



Finding Direction in Chaos

Data mining methods make sense out of millions of seemingly random data points.

by Thomas Hill, Ph.D.; Robert Eames; and Sachin Lahoti

Data mining methods have many origins, including drawing on insights into learning as it naturally occurs in humans (cognitive science), and advances in computer science and algorithm design on how to best detect patterns in unstructured data.

Know & Go

- Data mining is growing increasingly popular—and necessary—in many manufacturing and continuous-process applications.
- Data mining is intended to reveal the most pertinent pieces of data from the mountains of data surrounding the process of interest.
- Industries that have proven to be fertile grounds for data mining include semiconductors, heavy manufacturing, chemicals/pharmaceuticals, and power generation.
- Data mining techniques provide tools for process optimization and quality control.

Although traditional statistical methods for analyzing data, based on statistical theories and models, are now widely accepted throughout various industries, data mining methods have only been widely embraced in business for a decade or two. However, their effectiveness for root cause analysis, and for modeling, optimizing and improving complex processes, are making data mining increasingly

popular—and even necessary—in many real-world discrete manufacturing, batch manufacturing, and continuous-process applications.

There is no single, generally agreed-upon definition of data mining. As a practical matter, whenever data describing a process are available, in manufacturing for example, then any systematic review of those data to identify useful patterns, correlations, trends, and so forth, could be called “data mining.” Put simply, data mining uncovers nuggets of information from a sometimes vast repository of data describing the process of interest.

This article will focus on specific applications for data mining to improve discrete and continuous manufacturing processes, or to implement advanced process monitoring systems.

Hypothesis testing vs. pattern recognition

Typical statistical methods used in Six Sigma projects often involve analysis of variance (ANOVA), multiple regression, quality control charting, process capability analysis, and so forth. All of these methods are derived from statistical theory and involve hypothesis testing. For example, in designed experiments



of considering and varying systematically a fixed number of factors in a manufacturing process as you would in a designed experiment, one can extract all available historical data describing the process to identify reliable patterns and factor combinations associated with a problem or optimal process performance.

Following are some examples from different industries and domains where the data mining approach and methods have become increasingly popular and successful.

Semiconductor industry

The manufacture of semiconductors involves hundreds of process steps through various semiconductor processing tools. Any problems in a single tool or combination of tools can lead to serious and costly yield problems.

A typical dataset describing the successive processing steps applied to wafer lots will record the specific machine or tool that was used, the time when it was used, and the specific processing that was applied. Each lot goes through hundreds of steps, so when a yield problem occurs, there are almost an infinite number of potential root causes, i.e., individual tools or interactions between specific tools that could be responsible for the decrease in yield.

A first step is to “sift” through all tools to identify those that appear to discriminate between low-yield and high-yield lots. Put another way, the question is which specific tools (“features”) or processing steps are common to low-yield vs. high-yield lots.

The graph seen in figure 1, right, summarizes such a feature-selection graph. The longer the bar for a particular tool shown in this plot, the greater the likelihood that the respective tool is related to the yield problem. That is, the tools associated with the longest bars are the most likely root causes.

This graph is very similar to the common Pareto chart, except that instead of identifying the frequency of known failures attributable to a particular tool, part, or processing step, the data mining algorithms are used to identify the specific tools that appear to provide the best discrimination between high- and low-yield lots.

Of course, problems often cannot be traced to a single tool but are rather caused by the combination or interaction between different factors or tools. Now the possible list of combinations of tools to consider becomes extremely large and impossible to review without the help of data mining algorithms.

Recursive partitioning

One class of data mining algorithms that is particularly well-suited to identify interactions between inputs (tools, predictors, etc.) to a process that cause undesirable outputs are the so-called recursive-partitioning or tree algorithms. There are a number of such algorithms commonly implemented in software for data mining. All algorithms will partition the data, based on the available inputs, into subgroups of observations that are increasingly homogeneous with respect to the outcome, e.g., into subgroups that are all of high yield or low yield. In effect, these algorithms will create decision trees that will identify and represent interactions between inputs in the data.

