Computers and Chemical Engineering 35 (2011) 388–390

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Note

Outlier detection in large data sets

Guido Buzzi-Ferraris*, Flavio Manenti

Dipartimento di Chimica, Materiali e Ingegneria Chimica "Giulio Natta", Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

ARTICLE INFO

Article history:
Received 3 September 2009
Received in revised form
15 November 2010
Accepted 19 November 2010
Available online 27 November 2010

Keywords:
Outliers
Reliable parameter estimation
Robustness
Large data sets

ABSTRACT

In this paper we propose a method for correctly detecting outliers based on a new technique developed to simultaneously evaluate mean, variance and outliers. This method is capable of self-regulating its robustness to suit the experimental data set under analysis, so as to overcome shortcomings of: (i) non-robust methods such as the least sum of squares; (ii) the need of the user in defining a trimmed sub-set of experimental points such as in least trimmed sum of squares; and (iii) the possibility to read the data set only once to evaluate the mean, variance, and outliers of a population by preserving robustness.

1. Introduction

Outliers are a well-known problem in all experimental scientific and industrial fields. Many techniques for and approaches to detecting and identifying them have been proposed in the scientific literature.

Identifying outliers is particularly difficult and it is no coincidence that many robust methods have been proposed (Buzzi-Ferraris & Manenti, 2010; Draper & Smith, 1998; Rousseeuw & Leroy, 1987; Rousseeuw, 1984; Ryan, 2009; Seber & Wild, 2003) since certain well-established methods are not very effective in detecting outliers, while certain other methods, although robust enough, may require too much data analysis time if the data set itself is particularly large.

Large data sets are typical of process industries, where each measured process variable usually has a corresponding vector consisting of millions of elements stored in the historical database on the distributed control system (DCS). When the data set is very large, some nonrobust estimators are particularly performing at estimating parameters, but may provide biased values if the data set is affected by outliers. The computation of robust estimates, on the other hand, is much more computationally intensive. Even though this latter problem has become less relevant as computing power has exponentially increased in recent years, some effective robust methods cannot be used to analyze large data sets typically seen in the process industry.

2. Clever mean, clever variance and outlier detection

The general problem broached in this short paper is the detection of possible outliers in a set of n of experimental points $y_i (i = 1, \ldots, n)$ of the population Y with n being very large. The population must have a unimodal symmetrical distribution (for example Gaussian distribution) and with an expected value μ and variance σ^2 even though such parameters are unknown. Populations that are asymmetrically distributed or with multimodal distribution cannot undergo outlier detection analysis.

We propose a novel robust criterion that allows the mean and variance to be estimated while at the same time detecting possible outliers. These new estimators are called the *clever mean* and *clever variance*, respectively. Their estimation is quite trivial; when the data set is being read, the following quantities are calculated:

$$\text{sum} = \sum_{i=1}^{n} y_i \tag{1}$$

$$sq = \sum_{i=1}^{n} y_i^2 \tag{2}$$

and a predetermined number of maximum and minimum values is collected.

Let:

$$cm_0 = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} \tag{3}$$

* Corresponding author. Tel.: +39 02 2399 3257; fax: +39 02 7063 8173.
E-mail address: guido.buzziferraris@polimi.it (G. Buzzi-Ferraris).

0098-1354/$ – see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.compchemeng.2010.11.004

denote the zeroth-order clever mean and

$$cv_0 = \frac{\sum_{i=1}^{n}(y_i - cm_0)^2}{n-1} = s^2 \tag{4}$$

the zeroth-order clever variance. Assuming y_1^* as the first possible outlier and removing that value using the mean and variance, results in:

$$cm_1 = \frac{\sum_{i=1}^{n} y_i - y_1^*}{n-1} \tag{5}$$

$$cv_1 = \frac{\sum_{i=1}^{n}(y_i - cm_1)^2 - (y_1^* - cm_1)^2}{n-2} = \frac{sq + n(cm_1)^2 - 2cm_1\,\mathrm{sum} - (y_1^* - cm_1)^2}{n-2} \tag{6}$$

If:

$$\left|cm_1 - y_1^*\right| > \delta\sqrt{cv_1} \tag{7}$$

where δ is a threshold value (*i.e.*, 2.5), the experiment y_1^* can be considered an outlier and the values of cm_1 and cv_1 are an estimation of the first-order clever mean and clever variance, respectively.

