Data Mining Techniques

DATA MINING TECHNOLOGIES

Neural Networks/Pattern Recognition - Neural Networks are used in a black box fashion. One creates a test data set, lets the neural network learn patterns based on known outcomes, and then sets the neural network loose on huge amounts of data. For example, a credit card company has 3,000 test records, 100 of which are known fraud records and 2,900 non-fraud records. The data set updates the neural network to make sure it knows the difference between the fraud records and the legitimate ones. The network learns the patterns of the fraud records. Then the network is run against the company's billion record data set and the network spits out the records with patterns the same or similar to the fraud records. Neural networks are known for not being very helpful in teaching analysts about the data, just finding patterns that match. Neural networks have been used for optical character recognition to help the Post Office automate the delivery process without having to use humans to read addresses.

Memory Based Reasoning - This technique has results similar to neural network but goes about it differently. MBR looks for "neighbor" kinds of data, rather than patterns. If you look at insurance claims and want to know which the adjudicators should look at and which they can just let go through the system, you would set up a set of claims you want adjudicated and let the technique find similar claims.

Cluster Detection/Market Basket Analysis - This is where the classic beer/diapers bought together analysis came from. It finds groupings. Basically, this technique finds relationships in product or customer or wherever you want to find associations in data. Market Basket itself specifically refers to the groupings of items bought together at the same time – "what's in our customers basket at the checkout line?"

Link Analysis - This is another technique for associating like records. Probably not used as much, but there are some tools created just for this. As the name suggests, the technique tries to find links, either in customers, transactions, etc. and demonstrate those links.

Visualization - This technique helps users understand their data, often massive amounts of data. Visualization makes the bridge from text-based to graphical presentation. Such things as decision tree, rule, cluster and pattern visualization help users see data relationships rather than read about them. Many of the stronger data mining programs have made strides in improving their visual content over the past few years. This is really the vision of the future of data mining and analysis. Data volumes have grown to such huge levels, it is going to be impossible for humans to process it by any text-based method effectively, soon. We will probably see an approach to data mining using visualization appear that will be something like Microsoft's Photosynth. The technology is there, it will just take an analyst with some vision to sit down and put it together.

Decision Tree/Rule Induction - Decision trees use real data mining algorithms. Decision trees help with classification and spit out information that is very descriptive, helping users to understand their data. A decision tree process will generate the rules followed in a process. For example, a lender at a bank goes through a set of rules when approving a loan. Based on the loan data a bank has, the outcomes of the loans (default or paid), and limits of acceptable levels of default, the decision tree can set up the guidelines for the lending institution. These decision trees are very similar to the first decision support (or expert) systems.

Genetic Algorithms - GAs are techniques that act like bacteria growing in a petri dish. You set up a data set then give the GA ability to do different things for whether a direction or outcome is favorable. The GA will move in a direction that will hopefully optimize the final result. GAs are used mostly for process optimization, such as scheduling, workflow, batching, and process re-engineering. Think of GAs as simulations

CBIG Consulting

run over and over to find optimal results and the infrastructure around being able to both run the simulations and the ways to set up which results are optimal.

OLAP – Online Analytical Processing (OLAP) allows users to browse data following logical questions about the data. OLAP generally includes the ability to drill down into data, moving from highly summarized views of data into more detailed views. This is generally achieved by moving along hierarchies of data. For example, if one were analyzing populations, one could start with the most populous continent, then drill down to most populous country, then to the state level, then to the city level, then to the neighborhood level. OLAP also includes browsing up hierarchies (drill up), across different dimensions of data (drill across), and many other advanced techniques for browsing data, such as automatic time variation when drilling up or down time hierarchies. OLAP is by far the most implemented and used technique and is quite intuitive for users.

DATA MINING USES

Classification - This means getting to know your data. If you can categorize, classify, and/or codify your data, you can place it into chunks that are manageable by a human. Rather than dealing with 3.5 million merchants at a credit card company, if we could classify them into 100 or 150 different classifications that were virtually dead on for each merchant, a few employees could manage the relationships rather than needing a sales and service force to deal with each customer individually. Likewise, at a university, if an alumni group treats its donors according to their classifications, part-time students might be the representatives who work with minor donors and full-time professionals might receive incoming calls from the donors whose names appear on buildings on campus.

Estimation - This process is useful in just about every facet of business. From finance to marketing to sales, the better you can estimate your expenses, product mix optimization, or potential customer value, the better off you will be.

Prediction - Forecasting, like estimation, is ubiquitous in business. Accurate prediction can reduce

inventory levels (costs), optimize sales, etc. If you can predict the future, you will rule the world.

Affinity Grouping/Market Basket Analysis - This is a use that marketing loves. Product placement within a store can be set up based on sales maximization when you know what people buy together. There are several schools of thought on how to do it. For example, you know people buy paint and paint brushes together. One, do you make a sale on paint then jack up the prices on brushes, two do you put the paint in aisle 1 and the brushes in aisle 7 hoping that people walking from one to the other will see something else they will need, three do you set cheap stuff on the end of the aisle for everyone to see hoping they will buy it on impulse knowing they will need something else with that impulse buy (chips and dip, charcoal briquettes and lighter fluid, etc). As you can see, knowing what people buy together has serious benefits for the retail world.

Clustering/Target Marketing - Target marketing saves millions of dollars in wasted coupons, promotions, etc. If you send your promo to only the most likely to accept the offer, use the coupon, or buy your product, you will be much better served. If you sell acne medication, sending coupons to people over sixty is usually a waste of your marketing dollars. If, however, you can cluster your customers and know which households have a 75% chance of having a teenager, you are pushing your marketing on a group most likely to buy your product.

Description – This is similar to classification, but geared more toward explanation. Classification may put more women as candidates for breast cancer, while description will point out the reasons why that classification is the way it is. Users who deal with demographics are often concerned with description. Information services, for example, both classify and describe. They need to classify for obvious reasons. They need to describe so they can make their money. If they can tell a major manufacturer why advertising one way versus another will be more effective, that manufacturer is more likely to buy their services.

Author: Chris Ford, CBIG Consulting

For more information, call CBIG at (800) 334-2078 or visit our Web site at www.cbigconsulting.com.

CBIG Consulting