

IssueDate: May 2002, Posted on: 5/1/2002

Mining Factory Data

Larry Adams
adamsl@bnp.com



For manufacturers, making sense of the millions of bytes of data generated daily by many factory operations has always been a daunting task. Today, however, that task is becoming far less challenging, thanks to advances in data mining software.

Data mining, which uses statistical analysis and modeling techniques to uncover patterns and relationships hidden in large databases, has never been easier. Today's software integrates multiple complex analytics into its programming so that many of the jobs that previously required laborious data input and sophisticated programming can be accomplished by the click of a mouse. The software can now automatically convert raw data into nuggets of information that are available by drilling down through user-friendly menus, which frees up users to do even greater, more sophisticated statistical analysis.

In the last decade, as computer networks developed and more and more test, measurement and inspection equipment was tied into computer programs, the question of how best to use the data has become more important. For many, data mining is the answer. But data mining is only a single component, albeit the most important component, of a larger process called Knowledge Discovery in Databases. This cycle includes data selection, data preprocessing, data transformation, data mining, interpretation and evaluation.



In the past, these were difficult and time-consuming tasks. However, available software has taken on many of these jobs. Instead of developing data ware-houses, most software can extract data from such systems as ERP, MES, SCADA, HMI and SPC systems, from databases such as Oracle, MS-SQL and Sybase, from flat files such as Excel, Lotus 1-2-3, Quattro Pro, ASCII text files and equipment data files, and from Internet and Intranet pages and sites.

Today's software can also clean and filter the data that has been collected, which is a vital task. According to the Data Warehouse Institute (Seattle, WA), quality problems cost U.S. businesses more than \$600 billion a year.

Just click

Whether it is cleaning data or analyzing it, today's software can eliminate much of the work with the simple click of a mouse. Statistica Data Miner, supplied by StatSoft Inc. (Tulsa, OK), has nodes for data input and acquisition, nodes for data filtering and cleansing and nodes for data analysis. The software features 260 procedures, which the company calls Visual Basic scripts, that are used to specify relationships and control the flow of data.

Qualtrend, a software product from DataNet Quality Systems (Southfield, MI), lets users formulate and track Key Performance Indicators, but requires no data storage investment because it can mine existing databases and storage systems.

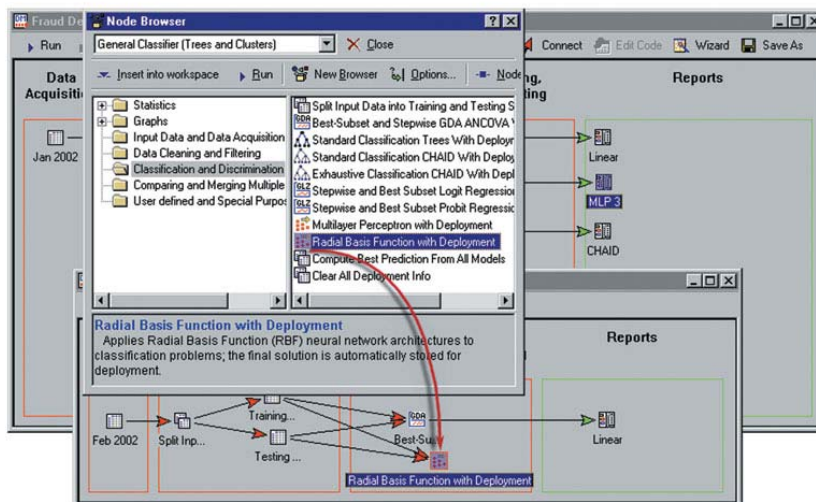
Another program, SAS Decision Trees and Tree Viewer, from the SAS Institute (Cary, NC), allows end users to make predictions and identify factors that can help provide interpretable rules or logic statements. This can help explain the cause of known manufacturing problems, unearth unknown problems and help a company make better business decisions.

"Usually, the data mining software has features to simplify the graphic representation of the data, plus interfaces to common database formats," says Herbert Edelstein, president of Two Crows Corp. (Potomac, MD) and author of *Introduction to Data Mining and Knowledge Discovery and Data Mining 2001: Technology Report*.

Operating on data

Intuitive Surgical (Sunnyvale, CA) is a company that has successfully integrated data mining software into its production operation. The company makes what is commonly referred to as robotic surgical equipment. Though the company's da Vinci Surgical System product isn't truly robotic, it does rely on mechanical manipulators that are controlled by the surgeon to perform minimally invasive procedures.