If an outlier exists, the procedure is iterated: a new possible outlier y_2^* is selected and cm_2 and cv_2 are both calculated also by simulating the removal of this new point y_2^*; if the elimination of this point satisfies the relation:

$$\left|cm_2 - y_2^*\right| > \delta\sqrt{cv_2} \tag{8}$$

it too has to be considered an outlier. The procedure goes on until y_k^* satisfies the condition:

$$\left|cm_k - y_k^*\right| > \delta\sqrt{cv_k} \tag{9}$$

when its removal is simulated, whereas y_{k+1}^* does not:

$$\left|cm_{k+1} - y_{k+1}^*\right| < \delta\sqrt{cv_{k+1}} \tag{10}$$

Please note:

- The selection of a possible outlier y_k^*, is very simple indeed: it is whichever one that minimizes cv_k between the two observations that currently represent the minimum and the maximum values after the removal of previous outliers.
- The clever mean might maintain its value while outliers are progressively removed for two reasons. Firstly, when there is a particularly large set of data, the arithmetic mean can change slightly even though an outlier is removed. Secondly, if two outliers that are symmetric with respect to the expected value are removed, the clever mean remains unchanged. It would be an error to check for outliers only by looking at the value of the clever mean.
- The clever variance, on the other hand, has a monotone decreasing trend as outliers are gradually removed; it is, moreover, practically unvaried or further increases when the observation removed is not a real outlier.

If the clever variance does not increase when the observation y_k^* is removed and if y_k^* satisfies the relation (9), y_k^* is an outlier.

3. Comparison with other techniques used to detect outliers in a population

As already demonstrated by Rousseeuw (1984), the procedure traditionally adopted to identify the outliers is inadequate and its use should be avoided.

The traditional procedure consists of the following steps: firstly, the expected value μ is evaluated using the arithmetic mean $\bar{y} = \sum_{i=1}^{n} y_i/n$; the variance σ^2 is calculated by means of the expression $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/n - 1$. Then, standardized residuals are calculated:

$$r_i = \frac{y_i - \bar{y}}{s} \tag{11}$$

If a standardized residual exceeds a specific threshold value, usually 2.5, the corresponding experimental point is considered suspicious.

As both the arithmetic mean in general and the standard deviation in particular are nonrobust estimators, they are biased even though there is only a single outlier.

Robust alternatives for evaluating the expected value are the median, remedian, and trimmed mean. The median of a finite list of numbers is the middle value obtained by sorting all the observations from the lowest to the highest values. The remedian (Rousseeuw, 1984; Rousseeuw & Leroy, 1987) is usually adopted for large data sets and is basically the median of a series of submedians. The trimmed mean (Rousseeuw & Van Driessen, 2006) is the arithmetic mean evaluated after discarding a given part of the lowest and the highest values.

A robust alternative for estimating the standard deviation is the *mad*, where *mad* stands for median absolute deviations (Rousseeuw & Leroy, 1987). It is defined as:

$$mad = 1.253313\,median\left|y_i - median(y_i)\right| \tag{12}$$

and corresponds to the standard deviation, but using the median rather than the mean.

The most important advantage of the median and *mad* is their robustness. Nevertheless, even they do have some shortcomings. Firstly, they require data to be sorted from the smallest value: if the experimental data set is so large that it is stored as an ASCII file, then that file will have to be read many times making the procedure particularly time-consuming, sometimes prohibitively so. Secondly, both the estimators are less efficient if compared to the arithmetic mean and the variance obtained by the mean. It is worth stressing that the efficiency of an estimator is based on the variance of that estimator. The arithmetic mean has a variance related to s^2/n, whereas the median to $\pi s^2/2n$ and therefore the mean is more efficient than the median. In particular, estimations obtained using the median or the *mad* can be inaccurate when all the outliers are positioned on one side of the two extremities of the experimental data.

The most important advantage of the remedian is that even if the data set is stored as an ASCII file, the latter only needs to be read once; unfortunately, however, this estimator is less efficient than the median.

The trimmed mean estimator, on the other hand, is more efficient than the median and remedian if the experimental points omitted in their evaluation are all real outliers. Its most significant shortcomings are that its estimation requires the sorting of data as per the median and also makes defining a proper trimmed percentage difficult. It is important to remember too that in trimmed mean evaluation, the same number of elements is trimmed for both the largest and the smallest values as this estimator acts symmetrically on the sorted data set. Hence, if a series of outliers, all of which are either very large or very small, affects the experimental data set, an equivalent amount of good points is also trimmed at the other extreme and the resulting estimation of the trimmed mean is less efficient than the arithmetic mean obtained by omitting the real outliers only.

