Introduction to Data Mining

Let’s start with some basic definitions. First of all, what the term “Data Mining” means? A well-formed definition was posted by Michael J. A. Berry & Gordon S. Linoff: “Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover patterns and rules”. Another definition comes from Raplh Kimball: “Data mining is a collection of powerful analysis techniques for making sense out of very large datasets”. Of course we have a definition by another guru, Bill Inmon: “Data Mining / Data Exploration is the usage of historical data to discover and exploit important business relationships”.

So we can say, we are deducing some hidden knowledge by examining, or training the data. Our unit of examination is called “case”, which can be interpreted as one appearance of an entity, or a row in a table. The knowledge is patterns and rules. In the process we are using attributes of a case, in the data mining terminology called “variables”.

Additional goal of DM is to perform predictions based on found patterns.

For better understanding we can compare Data Mining to OLAP. While OLAP is a model-driven analysis, where we build the model in advance (what if it is irrelevant?), the Data Mining is a data-driven analysis, where we search from the model in our data.

The Data Mining techniques are divided into two main classes: the directed and the undirected approach. With the directed approach we use known examples and apply gleaned information to unknown examples to predict selected target variable(s). Using the undirected approach, we are just trying to discover new patterns inside the data set as a whole. Analysis Services 2000 cover both approaches: Clustering is the undirected and Decision Trees is the directed one.

Another definition of the two styles has been posted by Bill Inmon. He divides the process of the analysis in two parts as well, called the Data Exploration and Data Mining. First is done by “Explorers” who access data infrequently, don’t know what they want, often find nothing, look at lots of data, look at things randomly and occasionally find huge nuggets. Second is done by “Farmers” who know what they are looking for, frequently look for things, have a repetitive pattern of access, look for small amounts of data and frequently find small flakes of gold. We can see that we are talking about the same things – the “Explorers” are doing the undirected Data Mining and the “Farmers” the directed one.

Some of the most important directed techniques include Classification, Estimation and Prediction. Classification means to examine a new case and assign it to predefined discrete class. Examples are assigning keywords to articles, customers to known segments and more. Very similar is Estimation, where we are trying to estimate a value of a variable of a new case in continuously defined pool of values. We can, for example, estimate number of children, family’s income… The Prediction is not far away as well. The main difference is that we can’t check the predicted value at the time of prediction. Of course we can evaluate it if we just wait enough. Examples include predicting which customers will leave in the future, which customers will order additional services …

The most common undirected techniques are Clustering and Affinity Grouping. An example of clustering is looking through a large number of initially undifferentiated customers and trying to see if they fall into natural groupings. This is a pure example of "undirected data mining" where the user has no preordained agenda and is hoping that the data mining tool will reveal some meaningful structure. An example of classifying is to examine a candidate customer and assign that customer to a predetermined cluster or classification. Another example of classifying is medical diagnosis. In both cases, a verbose description of the customer or patient is fed into the classification algorithm. We see that the previous activity of clustering may well be a natural first step that is followed by the activity of classifying.

Affinity grouping is a special kind of clustering that identifies events or transactions that occur simultaneously. A well-known example of affinity grouping is market basket analysis. Market basket analysis attempts to understand what items are sold together at the same time.

Some authors include simple methods like Description and Visualization in the Data Mining toolbox. The methods like frequencies distribution, cross-tabulation, graphical mapping of data, OLAP cubes are not methods of great statistical pretension. We use them for a quick preview of the data and to check if it is appropriate for more advanced methods.

A very well-known example of usage of Data Mining techniques is Amazon.com. It uses Data Mining to find cross-selling opportunities – books recommendations a customer gets based on her / his previous purchases and purchases of other customers. Association Rules and Decision Trees can be used for similar tasks.

Data Mining is popular in banks for detection of fraudulent usage of credit cards. Clustering technique can help you determine frauds.

Telecom, banking and insurance companies experience switches of their customers to competitors. This is called churn. Churn detection is an important task for them. Any directed technique, including Naive Bayes, Decision Trees and Neural Networks can be used for this task.

CRM systems should allow you to get more knowledge about your customers. A common way to comprehend your customers more is to segment them. Clustering algorithm is very appropriate for segmentation.

Nearly all Web sites count number of visits. But how many of them follow how visitors use the site? Which pages are visited more frequently and in what order? Sequence Clustering helps you understanding the usage of your Web site and thus gives you the knowledge how to organize pages in best possible way.

Also nearly any business needs some forecasting. This can be done by using the Time Series technique.

I hear so many times that Data Mining is cool; however, the time to implement it and use it in our daily operations is not appropriate yet. Well, my opinion is completely different. Take a look at business growth of Amazon.com; how long will regular bookstores be able to compete with Amazon.com without similar customer experience, namely without Data Mining? Actually, using Data Mining is becoming more a necessity for every company and not an advantage of some rare companies anymore.