Applying Genetic Algorithms to Outlier Detection

Kelly D. Crawford
Amoco Corporation
kcrawford@amoco.com

Roger L. Wainwright
The University of Tulsa
rogerw@penguin.utulsa.edu

Abstract

Researchers are accustomed to inexactness in physical measurements and have developed statistical methods to help deal with error. Detecting outlying observations in regression data (outliers) is an important step in analysis of these sets of data. This paper presents a genetic algorithm capable of generating subsets for multiple-case outlier diagnostics. The genetic algorithm uses the diagnostics as evaluation functions to drive the search for good subsets. Tests run on various data sets from the statistical literature demonstrate the excellent performance of the genetic algorithm in locating the best subsets. Post-analysis of these subsets ultimately determines which sets are actually outlying. The genetic algorithm performs superbly in using multiple-case diagnostics to drive the search for these outliers.

1 INTRODUCTION

Data is rarely perfect. Whether the problem is faulty equipment, fluctuations due to wind, or simple human error, the observations we record often contain errors that can result in misleading conclusions. These errors, called outliers, must be found and removed from the data.

Various regression techniques help locate outlying data by minimizing their effect or by locating them directly. A class of techniques called multiple-case outlier diagnostics provides a way to measure the "outlyingness" of sets of points relative to the entire data set. The difficult task is selecting the right subsets to test. The combinatorial nature of this problem requires a heuristic solution.

Section 2 discusses ways the outlier problem is handled in statistics. In Section 3, a genetic algorithm is presented that generates subsets of potentially outlying points in regression data. Tests on real data are shown in Section 4 followed by results of those tests in Section 5.

2 OUTLIERS

2.1 LEAST SQUARES

Least sum of squares, more commonly called least squares (LS), is a regression technique used to estimate a line (or hyperplane) fit to a set of data. Least squares regression constructs a line through data such that the sum of the squared residuals (distances from each point to the line in the y-direction) is minimized. Many good references for the least squares technique are available, such as (Cheney 1980, Hager 1988).

The data points are represented by the overdetermined system $y = X\theta + \epsilon$, where n is the number of data points, y is the n vector of dependent variables, X is the $n \times p-1$ matrix of $p-1$ independent variables, θ is the $p-1$ vector of coefficients and ϵ is the n vector of independent errors. Taking $\hat{y} = y - \epsilon$, a least squares solution to the system is $X^tX\theta = X^t\hat{y}$. This is equivalent to minimizing the residual error as follows:

$$LS = minimize \sum_{i=1}^{n} \epsilon^2$$

As an example, consider the Hertzsprung-Russell diagram of the CYG OB1 star cluster, an often analyzed dataset taken from the field of astronomy (Rousseeuw and Leroy 1987). For reference, we will call this the CYG dataset. The data, plotted in Figure 1, is taken from 47 stars in the direction of Cygnus.

The x-axis (shown numerically in reverse) represents the log of the effective temperature at the surface of each star, while the y-axis shows the log of the star's light intensity. There are two lines on this plot. The line labeled LS was produced from a least squares fit of the entire dataset.

Notice the 4 points in the upper right corner of the

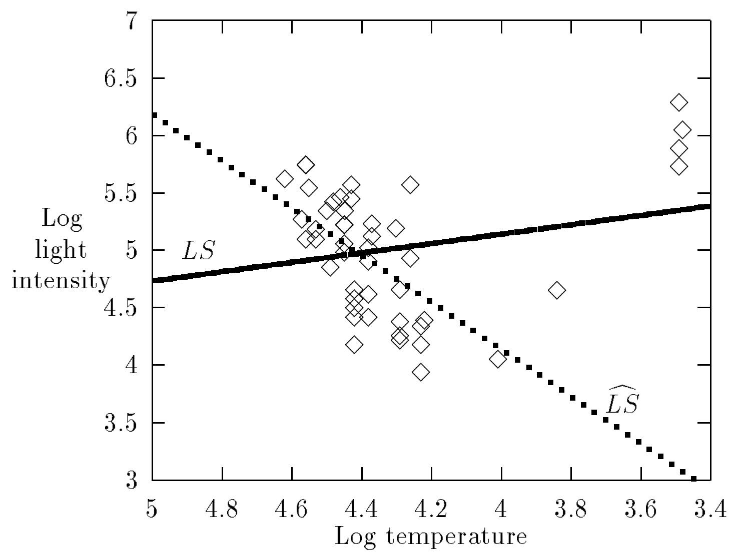

Figure 1: Hertzsprung-Russell diagram for CYG OB1 star cluster

plot. They correspond to giant stars whose measurements do not correlate linearly with those of the other stars. These points are considered to be outliers. Removing these points and recomputing the least squares line results in the line labeled $\hat{LS}$. Notice in this example how the original least squares line was dramatically affected by the outlying points and provided no help in finding any outliers.

