **Prediction:** This model predicts the value for a specific attribute of the data item. For example, given a predictive model of credit card transactions, predict the likelihood that a specific transaction is fraudulent. Prediction may also be used to validate a discovered hypothesis.
- **Regression:** This model is used for the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic prediction of these attribute values for new records. For example, given a data set of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions.

- **Time Series:** This model describes a series of values of some feature or event that are recorded over time. The series may consist of only a list of measurements, giving the appearance of a single dimension, but the ordering is by the implicit variable, time. The model is used to describe the component features of a series data set so that it can be accurately and completely characterized.
- **Clustering:** This model is used to divide the database into subsets, or Clusters, based on a set of attributes. For example, in the process of understanding its customer base, an organization may attempt to divide the known population to discover clusters of potential customers based on attributes never before used for this kind of analysis (for example, the type of schools they attended, the number of vacation per year, and so on). Clusters can be created either statistically or by using artificial intelligence methods. Clusters can be analyzed automatically by a program or by using visualization techniques.
- **Association:** This model is used to identify affinities among the collection, as reflected in the examined records. These affinities are often expressed as rules. For example: “60% of all the records that contain items A and B also contain items C and D.” The percentage of occurrences (in this case, 60) is the confidence factor of association. Association model is often applied to Market Basket Analysis, where it uses point-of-sale transaction data to identify product affinities.
- **Sequencing:** This model is used to identify patterns over time, thus allowing, for example, an analysis of customer purchases during separate visits. It could be found, for instance, that if a customer buys engine oil and filter during one visit, he will buy gasoline additive the next time. This type of models is particularly important for catalog companies. It’s also applicable in financial application to analyze sequences of events that affect the prices of financial instruments.
- **Characterization:** This model is used to summarize the general characteristics or features of a target class of data. The data corresponding to the user-defined class are typically collected by a database query. For example, to study the characteristics of software products whose sales increased by 10% in the last year, the data related to such products can be collected by executing an SQL query.
- **Comparison (Discrimination):** The model is used for comparison of the general features of target class data objects with the general features of objects from one or a set of

comparative (contrasting) classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through SQL queries. For example, the model can be used to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period

There are two classes of data to be mined, qualitative and quantitative. Qualitative data use descriptive terms to differentiate values. For example, gender is generally classified into “M” or male and “F” or female. Qualitative data can be used for segmentation or classification, where quantitative data is characterized by numeric values. Gender could also be quantitative if prior rules are established. For example, we could say the values for gender are 1 and 2 where 1=“M” or male and 2=“F” or female. Quantitative data is used for developing predictive models.

Quantitative data falls into four types (Han, Kamber, 2003):

- Nominal Data is numeric data that represents categories or attributes. The numeric values for gender (1 and 2) would be nominal data values. One important characteristic of nominal data is that it has no relative importance. For example, even though male = 1 and female = 2, the relative value of being female is not twice the value or higher value than that of being male. For modeling purposes, a nominal variable with only two values would be coded with values 0 and 1 (Binary). Other examples are geographical_location, map_color, and item_type.
- Ordinal Data is numeric data that represent categories that have relative importance. They can be used to rank strength or severity. For example, a list a company assigns the values 1 through 5 to denote financial risk. The value 1, characterized by no late payment, is considered low risk. The value 5, characterized by a bankruptcy, is considered high risk. The values 2 through 4 are characterized by various previous delinquencies. A prospect with a ranking of 5 is definitely riskier than a prospect with a ranking of 1. But he or she is not 5 times as risky. And the difference in their ranks of $5-1 = 4$ has no meaning.
- Interval Data is numeric data that has relative importance and has no zero point. Also, addition and subtraction are meaningful operations. For example, many financial institutions use a risk score that has much finer definition than the values 1

through 5, as in the previous example. A typical range is from 300 to 800. It is therefore possible to compare scores by measuring the difference.

- Continuous Data is the most common data used to develop predictive models. It can accommodate all basic operations, including addition, subtraction, multiplication, and division. Most business data, such as sales, balances, and minutes, are continuous data.

In general, common goals of data mining applications and models include the detection, interpretation, and prediction of qualitative and/or quantitative patterns in data. To achieve these goals, data mining solutions employ a wide variety of techniques of machine learning, artificial intelligence, statistics, and database query processing. These algorithms are also based on mathematical approaches such as multi-valued logic, approximation logic, and dynamic systems. There are relatively many Data Mining Algorithms. Some of them are mentioned below together with the type of models they are capable of solving (Daimi, 2007).