The illustration in figure 2 at the top of page 22 shows a simple tree where each split produces two child nodes. The bars inside each node (box) show the proportions of lots that are classified as high yield vs. low yield; at the root node there is approximately an equal number of each, representing the entire data set for this study. As you traverse down the

(DOE), factors are systematically varied to test the hypothesis that a specific factor or combination of factors has a statistically significant effect on the outcome variable or process of interest.

To summarize, traditional statistical data analysis methods involve hypothesis testing to identify statistically significant factors, points, deviations etc., or the testing of expectations about the process under consideration.

Data mining methods fundamentally do not rely on or test any *a priori* expectations. Instead, the data mining process can best be described as pattern recognition: A typically large amount of data are available to describe a process or specific problem, and the goal is to detect in the data the relevant patterns that allow us to improve the process. For example, instead

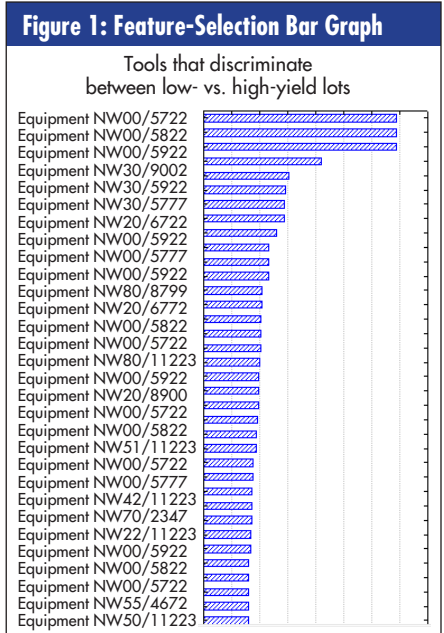
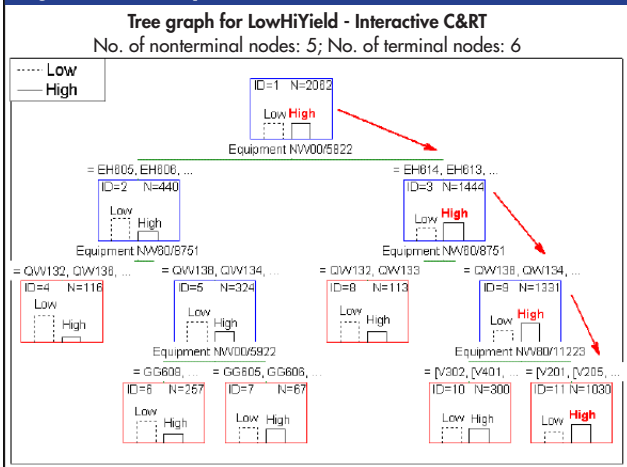


Figure 2: Tree Graph



tree, you can see that a relatively greater number of low-yield cases are partitioned on the left side of the tree (the initial split).

For example, if Equipment NW00/5822 applied the process EH605 or EH606 and Equipment NW60/8751 applied process QW132 or QW136, then many more lots were in the low-yield category than the high-yield category. On the other hand, if you follow the right-hand branch in the illustration down to the final node ID=11, then the respective split (“decision”) criteria associated with each split identify the combinations of tool steps (“commonalities”) associated with a greater chance of high yield. In node partition ID=11 there are clearly more high-yield lots than low-yield lots. Hence that branch of the tree identifies an interaction between these three pieces of equipment.

A detailed description of recursive partitioning methods is beyond the scope of this article. These data mining algorithms are, however, extremely useful for identifying and exploring complex factor interactions in complex high-dimensional data. They are therefore very useful for root cause analysis when all simpler methods have been exhausted.

General applicability

Most automated manufacturing involves the application of a large number of processing steps to move from raw materials to final product. Because data collection and storage has become very inexpensive during the last few years, comprehensive historical data typically exists describing

exactly the steps that were applied to each batch, part, lot, or machine. Data mining methods for feature selection, as described above, and recursive partitioning are generally very useful for finding the needle in the haystack, that is, identifying the specific tool or processing step, or interaction between two or more steps or tools, responsible for quality problems.

