Automated Anomaly Detection

Brad Morantz
Georgia State University, USA

INTRODUCTION

Preparing a dataset is a very important step in data mining. If the input to the process contains problems, noise, or errors, then the results will reflect this, as well. Not all possible combinations of the data should exist, as the data represent real-world observations. Correlation is expected among the variables. If all possible combinations were represented, then there would be no knowledge to be gained from the mining process.

The goal of anomaly detection is to identify and/or remove questionable or incorrect observations. These occur because of keyboard error, measurement or recording error, human mistakes, or other causes. Using knowledge about the data, some standard statistical techniques, and a little programming, a simple data-scrubbing program can be written that identifies or removes faulty records. Duplicates can be eliminated, because they contribute no new knowledge. Real valued variables could be within measurement error or tolerance of each other, yet each could represent a unique rule. Statistically categorizing the data would eliminate or, at least, greatly reduce this.

In application of this process with actual datasets, accuracy has been increased significantly, in some cases double or more.

BACKGROUND

Data mining is an exploratory process looking for as yet unknown patterns (Westphal & Blaxton, 1998). The data represent real-world occurrences, and there is correlation among the variables. Some are principled in their construction, one event triggering another. Sometimes events occur in a certain order (Westphal & Blaxton, 1998). Not all possible combinations of the data are to be expected. If this were not the case, then we would learn nothing from this data. These methods allow us to see patterns and regularities in large datasets (Mitchell, 1999).

Credit reporting agencies have been examining large datasets of credit histories for quite some time, trying to determine rules that will help discern between problematic and responsible consumers (Mitchell, 1999). Datasets have been mined looking for indications for boiler explosion probabilities to high-risk pregnancies to consumer purchasing patterns. This is the semiotics of data, as we transform data to information and finally to knowledge.

Dirty data, or data containing errors, are a major problem in this process. The old saying is, "garbage in, garbage out" (Statsoft, 2004). Heuristic estimates are that 60-80% of the effort should go into preparing the data for mining, and only the small remaining portion actually is required for the data-mining effort itself. These data records that are deviations from the common rule are called anomalies.

Data are always dirty and have been called the curse of data mining (Berry & Linoff, 2000). Several factors can be responsible for attenuating the quality of the data, among them errors, missing values, and outliers (Webb, 2002). Missing data have many causes, varying from recording error to illegible writing to just not supplied. This is closely related to incorrect values that also can be caused by poor penmanship as well as measurement error, keypunch mistakes, different or incorrect metrics, misplaced decimal, and other similar causes.

Fuzzy definitions, where the meaning of a value is either unclear or inconsistent, are another problem (Berry & Linoff, 2000). Often, when something is being measured and recorded, mistakes happen. Even automated processes can produce dirty data (Bloom, 1998). Micro-array data has errors due to base pairs on the probe not matching correctly to genes in the test material (Shavlik et al., 2004). The sources of error are large, and it is necessary to have a process that finds these anomalies and identifies them.

In real valued datasets, the possible combinations are (almost) unlimited. A dataset with eight variables, each with four significant digits, could yield as many as 10^{32} combinations. Mining such a dataset would not only be tedious and time-consuming, but possibly could yield an overly large number of patterns. Using (six-range) categorical data, the same problem would only have 1.67×10^{6} combinations. Gauss normally distributed data can be separated into plus or minus 1, 2, or 3 sigma. Other distributions can use Chebyshev or other distributions with similar dividing points. There is no real loss of data, yet the process is greatly simplified.

Finding the potentially bad observations or records is the first problem. The second problem is what to do once they are found. In many cases it is possible to go back and verify the value, correcting it, if necessary. If this is

possible, the program should flag values that are to be verified. This may not always be possible, or it may be too expensive. Not all situations repeat within a reasonable time, if at all (i.e., observation of Halley's comet).

There are two schools of thought, the first being to substitute the mean value for the missing or wrong value. The problem with this is that it might not be a reasonable value, and it can create a new rule, one that could be false (i.e., shoe size for a giant is not average). It might introduce sample bias, as well (Berry & Linoff, 2000).

Deleting the observation is the other common solution. Quite often, in large datasets, a duplicate exists, so deleting causes no loss. The cost of improper commission is greater than that of omission. Sometimes an outlier tells a story. So, one has to be careful about deletions.

THE AUTOMATED ANOMALY DETECTION PROCESS

Methodology

To illustrate the process, a public dataset is used. This particular one is available from the University of California at Irvine Machine Learning Repository (University of California, 2003). Known as the Abalone dataset, it consists of 4,400 observations of abalones that were captured in the wild with several measurements of each one. Natural variation exists, as well as human error, both in making the measurements and in the recording. Also listed on the Web site were some studies that used the data and their results. Accuracy in the form of hit rate varied between 0-35%.

While it may seem overly simple and obvious, plotting the data is the first step. These graphical views can provide much insight into the data (Webb, 2002). The data for each variable can be plotted vs. frequency of occurrence to visually determine distribution. Combining this with knowledge of the research will help to determine the correct distribution to use for each included variable. A sum of independent terms would tend to support a Gauss normal distribution, while the product of a number of independent terms might suggest using log normal. This plotting also might suggest necessary transformations.

It is necessary to understand the acceptable range for each field. Some values obtained might not be reasonable. If there is a zero in a field, is it indicative of a missing value, or is it an acceptable value? No value is not the same as zero. Some values, while within bounds, might not be possible. It is also necessary to check for obvious mistakes, inconsistencies, or out of bounds.

Knowledge about the subject of study is necessary. From this, rules can be made. In the example of the abalone, the animal in the shell must weigh more than when it is shucked (removed from the shell) for obvious reasons. Other such rules from domain knowledge can be created (abalone.net, 2004; University of Capetown, 2004; World Aquaculture, 2004). Sometimes, they may seem too obvious, but they are effective. The rules can be programmed into a subroutine specific to the dataset.

Regression can be used to check for variables that are not statistically significant. Step-wise regression is a handy tool for identifying significant variables. Other ratio variables can be created and then checked for significance using regression. Again, domain knowledge can help create these variables, as well as insight and some luck. Insignificant variables can be deleted from the dataset, and new ones can be added.

If the dataset is real valued, it is possible that records exist that are within tolerance or measurement error of each other. There are two ways to reduce the number of unique observations. (1) Attenuate the accuracy by rounding to reduce the number of significant digits. Each variable rounding to one less significant digit reduces the number of possible patterns by an order of magnitude. (2) Calculate a mean and standard deviation for the cleaned dataset. Using an appropriate distribution, sort the values by standard deviations from the mean. Testing to see if the chosen distribution is correct is accomplished by using a Chi square test, a Kolmogorof Smirnoff test, or the empirical test. The number of standard deviations replaces the real valued data, and a simple categorical dataset will exist. This allows for simple comparisons between observations. Otherwise, records with values as little as .0001% differences would be considered unique and different. While some of the precision of the original data is lost, this process is exploratory and finds the general patterns that are in the data. This allows one to gain insight into the database using a combination of statistics and artificial intelligence (Pazzani, 2000), using human knowledge and skill as the catalyst to improve the results.

The final step before mining the data is to remove duplicates, as they add no additional information. As the collection of observations gets increasingly larger, it gets harder to introduce new experiences. This process can be incorporated into the computer program by a simple process that is similar to bubblesort. Instead of comparing to see which row is greater, it just looks for differences. If none are found, then the row is deleted.

Example Results

A few variables were plotted producing, some very unusual graphs. These were definitely not the graphs that were expected. This was the first indication that the dataset was noisy. Abalones are born in very large numbers, but with an extremely high infant mortality rate (over