- Association Algorithms: These are used to solve Association Models. Association Algorithms include The Apriori Algorithm, PCY Algorithm, Iceberg Algorithm, AIS Algorithm, STEM Algorithm, AprioriHybird Algorithm, Toivonen Algorithm, and Frequent Pattern Growth Algorithm.
- Clustering Algorithms: These are used to cluster or segment data. They include K-Means, BFR Algorithm, BIRCH Algorithm, CURT Algorithm, Chamelon Algorithm, Incremental Clustering, DBSCAN Algorithm, OPTICS Algorithm, DENCLUE Algorithm, Fast Map Algorithm, GRGPF Algorithm, STING Algorithm, Wave Cluster Algorithm, CLIQUE Algorithm, and COBWEB Algorithm.
- Decision Trees: These are analytical tools used to discover rules and relationships by systematically breaking down and subdividing the information contained in the data. These algorithms seek to find those variables or fields in the data set that provide maximum segregation of the data records. Decision trees are useful for problems in which the goal is to make broad categorical Classifications and Predictions.

The following statistical methods are available in the data mining process.

- Linear Regression: Maps values from a predictor, so that the fewest errors occur when making a prediction. Linear regression contains only one predictor (X) and a prediction (Y) using the equation $Y = a + bX$. If we have more than one predictor that are still

linear, then we have Multiple Linear Regression, $Y = a + b X_1 + c X_2 + d X_3 + \dots$. Linear Regression is used for Prediction.

- **Logistic Regression:** This is a regression method in which the prediction (Y) just has a yes/no or 0/1 value. Logistic Regression is used to predict response problems (just yes or no). For example, customers bought the product or did not buy it.
- **Nonlinear Regression:** In this method, the predictors are nonlinear, $Y = a + b X + c X^2 + \dots$. Nonlinear Regression is also used for Prediction models.
- **Discriminate Analysis:** Finds a set of coefficient or weights that describe a Linear Classification Function (LCF), which maximally separates groups of variables. A threshold is used to classify an object into groups. The LCF is compared to this threshold to decide the group. It is used for Segmentation.
- **Bayesian Method:** This algorithm tries to find an optimum classification of a set of examples using the probabilistic theory of Bayesian Analysis. It can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. It is used for Classification.
- **Neural Networks:** This algorithm is based on the architecture of the brain consisting of multiple simple processing units connected by adaptive weights. It is a collection of processing units and adaptive connections designed to perform a specific processing function. A neural network is particularly used for Classification but can also be used for Prediction. A version of neural networks, called Kohonen Feature Map, is used for Clustering. Neural networks are also used for Estimation.
- **K-Nearest Neighbor:** This is a Predictive technique. In order to predict what a prediction value is in one record, look for records with similar predictor values in the historical database, and use the prediction value from the record that is “nearest” to the unclassified record. In other words, it performs prediction by finding the prediction value of records similar to the record to be predicted. The data used by the K-nearest algorithm is numeric.
- **Genetic Algorithms:** These algorithms solve problems by borrowing a technique from nature. GAs use Darwin’s basic principles of survival of the fittest, mutation, and crossover to create solutions for problems. When a GA finds a good solution, it percolates some of that solution’s features into a population of competing solutions.

Over time, the GA “breeds” good solutions. Genetic Algorithms are optimization techniques. They are used for Classification. Another important use for them is in finding the best possible combination of link weights for a given neural network architecture.

- **Case Based Reasoning:** This is an intelligent-systems method that enables information managers to increase efficiency and reduce cost by substantially automating some processes such as scheduling, diagnosis and design. It works by matching new problems to “cases” from historical database and then adapting successful solution from the past to the current situations. Case-Based Reasoning is used for Classification and Prediction. Unlike the K-nearest algorithm, Case-based reasoning works on symbolic data.
- **Fuzzy Sets:** Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today’s databases including imprecision, inconsistency, and vagueness. They constitute a powerful approach to deal not only with incomplete noisy or imprecise data, but may also be helpful in developing uncertain models of data that provide smarter and smoother performance than traditional systems. They are used for Prediction in situations where precise input is unavailable or too expensive.
- **Rough Sets:** A rough set is defined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every nonmember of the upper bound is a certain nonmember of the set. The upper bound of a rough set is the union between the lower bound and the boundary region. A member of the boundary region is possibly a member of the set. Rough sets may be viewed as fuzzy sets with three-valued membership function (yes, no, perhaps). Rough sets are seldom used as a stand-alone algorithm. They are usually combined with other methods such as classification, clustering, or rule induction.

