Automated Anomaly Detection

By Brad Morantz

Introduction

Preparing a dataset is the single most important step in data mining. Since the data represents real world occurrences, there is correlation amongst the variables and not all possible combinations of the data are to be expected. The first step after categorizing the data would be to eliminate the duplicates. They can be eliminated because they contribute no new knowledge.

The next step is to remove the bad data 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. The author used Fortran for his studies, but other languages will work as well.

In application of this process with real world data sets, 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] In real valued data sets, the possible combinations are (almost) unlimited. A data set with 8 variables, each with 4 significant digits could yield as many as 10^{32} combinations. Mining such a data set would not only be tedious and time consuming, but could possibly yield an overly large number of patterns. Using (6 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 with similar dividing points. There is no real loss of data, yet the process is greatly simplified.

The data represents real world occurrences and there is correlation amongst the variables. Some are principled in their construction, one event triggering another. Sometimes events occur in a certain order. [Westphal & Blaxton] Not all possible combinations of the data are to be expected. If this was not the case, then we would learn nothing from this data. These methods allow us to 'see' patterns and regularities in large data sets [Mitchell]. Credit reporting agencies have long been examining large data sets of credit histories trying to determine rules that will help discern between problematic and responsible consumers [Mitchell]. This is the semiotics of data, as we transform data to information, and finally to knowledge.

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

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

Another problem is fuzzy definitions, where the meaning of value is either unclear or inconsistent. {Berry & Linoff] Whenever something is being measured and recorded, mistakes happen. There is checking for obvious mistakes, inconsistencies, or out of bounds. [Bloom] Even automated processes can produce dirty data. Microarray data has errors due to base pairs on the probe not matching correctly to genes in the test material. [Shavlik et al] The sources of error is large, and it is necessary to have a process that finds these anomalies and flags them.

The first problem is how to find the potentially bad observation or record. And the second problem is what to do after they are found. In many cases it is possible to go back and verify the value, correcting it if necessary. This is not always possible, or it may be too expensive. Not all situations repeat within a reasonable time, if at all. (i.e. Observance of Haley'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 is creating 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]

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

The Automated Anomaly Detection Process

Methodology

The data set used for this study was a public data set, available on the Internet from University of California at Irvine. [www.uci.edu] It is known as the Abalone data set, a set of 4400 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 was some studies that used the data and their results. Accuracy in the form of hit rate varied between 0 and 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] The data for each variable can be plotted versus frequency of occurrence to visually determine distribution. Combining this with knowledge of the research will help 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. Each variable is plotted against the criterion, to see the distribution. They were also plotted versus time, and using a multidimensional plotting program, were plotted together. This plotting might also suggest necessary transformations.

Once it was determined that the data set was indeed dirty, the task was to first identify the observations that indeed had one or more incorrect (or missing) data. It is first 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. In many cases, an observation that has been identified as anomalous can be checked or verified. If an error was found, the correct value can inserted into the file.

Fortran, short for ‘FORmula TRANslator’, was used as it does not require extensive programming skills and is very powerful mathematically. It was simple to create a small program that inputted the data set into a matrix, and then in order of need did the proper filtering and flagging. The majority of the program is the same for almost any data set, but it does require a domain specific subroutine to check for knowledge dependent rules.

Knowledge about the subject of study is necessary. From this, rules can be made. In the case 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. [www.abalone.net] [www.sardi.sa.gov] [www.fishtech.net] Sometimes they may seem too obvious, but they are effective. The rules were programmed into a subroutine specific to this problem.

Regression was used to check for variables that were not statistically significant. Stepwise 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 are deleted from the data set and the new ones are created. The Fortran program calculated the values of the new variables and inserted them into the file.

Since this data set has over 4400 observations, and there was no way to verify, check, or correct flagged observations, the anomalous ones were deleted. The mathematical odds are in favor of there being another correct observation in the data set.

The final step before mining this 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'. With the potentially large number of values and combinations, it was decided to apply statistical methodology to the data set. Mean and standard deviation for the cleaned data set were calculated. The data was verified to be approximately gauss normal distribution. This could be done by using the Chi square test, a Kolmogorof Smirnoff test, or for this application, the empirical test is more than adequate.