2.2 ROBUST TECHNIQUES

Robust techniques are so named because they attempt to minimize the effect of outliers. One of the best examples of this is least median of squares (LMS) (Rousseeuw 1984). With this technique, the median of the squared residuals is minimized rather than the sum. This works well because for any line under consideration, only the point corresponding to the median of the sorted squared residuals is used. This results in throwing out the high and low residual values that might tend to throw least squares off of the trail.

Many other robust estimators are available, such as the least absolute values estimator, least trimmed squares, weighted and reweighted least squares, M-estimators and S-estimators. Rousseeuw and Leroy (1987) provide an excellent history of the development of these techniques, together with a discussion of the various strengths and weaknesses of each.

2.3 OUTLIER DIAGNOSTICS

This paper focuses on a class of techniques called outlier diagnostics. Rather than minimizing the effect of outlying data, outlier diagnostics try to assess the influence of each point in the dataset. Single-case diagnostics focus on the influence of individual points, while multiple-case diagnostics work with sets of points.

The least squares technique, described in Section 2.1, can be thought of as a simple, multiple-case outlier diagnostic technique. Thus, $LS(I)$ is the least squares of the dataset with the subset of points I removed (Crawford, Wainwright and Vasicek 1995). We will refer to this measure as LS.

Cook's squared distance formula for multiple-case diagnostics (Cook and Weisberg 1982) is a generalization of the single-case version (Cook 1977). Details of the formula are left to the references, but in short, the formula is

$$CD^2(I) = \frac{(\hat{\theta} - \hat{\theta}(I))^t X^t X(\hat{\theta} - \hat{\theta}(I))}{ps^2}$$

where $\hat{\theta}$ is the line equation formed by least squares, and $\hat{\theta}(I)$ is the least squares line formed after the subset of points I is removed from the dataset. X is the overdetermined matrix constructed from the dataset, p is effectively the dimension of the dataset, and s^2 is the sum of the squared residuals divided by the reciprocal of the degrees of freedom ($\frac{1}{n-p}$). We will refer to this measure as CD.

Andrews and Pregibon (1978) developed a multiple-case outlier diagnostic based on the following determinantal ratio

$$AP(I) = \frac{det(Z^t(I)Z(I))}{det(Z^tZ)}$$

where Z is the X matrix with the response variable y appended, and $Z(I)$ is Z with the subset of points I removed. Geometrically, $AP(I)$ is a measure of the "remoteness" of the subset of points I. We will refer to this measure as AP.

The outlyingness of selected subsets of points from the dataset can be estimated using LS, CD and AP. The difficult part is figuring out which subsets to evaluate. All subsets of points will need to be checked unless there is some mechanism to choose appropriate subsets. This is a fundamental combinatorial problem shared by all such multiple-case outlier diagnostics (Rousseeuw and Leroy 1987). Consider a dataset of size n. Checking all sets of $1, 2, \cdots, \lfloor \frac{n}{2} \rfloor$ involves computation on the order of

$$\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i}.$$

3 A GENETIC ALGORITHM FOR GENERATING SUBSETS

We developed a genetic algorithm (GA) to manage the selection of point subsets. Given a dataset containing n points, we need to be able to represent a subset of k points as a potential outlier set. We use a unique list of k point indices to accomplish this. Logically, this can be considered as an order-based chromosome where the leftmost k point indices are the outliers, and the remaining $n-k$ values are non-outlying. However, in practice, we only keep the k outlying point indices in the chromosome. This results in much faster execution for the larger datasets.