Optimize product quality

The following case study is discussed in greater detail in a previous article (Grichnik, T., Hill, T., and Seskin, M., “Predicting Quality Outcomes Through Data Mining,” *Quality Digest*, September 2006, pp. 42–47). In short, to control vibration problems during final testing, the manufacture of gas turbines requires that very tight tolerances be carefully observed. In this case, the goal was not only to identify specific root causes and problems during the complex manufacturing process, which could lead to problems in final product testing, but also to build comprehensive data-mining prediction models that would provide a clear line-of-site from the individual manufacturing steps to final product quality.

Caterpillar engineers used data mining modeling algorithms to build accurate predictive models of key performance and quality parameters, based on manufacturing data. Such prediction models—accurately describing the relationship between the inputs to the manufacturing process to the quality outcomes of the process—can then be used in computer simulation and optimization to identify specific opportunities for improving final product quality. For example, Caterpillar was able to reduce the occurrence of trim balance problems by

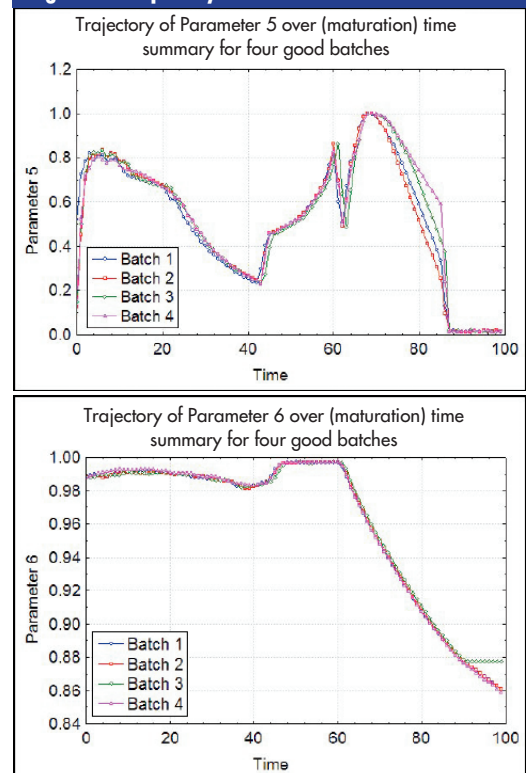
nearly 45 percent using these data mining methods.

Chemical and pharmaceutical manufacturing

The manufacture of chemicals or pharmaceuticals often involves monitoring continuous processes of raw materials batches through a lengthy and complex manufacturing process. Multiple process parameters are repeatedly measured over time as the raw materials slowly mature, combine, react, ferment, etc. into the final product.

It is obviously desirable to detect any problems in the maturation process well before final product testing. At that point, repairs, in the discrete manufacturing sense, are usually not possible, and an entire batch of vaccine must be discarded at great cost. The problem, however, is that the maturation process is reflected in the process data as gradual and nonlinear changes over time, often simultaneously in different parameters. Therefore, standard control-charting methods are not suitable for quality control because parameters are expected to be correlated and drift over time, so no fixed control or process capability limits can be set.

Figure 3: Trajectory of Parameter 5 and 6



Tracking “distance from model”

An increasingly popular and successful method of implementing effective process monitoring methods for continuous and batch processes is based on a fundamentally simple approach: First, using successful product runs or batches and the data collected during their manufacture, build accurate data mining prediction models that summarize the relationship between the process parameters and maturation (e.g., time). Then use this prediction model of time (or a proxy-measure of time) for process monitoring by charting deviations or distances of the predictions from the actual observed maturation.

The following example is based on a data set describing an industrial batch polymerization reactor (Nomikos, P., and MacGregor, J. F., “Multivariate SPC Charts for Monitoring Batch Processes,” *Technometrics*, February 1995, Vol. 37, No. 1). The batch duration is approximately two hours, divided into 100 equally spaced time intervals each. At each time interval, 10 process variables were measured.

For example, the graphs in figure 3 at the bottom of page 22 show the trajectories over time for Parameter 5 and Parameter 6, for four good batches.

Obviously, standard control charting is not applicable here. In addition, the parameter trajectories are correlated, i.e., the maturation process happens simultaneously for all parameters, each moving according to its own trajectory.