The da Vinci system has three mechanical arms that are controlled by the surgeon who sits a few feet away looking into a viewfinder at 3-D images sent by cameras attached to the arms. The arms have what the company calls Endowrists, which mimic the movements of the surgeon. The surgeon manipulates the arms and wrists with something akin to a joystick. The arm and viewfinder allow the surgeon to work in extremely small areas. A gallbladder operation, for instance, would require three incisions, each no larger than the diameter of a pencil.



Menus such as this one allow users to point-and-click to do a variety of tasks such as data cleaning and filtering, regression analysis or graphing data. Photo: StatSoft

The sophisticated product has more than 2,500 parts on its bill of materials that include mechanical parts, electronics, optics and other vision components as well as myriad materials. "It has a metal frame, but it has just about every material and part that you can think of," says Steve Lucchesi, director of information services for Intuitive Surgical.

The company is regulated by the Food and Drug Administration (FDA) and is required to track the manufacturing history in detail for every unit shipped. This task was made more challenging in July 2000 when the FDA approved the da Vinci Surgical System and sales began to ramp up, Lucchesi says.

"We record every quality incident on the manufacturing line in detail for every unit we ship to the field. In addition, when we have field issues, we put them into the data mining software and we track them through a full failure analysis process," Lucchesi says. "With so many different parts and different qualifications for the different parts, the amount of data we collect can be overwhelming and be of very little use to us because we would not be able to see any trends."

The company uses Datasweep Advantage 5.0 from Datasweep (San Jose, CA), which is a Web-based integrated plant system, to see these trends. The company is able to compare both field issues and manufacturing issues to "see the total picture for a given subassembly or unit," Lucchesi says. It allows data to be taken from databases and suppliers and integrate them for analysis. The software features an enhanced manufacturing dashboard and allows for global reporting and analysis. It allows for automatic, prioritized alerts via e-mail, pager and phone, and multilevel drill down to pinpoint production problems anywhere in the world. It provides a global view of operations including quality, supplier performance, inventory and overall plant performance across the enterprise.

A typical application at Intuitive is to run Pareto charts on any given subassembly or part number and then drill down to find the problem. "Initially, we look at specific subassemblies that we are having problems with," Lucchesi says. "Then, we have a four-level failure code that is increasingly more detailed in terms of what is

wrong with the part. This allows us to drill down from the part number to the major field symptoms or factory floor symptoms. We can drive down from the top level parts to the lower level parts of the subassembly, all the way down the bill of materials."

Sharing information

Another key to today's data mining software is the end-users' ability to share information that has been mined. For example, Cymer Inc. (San Diego, CA) allows any of its employees to drill down into data to try and find problem areas and solutions.

Cymer builds excimer laser illumination sources, which are the essential light sources for deep ultraviolet photolithography systems used in manufacturing semiconductors. An excimer laser uses a noble-gas halide to generate radiation, usually in the ultraviolet region of the spectrum, and has several critical components, called consumables, that are closely monitored both on the factory floor and out in the field. Tracking the parts and predicting potential problems is important, because in the semiconductor industry, any downtime can cost Cymer's customers millions of dollars, says Sashi Murty, Cymer's manager of failure analysis.

The company uses the Statsserver software product from Insightful Corp. (Seattle, WA). Statsserver is a Web-based system that uses Insightful's S-Plus software for data analysis, data mining and statistical modeling. It enables users to deploy analytical models, view custom reports and generate graphics from a Web browser or spreadsheet program such as Excel.

"To share the analysis was time consuming and most people didn't understand all the intricacies of the analysis, they just wanted the results," says Murty. "We needed something that people could just click on and get the information very quickly instead of having to come to our statistician, Chris Wilson, every time. This program allows Chris to do more of his creative work and allows all the users to get into these results as quickly as possible without having to send e-mails and do all these other things."

Murty and his analytic team are responsible for analyzing and responding to product field failures from around the globe. His team collects terabytes of data each day from field service engineers who visit customer sites, from Cymer's manufacturing and test sites and from R&D scientists who are responsible for developing products that meet stringent manufacturing specifications. The company has roughly 1,500 to 1,600 lasers in the field and each diagnostic download is a small file that is e-mailed over the Internet to a database maintained at Cymer's San Diego headquarters. Each night, an automatic program pulls the data from the day's e-mails and puts it into the database. The next morning, the Statsserver program analyzes the data looking to spot trends.