The crossover used in our GA is a variant of two-parent uniform order-based crossover (UOX) (Davis 1991). At each position in the string, a value is randomly chosen from one of the two parents to form the new children. Duplicate values are removed by selecting a new unique random value. For mutation, we randomly choose a point in the chromosome and exchange it with a unique random value. In addition, we sort the point indices in the chromosome for improved performance.

We compared three different evaluation functions. Each is based on the multiple-case outlier diagnostics formulas LS, CD and AP, previously described in Section 2.1 and Section 2.3.

4 TEST DATA

We tested 5 different datasets with the LS, CD and AP versions of the genetic algorithm. We first ran tests for $k = 2$. Then we ran tests for $k = 3$, separately, and so on, up to $k = \lfloor \frac{n}{2} \rfloor$.

We will refer to the datasets as *belgium*, *brain*, *china*, *cyg* (described in Section 2), and *siegel*. The belgium dataset shows the number of international calls from Belgium between 1950 and 1973. The brain dataset contrasts body weight against the brain weight of 18 different animal species. The china dataset is a record of the annual rates of growth of average prices in the main free cities of free China from 1940 to 1948. The siegel dataset is an exact-fit dataset. In other words, when the outliers are removed, all of the remaining points will exactly lie on a straight line. Table 1 provides detail on the size and dimension of each dataset, as well as how many known outliers each contains.

Table 1: GA test data

DATASET NAME	NUMBER OF POINTS	NUMBER OF OUTLIERS
belgium	24	6
brain	28	3
china	9	2
cyg	47	4
siegel	9	3

For each dataset, the goal is to remove any outlying data points and find a linear fit for the data. Each of these datasets have been analyzed thoroughly in the statistical literature, and each have known outliers. For a complete description of each dataset, see Rousseeuw and Leroy (1987).

We used the LibGA package to construct our genetic algorithm (Corcoran and Wainwright 1993). Each dataset was tested with 100 different random seeds. The GA parameters are: representation = integers, string length = k (separate fixed length runs for $k = 2, ..., \lfloor \frac{n}{2} \rfloor$), population size = 100, selection bias = 1.8, mutation rate = 0.05, and maximum number of iterations = 10000.

5 RESULTS

Table 2 shows the performance of each method on the test datasets in finding the one single subset containing the actual outliers. "% of time outliers found" refers to how many of the 100 test runs actually found the best subset, while the "average # iters" indicates how many subsets were evaluated before the best one was found.

Overall, CD performed best, with the exception of the siegel dataset. In this dataset, CD ranked a subset with two outlying points and one nonoutlying point higher than the fitness of the correct three points. Thus, while the GA worked correctly, CD led us to an incorrect answer (in fact, the GA found this dataset

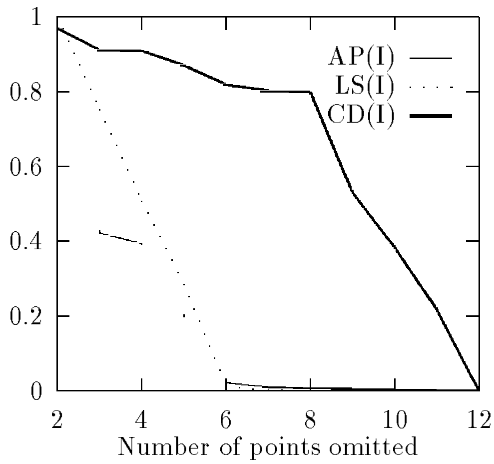

Figure 2: Belgium dataset

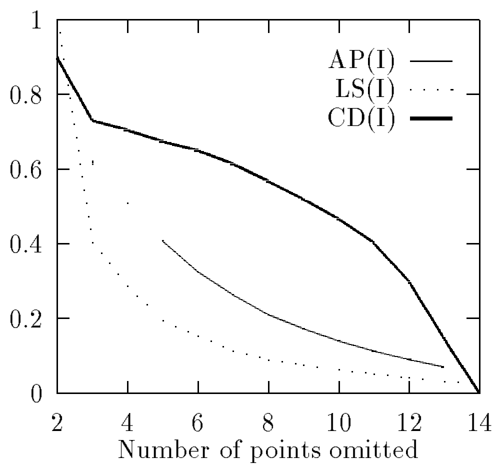

Figure 3: Brain dataset

for all 100 runs). The most likely explanation is that CD does not handle exact-fit datasets properly.