At the data collection stage, identification of available data and required data is a must. Sources of the collected data will be databases, data warehouses, purchased data, and public data. In addition, data should be collected from surveys, and if applicable, additional data elements are obtained from forms. Data consolidation combines data from different sources into a single mining database. It requires reconciling differences in data. Improperly reconciled data is a source of quality problems.

The process should take care of data heterogeneity (differences in the way data is defined and used in different databases). Among these differences are (Daimi, 2007):

- Homonyms: same name for different things
- Synonyms: different names for the same thing
- Unit incompatibilities: English vs. metric units
- Different attributes for the same entity
- Different ways of modeling the same fact

The data quality has great impact on the algorithm and the analysis. Data quality is reduced by incorrect data (noise) due to content of a single field, inconsistent data, and integrity violation, in addition to missing data. Noise is the part of data that is not explained in the model ($\text{Data} = \text{Model} + \text{Noise}$). Sources of noise include erroneous data, exceptional data, and extraneous columns and rows. Algorithms should degrade gracefully as noise increases. Another source of errors is missing data. When values are missing, we can model the mechanism that causes them to be missing and include these terms in our overall model. In order to model this mechanism, we must know the values of the missing data. We should not let the data go missing. There are several techniques to deal with missing data: ignoring the tuple, filling the missing value manually, and fill in the missing value with the attribute mean (Grossman, 2005).

Sampling must be done intelligently by analyzing your data and data mining task. You can also do random sampling by making a random probability selection from a database. Most model building techniques require separating the data into training and testing samples. It is very important not to sample when information loss is too great. This occurs when the database is small or the sampling has small effects in large databases (Grossman, 2005).

A number of Data Mining software tools exist in the market. These tools are freeware, shareware, or commercial. The price for these tools varies depending on the functionality they provide. Most of these tools provide demos. When considering buying or using a software tool, the computer architecture on which the software runs should be considered. This architecture could be a Standalone, Client/server, or Parallel Processing. Also, the operating system, for which the run time version of the software can be obtained, must be considered. The following points should be considered when selecting a Data Mining software tool (Daimi, 2007).

Running Head: DATA MINING

The follow are steps in the Data Mining Process. Each step must be taken, and in the proper order in which to have use full data to evaluate (Squire, date unknown).

- Project Goal: Develop an understanding of the application domain, the relevant prior knowledge, and the goals of the end user
- Model Goal: Discern whether to estimate, classify, describe, control, detect changes, find dependencies, or cluster.
- Data: Create or select the target data set, focusing on a subset of potential data sources and/or subsamples of huge databases
- Filter: Clean and preprocess data (handle noise, outliers, missing data, time dependencies, normalization, etc.)
- Reduce: Select meaningful subsets of variables, eliminate redundant dimensions, combine or project groups of inputs.
- Expand: Hypothesize useful transformations and combinations
- Mine: Choose a data mining algorithm and monitor its execution
- Evaluate: Measure model accuracy on training and evaluation data (cross-validate), and note its simplicity, robustness, and clarity
- Implement: Document the model and embed it in the decision system

The data mining tasks involved in processing information are enormous. The ability to even gather and process and evaluate data is no longer an impossible task. With the algorithms available, statistical analysis has taken on another face through data mining. We need look no further than what data mining is doing in the retail sector to see the benefits in specific customer feedback.

Bibliography

Abbass, Hussein A. , Ruhul A. Sarker, Charles Sinclair Newton Data Mining: A Heuristic Approach, Idea group Publishing inc. , 2005

Daimi, Kevin, Data Mining software, University of Detroit Mercy, 2007

Grossman, Robert L. ,DataMining for Scientific and engineering applications, Springer Publisher, 2005

Han, Jaiwei, Micheline Kamber, Data Mining: concepts and techniques, Elsevier publishing, 2003

Squire, Linda, What is Data Mining, SBSS-DAMA-NCR, Date Unknown