Wilson, the Cymer statistician, has programmed integrated life curves and survival curves into the data mining program. "We have a number of tests that insure that our components are up to specification. We get data from our lasers out there that we can monitor. A lot of data we get is downloaded from lasers sent to us from field service engineers," Wilson says.

The amount of data is sure to go up, as Cymer has introduced a new service for its customers called Cymer Online. This allows Cymer to collect and monitor data in real-time. Data is collected and sent every five minutes to a server.

Some of the data analysis operations that Wilson runs with Statserver are Pareto, Shewart and CuSum charts. These are used to monitor data to make sure that they are within specified standards. When looking for problems, Wilson runs various data mining analytics to uncover them. "We use a certain amount of regression analysis and we can also go into these diagnostics and do a generic plot. We can go in and plot different variables against each other and look at the performance of the lasers."

By monitoring laser performance, the company gets a good idea of the average lifetime of the laser and its components. "For one of the optical components, our analysis showed that it could go 50 percent longer than we had assumed," Murty says. "This saves costs because these expensive parts do not need to be replaced as often."

By freeing up Wilson from having to do more routine data analysis functions, he can work on more sophisticated analysis. "I can see us getting into using more things such as variance analysis," he says. "Especially if we get involved in projects that require Design of Experiments to analyze the data."

Checking up on you

Not all data mining is done by manufacturers, of course. Data mining is used in just about every field imaginable, from insurance to groceries and automobiles to power generation. One manager of a power generation plant uses data mining to check on remanufacturing work done on some equipment.

Lloyd Pentecost is a manager at Southern California Edison's San Onofre Nuclear Generating Station and uses eDNA software from InStep Software (Chicago). While the software is primarily used to track the operations of the plant by collecting up to 5,500 data points a minute, he also uses the software to check on which company has done better remanufacturing work on his motors.

"We have some motors that are 30 years old and they ran without problems Now they are going through a remanufacturing process, Pentecost says. "Data mining may be able to show that the motors that have been rebuilt by one vendor are better or worse than motors rebuilt by another manufacturer." Q

TECH TIPS

- Data mining can scour disparate data from a variety of sources to find patterns in the data.
- Data mining software allows users to drill down into data to explain known patterns, uncover hidden trends and predict future problems.
- Software can generate text reports and hundreds of graphs such as Pareto, Shewart, CuSum and others.

WEB EXTRA

What is Data Mining?

According to Herbert Edelstein, president of the consulting company Two Crows Corp. (Potomac, MD), there are two main kinds of models in data mining: predictive and descriptive. Predictive models can be used to forecast explicit values, based on patterns determined from known results. Descriptive models describe patterns in existing data and are generally used to create meaningful subgroups such as demographic clusters.

Data mining uses data to build a model of the real world. It is used to build six types of models: classification, regression, time series, clustering, association analysis, and sequence discovery.

According to A Perspective on Data Mining, written by Dr. Kenneth Collier, Dr. Bernard Carey, Ellen Grusy, Curt Marjaniemi and Donald Sautter, from the Center for Data Insight at Northern Arizona University, some of the basic types of data mining algorithms include:

- Rule association: identifies the cause and effect relationships and assigns probabilities or certainty factors to support the conclusions. Rules are in the form of if-then and can be used to make predictions or estimate unknown values.
- Memory-based reasoning or Case-based reasoning: These algorithms find the closest past analogs to a present situation in order to estimate an unknown value or predict an unknown outcome.
- Cluster analysis: Separates heterogeneous data in homogeneous and semi-homogeneous subgroups. These are based on the assumption that observations or data sets tend to be like their neighbors. Clustering increases the ability to make predictions.
- Classification Algorithms and Decision Trees: Determine natural splits in data based on a target value. First splits occur on the most significant variables. A branch in a decision tree can be viewed as the conditional side of a rule.
- Artificial Neural Networks: uses a collection of input variables, mathematical activation functions, and weightings of inputs to predict the value of target variables. Through an iterative training cycle, a neural network modifies its weights until the predicted output matches actual values. Once trained, the network is a model that can be used against new data for predictive purposes.
- Genetic algorithms: Uses a highly iterative process of selection, crossover and mutation operations to evolve successive generation of models. A fitness function is used to keep certain members and discard others. Genetic algorithms are primarily used to optimize neural network topologies and weights. However, they can be used by themselves for modeling.

Data mining is only one step in the knowledge discovery process. The other steps include identifying the problem to be solved, collecting and preparing the right data, interpreting and deploying models and monitoring the results.



Copyright © 2003 Business News Publishing Company