Normalized plots of the GA results are used to locate the most likely outlier set. The GA locates the best subset of size k (call it I_k), for $k = 2, ..., \lfloor \frac{n}{2} \rfloor$. The normalized values of $LS(I_k)$, $CD(I_k)$ and $AP(I_k)$ are plotted as three lines in Figure 2, Figure 3 and Figure 4 for the belgium, brain and cyg datasets, respectively. Relatively small y-axis values indicate a better fit for the data. A perfect fit, where all points in the set are exactly on a single line (or hyperplane), would have a y value of 0.

A relatively large slope change in the curves shown in Figure 2, Figure 3 and Figure 4 indicates a probable outlier set. Note the relative slope change in the LS and AP lines when 6 points are omitted for belgium (Figure 2), 3 points are omitted for brain (Figure 3) and 4 points are omitted for cyg (Figure 4). These correspond directly with the actual outliers. CD proves disappointing for the purposes of post-analysis.

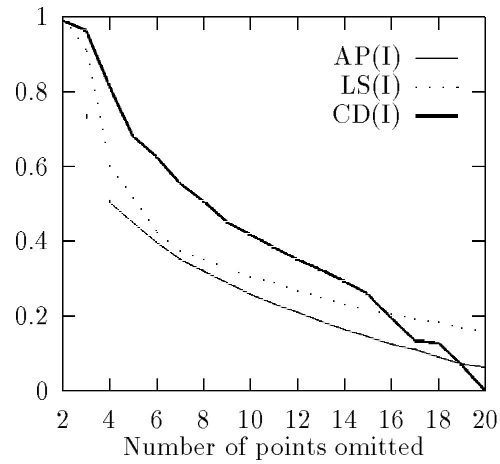

Figure 4: CYG dataset

6 CONCLUSIONS

Our findings indicate that genetic algorithms provide a useful means for generating subsets for multiple-case outlier diagnostics. The ability of genetic algorithms to overcome combinatorial search spaces makes it a prime candidate for this sort of research. It is the understanding of the authors that multiple-case outlier diagnostics are not widely used on large data sets because of this limitation. Our genetic algorithm makes such analysis feasible.

Each multiple-case diagnostic has strengths and weaknesses. Different data models and data sets produce different "fitness landscapes". As in robust regression and outlier diagnostics, there is no catch-all formula for detecting outliers. The GA, however, performed superbly in using multiple-case diagnostics to drive the search for outliers. Further research with other multiple-case outlier diagnostic techniques will certainly offer additional insights into this intriguing problem.

Acknowledgements

This research has been partially supported by OCAST Grant AR2-004. The authors also wish to acknowledge the support of Sun Microsystems, Inc.

Table 2: GA results

DATASET NAME	EVALUATION FUNCTION	% OF TIME OUTLIERS FOUND	AVERAGE # ITERS (100 RUNS)
belgium	LS	96	769
	CD	91	1526
	AP	81	4846
brain	LS	100	245
	CD	100	217
	AP	100	391
china	LS	100	217
	CD	100	110
	AP	100	100
cyg	LS	87	2925
	CD	100	323
	AP	100	759
siegel	LS	100	261
	CD	0	195
	AP	96	2438

References

D. F. Andrews and D. Pregibon (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society B*, 40:85–93.

W. Cheney and D. Kincaid (1980). *Numerical Mathematics and Computing*. Brooks/Cole, Monterey, California.

R. D. Cook (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

R. D. Cook and S. Weisberg (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

A. L. Corcoran and R. L. Wainwright (1993). LibGA: A user-friendly workbench for order-based genetic algorithm research. In *Proceedings of the 1993 ACM/SIGAPP Symposium on Applied Computing*, pages 111–117.

K. D. Crawford, R. L. Wainwright and D. J. Vasicek (1995). Detecting multiple outliers in regression data using genetic algorithms. In *Proceedings of the 1995 ACM/SIGAPP Symposium on Applied Computing*.

Lawrence Davis (1991). *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, NY.

W. W. Hager (1988). *Applied Numerical Linear Algebra*. Prentice Hall, Englewood Cliffs, New Jersey.

P. J. Rousseeuw (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

P. J. Rousseeuw and A. M. Leroy (1987). *Robust Regression & Outlier Detection*. John Wiley & Sons, New York, NY.