Using time as a proxy for batch maturity over time, we can use data mining methods to build a prediction model of maturity (time), that is, a model that, given the values of all the parameters, predicts at which stage of maturation the product should be. If that prediction is “bad,” i.e., if the actual time at which the measurements were taken is entirely different than that predicted according to the data mining model, then probably something is wrong with the batch, and its maturation is not compatible with that observed in the successful batches from which the model was built.

A simple control chart to implement this type of process-monitoring method would be to chart the deviations of the predictions from the observed maturation (i.e., time at which the measure-

ments were taken). If a batch of product is compatible with the maturation process of the successful batches from which the models were built, then the prediction residuals (predicted vs. residual time) will vary around a mean of zero, and standard X-and-MR charts or cumulative sum (CUSUM) charts can be used to detect when the process is shifting or drifting.

Continuous processes: power generation

The methods described here are applicable to practically all continuous- or batch-process manufacturing, when detailed data are collected describing the process. Another good example where data mining methods have been shown to be very successful is in the area of furnace and power plant optimization. Power generation from coal-fired furnaces requires the coordination of numerous complex subsystems, from coal preparation, through combustion, to the complex treatment of exhaust gases and waste materials to minimize or eliminate harmful emissions. It is not uncommon for a single unit in a power generating station to collect several thousand parameters every minute into a detailed data history, and often years of historical data are available describing all operations.

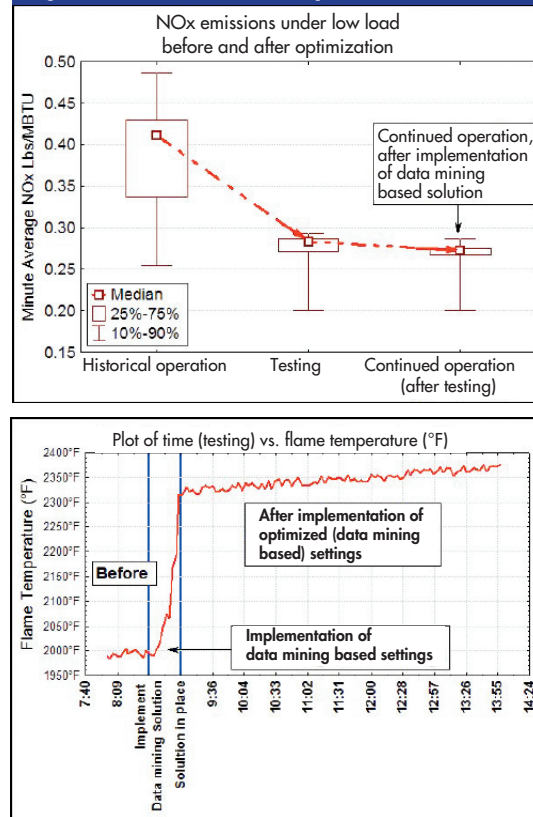
A recent study (Hill, T., *EPRI/StatSoft Project 44771: Statistical Use of Existing DCS Data for Process Optimization*, EPRI, Palo Alto, CA, 2008) provides a good illustration of how data mining models can be used to optimize simultaneously the combustion processes and emission control systems for significantly improved overall process performance. (See figure 4, above.)

Whenever a complex continuous process is well instrumented, providing detailed data of all parameters controlling the process, data mining methods can provide effective tools for optimization and control.

About the authors

Thomas Hill, Ph.D., is vice president of analytic solutions development at Stat-

Figure 4: NOx and Flame Temperature



Soft Inc. and StatSoft Power Solutions (a wholly owned subsidiary of StatSoft Inc.). Hill has more than 20 years of experience in data analysis for applied and basic research, quality and process control, organizational development, and research. He's the co-author with Pawel Lewicki of the textbook *Statistics: Methods and Applications* (StatSoft Inc., 2006).

Robert Eames is a product manager at StatSoft Inc. with a focus on enterprise analytic software solutions. He has 10 years of experience providing data analysis consulting in pharmaceutical, chemical, discrete manufacturing, and other industries.

Sachin Lahoti is senior data mining consultant at StatSoft Inc. He holds a bachelor's degree in mechanical engineering and a master's degree in management information systems and data mining. Lahoti has eight years of experience with the application of advanced data mining algorithms to solve various manufacturing and engineering problems. **QD**

Comments

Send feedback to comments@qualitydigest.com.