The actual real valued data were then replaced by the number of standard deviations and a simple categorical data set existed. 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. Looking for duplicates becomes simple. This is an easy Fortran process of trying to see differences and then going to the next observations if they are found. It is basically a modified bubblesort routine. This allows one to gain insight into the database using a combination of statistics and artificial intelligence [Pazzani], using human knowledge and skill as the catalyst to improve the results.

The last step is to do the actual mining. Many programs exist, but Answer Tree from SPSS was available so it was used. Because this is a classification problem, hit rate became the method of evaluating correct classification.

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 data set was 'noisy'. Abalones are born in very large numbers, but with an extremely high infant mortality rate (over 99%). This graph did not reflect that.

An initial scan of the data showed some inconsistent points, like a 5 year old infant, a shucked animal weighing more than a complete one, and things like that. Another problem with most analysis of these data sets is that gender is not ratio or ordinal data, and therefore had to be converted to dummy variables.

Stepwise regression tossed out all but 5 variables. The remaining variables were: diameter, height, whole weight, shucked weight, and viscera weight. Two new variable were created: shell ratio (whole weight divided by shell weight) and weight to diameter ratio. Since the diameter is directly proportional to volume, this variable is proportional to density. The proof of its significance was a 't' value of 39 and an F value of 1561. These are both statistically significant. A plot of shell ratio versus frequency yielded a fairly gauss normal looking curve.

As these are real valued data with 4 digits given, it is possible to have observations that vary by as little as 0.01%. This value is even less than the accuracy of the measuring instruments. In other words, there are really a relatively small number of possibilities, described by a large number of almost identical examples, some within measurement tolerance of each other.

The mean and standard deviation were calculated for each of the remaining and new variables of the data set. A simple test was done to verify approximate meeting of gauss normal distribution. Each value was then replaced by the integer number of standard deviations it is from the mean, creating a categorical data set. Simple visual inspection showed two things: 1) that there was indeed correlation among the observations; and 2) it became increasingly more difficult to introduce a new pattern.

Duplicate removal process was the next step. As expected, the first 50 observations only had 22% duplicates, but by the time the entire data set was processed, 65 percent of the records were removed because they presented no new information.

To better understand the quality of the data, least squares regression was performed. The model produced an ANOVA F value of 22.4, showing good confidence in it. But the Pearsonian correlation coefficient R^2 of only 0.25, indication that there was some problem. Visual observation of the data set and its plots led to some suspicion of the group with one ring (age = 2.5 years). OLS regression was performed on this group yielding and F of 27, but an R^2 of only 0.03. This tells us that this portion of the data is only 'muddying the water' and attenuating the performance of our model.

Upon removal of this group of observations, OLS regression was performed on the remaining data giving an even better F of 639 (showing that indeed it is a good model) and an R^2 of 0.53, an acceptable level, and one that can adequately describe the variation in the criterion.

The data is now 'clean' and ready to proceed to the data mining step. The new clean data set was converted back to Excel format and then given to Answer Tree (from SPSS, as this is the program that was available.)

The results listed at the web site where the data set was obtained are as follows:

Sam Waugh in the Computer Science Department at the University of Tasmania used this data set in 1995 for his doctoral dissertation. His results, while the first recorded attempt, did not have good accuracy at predicting the age. The problem was encoded as a classification task.

24.86%	Cascade Correlation (no hidden nodes)
26.25%	Cascade Correlation (5 hidden nodes)
21.5%	C4.5
0.0%	Linear Discriminant Analysis
3.57%	k=5 Nearest Neighbor

David Clark [Clark et al] did further work on this data set. They split the ring classification into three (3) groups: 1 to 8, 9 to 10, and 11 and up. Their results were much better as shown below.

64%	Back propagation
55%	Dystal

This reduced the number of targets, and made each one bigger, in effect, making an easier target.

The results obtained from Answer Tree using the new cleaned data set are as follows:

Hit rate Table

			Actual Category		
Predicted	**1**	**2**	**3**	**4**	**Total**
1	0	0	0	0	0
2	11	269	79	2	361
3	231	140	953	280	1604
4	22	3	27	81	133
Total	264	412	1059	363	2098

All of the 1 ring observations were filtered out in a previous step, and the extraction was 100% accurate in not predicting any as being 1 ring. The hit rates are as follows:

1 ring	100.0% correct
2 ring	74.5% correct
3 ring	59.4% correct
4 ring	60.9% correct
Overall accuracy	62.1% correct

Future Trends

Finding a data set where it would be possible to go and verify the individual observations would allow one to see how many erroneous records were found, and how many were indeed incorrect. The accuracy of the data mining could be compared both before and after the automated anomaly detection process.