The main advantage of the clever mean and clever variance is that, even though these estimators are robust, they are also very

efficient (low variance) as they are just as efficient as traditional estimators when no outliers are included in the data set.

The clever mean and clever variance are particularly useful when there is a large number of observations. In fact, unlike the other alternatives, it is possible to get a robust estimation of the mean and variance with a single reading of the data set while simultaneously detecting the presence of outliers.

This procedure also bypasses the need to define an appropriate level of robustness compared to the one required by the user (trimmed percentage) to implement the trimmed mean: the clever mean and clever variance self-regulate their own robustness according to the data set being processed.

Another benefit of the clever mean over the trimmed mean is that the former does not remove the same number of largest and smallest values (the removal may be asymmetric). The mean is therefore evaluated by excluding real outliers but without sacrificing any good observations. For example, if there were m outliers all with large values, the trimmed mean would remove even the smallest m values even though they are good observations.

As a simple comparative example, let us consider the following set of measures (the data set does not contain any outlier):

$$\mathbf{y} = \{31.1, 31.2, 31.2, 31.3, 31.3, 31.1, 31.4, 31.3, 31.0\} \quad (13)$$

The arithmetic mean is 31.2111 and the variance 0.01611.

If three outliers are added to the original data set (13), giving:

$$\mathbf{y} = \{31.1, 31.6, 31.2, 31.2, 31.3, 311.1, 31.3, 31.1, 31.4, 31.3, \mathbf{32.1}, 31.0\} \quad (14)$$

the following results are obtained:

- $\bar{y} = 54.642$
- Median: 31.3
- $cm_0 = 54.642$, $cm_1 = 31.327$, $cm_2 = 31.25$, $cm_3 = 31.2111$
- $s^2 = 6522.8$
- $(mad)^2$: 0.049457
- $cv_0 = 6522.8$, $cv_1 = 0.09218$, $cv_2 = 0.02944$, $cv_3 = 0.01611$

Outliers sorted by relevance are the no. 6, 11, and 2.

Standard residuals obtained using the mean $\bar{y}$ and the variance s^2 highlight only the point no. 6 as outlier. Robust residuals obtained by the median and the *mad* identify the points no. 6 and no. 11 as outliers, but not the point no. 2. This happens since *mad* is inaccurate.

Note that the clever mean and clever variance preserve their values in both the cases and coincide with the traditional variance estimation of the data set without any outliers (13). Classical mean and variance estimators result in erroneous values when outliers are present as they are nonrobust estimators; the median and $(mad)^2$ are robust estimators, but less accurate than the clever mean and clever variance; the $(mad)^2$, in particular, is far less efficient and accurate than the clever variance.

This example involves a small data set and either the median or the trimmed mean and the mad can be used to detect outliers; conversely, when the data set is so large that it must be collected on a file, the evaluation of the median or the trimmed mean and of the mad is computationally too heavy.

Suppose we have a population with $\mu = 24.3376$ and $\sigma^2 = 197.5403$ and that we generated 10^{10} random numbers; in addition, suppose we introduced the following 4 outliers:

No.:	15	Value: 562.95
No.:	153	Value: −6488.79
No.:	1500	Value: 10912.88
No.:	9532	Value: 67.86

While the median and the trimmed mean are computationally too heavy, the clever mean requires relatively few additional seconds to the single file reading time to properly detect outliers and give their sequence in order of relevance: outliers on no. 1500, 153, 15, and 9532.

Removing the outliers from the most relevant, the clever mean is: 24.337008, 24.337659, 24.337606, and 24.337601. Analogously, the clever variance gradually decreases: 201.81148, 197.56941, 197.54042, and 197.54025. Note that the clever variance is also more sensitive than clever mean to the outlier removal for large data sets.

4. Conclusions

This short note deals with the problem of outlier detection in large data sets. A novel robust method designed to efficiently evaluate mean, variance, and outliers simultaneously is proposed and compared to some of the most popular outlier identification methods.

References

Buzzi-Ferraris, G., & Manenti, F. (2010). *Interpolation and regression models for the chemical engineer: Solving numerical problems*. Weinheim, Germany: Wiley-VCH.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd Ed.). New York: Wiley.

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

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, US: John Wiley & Sons.

Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, *12*(1), 29–45.

Ryan, T. P. (2009). *Modern regression methods* (2nd Ed.). NJ: Wiley.

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear regression*. John Wiley and Sons.