A program could be written that would input the simple things like the number of variables, the number of observations, and some classification results. Then a rule input mechanism would accept the domain specific rules and make them part of the analysis. Further improvements would be the inclusion of fuzzy logic. Type I would allow the use of lingual variables (i.e. Big, small, hot, cold) in the records and type II would allow for some fuzzy overlap and fit.

Conclusion

Data mining is an exploratory process, to see what is in the data, what patterns can be found. The original data set was dirty and the real values created countless unique rules. The data set was cleaned, filtered, and categorized with accepted statistical techniques. This process, after some manual viewing and analysis, was automated in a Fortran program. Part of the program was specific to the knowledge domain of the original data, and part could be standardized. It greatly improved the regression model. Rule extraction was performed on the data. The resulting rules defined patterns found in the data, more accurate than other researchers had found not utilizing these anomaly detection and filtering techniques.

If it was possible to return to the source of the data, then the program would produce a list of observations that needed checking. Verifying and correcting, if necessary, these identified observations that were marked as anomalous could produce even more accurate results.

References

Berry, M & Linoff, G; *Mastering Data Mining The Art and Science of Customer Relationship Management,* Wiley & Sons, New York: 2000

Bloom, D; *Technology, Experimentation, and The Quality of Survey Data;* Science, Vol. 280, No. 5365, May 8, 1998

Clark, D; Schreter, Z; and Adams, A; *A Quantitative Comparison of Dystal and Backpropagation,* submitted to Australian Conference on Neural Networks (ACNN '96)

Mitchell, T; *Machine Learning and Data Mining,* Communications of the ACM, Volume 42, No. 11, November 1999, pp. 30-36

Pazzani, Michael J; *Knowledge Discovery from Data?,* IEEE Intelligent Systems, Volume 15, number 2, March/April 2000, pp. 10-13

Shavklik, Jude,; Molla, M; Waddell, M; & Page, David; *Using Machine Learning to Design and Interpret Gene-expression Microarrays*; AI Magazine, AAAI (American Association for Artificial Intelligence) Volume 25, No. 1, spring 2004, page 23 to 44.

University of California at Irvine, Machine Learning Repository

Webb, A; *Statistical Pattern Recognition*, Wiley & Sons, West Sussex England: 2002

Westphal, C & Blaxton, T; *Data Mining Solutions Methods and Tools for Solving Real-World Problems,* Wiley & Sons, New York: 1998

www.abalone.net

www.fishtech.net world wide abalone consultants

www.sardi.sa.gov a web site maintained by the South Australian Research and Development Institute

www.statsoft.com Electronic Textbook

Key Words

Anomaly - A value or observation that deviates from the rule or analogy. A potentially incorrect value.

ANOVA or analysis of variance - a powerful statistical method for studying the relationship between a response or criterion variable and a set of one or more predictor or independent variable(s)

Correlation - amount of relationship between two variables, how they change relative to each other, range: -1 to +1
F value - Fisher value, a statistical distribution, used here to indicate the probability that an ANOVA model is good. In the ANOVA calculations it is the ratio of squared variances. A large number translates to confidence in the model.

Ordinal data - data that is in order, but has no relationship between the values or to an external value.

Pearsonian correlation coefficient - defines how much of the variation in the criterion variable(s) is caused by the model. Range: 0 to 1

Ratio data - data that is in order, and has fixed spacing, a relationship between the points, and is relative to a fixed external point.

Stepwise regression - an automated procedure on statistical programs that adds one predictor variable at a time, and if it is not statistically significant, it removes it from the model. Some work in both directions, by either adding or removing from the model, one at a time.

'T' value - also called 'student's t' - a statistical distribution for smaller sample sizes. In regression routines in statistical programs it indicates whether or predictor variable is statistically significant, if it is truly contributing to the model. A value more than about 3 is required for this indication.