2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

Robust Estimators of Scale and Location Parameter Using Matlab

Gandhiya Vendhan. S [1] and K.K. Suresh [2] [+]

[1] Department of Information Technology, Anna University, India.
[2] Director, School of Mathematics and Statistics, Bharathiar University, Coimbatore, India.
Corresponding author. [1]Research Scholar, Department of Statistics, Bharathiar University, India,

Abstract. In this paper an attempt has been made to analyze the data for the specific distributions and estimated the values of robust estimators under scale, shape and location parameter using MATLAB software. The concept of outliers and their effect on these estimators have also discussed in the scale and location parameter which are used in the Normal, Exponential and Poisson distribution.

Keywords: Robust Estimator, Outlier Analysis, MATLAB Software.

1. Introduction

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than more regularly occurring ones. The analysis of outlier data is referred to as outlier mining. Outliers may be detected using statistical tests that assume distribution or probability model for the data or using distance measures where objects that has substantial distance from any other cluster which are considered as outliers. Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group. Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase or purchase frequency. In general, concept description, association and correlation analysis, classification, prediction and clustering mine data regularities, rejecting outliers as noise. These methods may also help detect outliers. If one has to take the square root of the mean square error, the resulting error measure is called the root mean square error. The above minimum and maximum measures represent two extremes in measuring the distance between clusters. They tend to be overly sensitive to outliers or noisy data. The use of mean or average distance is a compromise between the minimum and maximum distances and overcomes the outlier sensitivity problem. The Conference is a primary international forum for scientists and technicians working on topics relating to computer and applications. It will provide an excellent environment for participants to meet fellow academic professionals, to enhance communication, exchange and cooperation of most recent research, education and application on relevant fields. It will bring you new like-minded researchers to have fresh idea. It also provides a friendly platform for academics and application professionals from crossing fields to communication together.

2. Outlier Analysis

[+] Corresponding author.
E-mail address: (gandhiyavendhan@yahoo.com, sureshkk1@rediffmail.com.).

Robust statistics is the stability theory of statistical procedures. Many researchers have worked in this field and described the method for evaluating robust estimators. This process was considered by Gauss and unknown error distribution. The normal distribution was assumed to follow error distribution and deeper justifications other than simplicity of the method of least squares. The central limit theorem, being a limit theorem, only suggests approximate normality under well-specified conditions in real, show that typical error distributions of high-quality data are slightly but clearly longer tailed with higher kurtosis or standardized 4th moment than the normal. Gauss had been careful to talk about observations of equal accuracy.

2.1. Statistical Distribution-Based Outlier Detection

Contributions to the congress are the statistical distribution-based approach to outlier detection assumes a distribution or probability model for the given data set using normal or Poisson distribution and then identifies outliers with respect to the model using a discordancy test. Application of the test requires knowledge of data set parameters such as assumed data distribution, knowledge of distribution parameters such as mean and variance, and the expected number of outliers. Significance probability, $SP(vi) = Prob(T > vi)$, is evaluated. If $SP(vi)$ is sufficiently small, then $\boldsymbol{o}_i$ is discordant and the working hypothesis is rejected. An alternative hypothesis, H, which states that $\boldsymbol{o}_i$ comes from another distribution model, G, is adopted. The result is very much dependent on which model F is chosen because $\boldsymbol{oi}$ may be an outlier under one model and a perfectly valid value under another. A major drawback is that most tests are for single attributes, yet many data mining problems require finding outliers in multidimensional space. Moreover, the statistical approach requires knowledge about parameters of the data set, such as data distribution. However, in many cases, the data distribution may not be known. Statistical methods do not guarantee that all outliers will be found for the cases where no specific test was developed, or where the observed distribution cannot be adequately modelled with any standard distribution.

2.2. Distance-Based Outlier Detection

The notion of distance-based outliers was introduced to counter the main limitations imposed by statistical methods. In other words, rather than relying on statistical tests, one can think of distance-based outliers as those objects that do not have enough neighbours, where neighbours are defined based on distance from the given object. In comparison with statistical-based methods, distance based outlier detection generalizes the ideas behind discordancy testing for various standard distributions. Several efficient algorithms for mining distance-based outliers have been developed. These are outlined as follows.

Index-based algorithm: Given a data set, the index-based algorithm uses multidimensional indexing structures, such as, Let M be the maximum number of objects within the dmin neighbourhood of an outlier. The index-based algorithm scales well as k increases.

Nested-loop algorithm: The nested-loop algorithm has the same computational complexity as the index based algorithm but avoids index structure construction and tries to minimize the number of I/Os.

Cell-based algorithm: To avoid $O(n2)$ computational complexity, a cell-based algorithm was developed for memory-resident data sets. Its complexity is $O(ck+n)$, where c is a constant depending on the number of cells and k is the dimensionality. In this method, the data space is partitioned into cells with a side length equal to dmin 2p*k*. Distance-based outlier detection requires the user to set both the pct and dmin parameters. Finding suitable settings for these parameters which can involve much trial and error.

2.3. Density-Based Local Outlier Detection

Statistical and distance-based outlier detection both depend on the overall or “global” distribution of the given set of data points, D.

Sequence Exception Technique: The sequential exception technique simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects. It uses implicit redundancy of the data.

Local outlier factor (LOF) is an interesting notion for the discovery of local outliers in an environment where data objects are distributed rather unevenly. However, its performance should be further improved in order to efficiently discover local outliers. The statistical approach and discordancy tests are described in

Barnett and Lewis. Distance-based outlier detection is described in Knorr and Ng. The detection of density based local outliers was proposed by Breunig,Kriegel, Ng, and Sander. Outlier detection for high-dimensional data is studied by Aggarwal and Yu. The sequential problem approach to deviation based outlier detection was introduced in Arning, Agrawal, and Raghavan. Sarawagi, Agrawal, and Megiddo introduced a discovery-driven method for identifying exceptions in large multidimensional data using OLAP data cubes. Jagadish, Koudas, and Muthukrishnan introduced an efficient method for mining deviants in time-series databases. It also provides a friendly platform for academic and application professionals from crossing fields to communication together.

3. Coverage Estimators

3.1. Θ notation

Contributions to the congress are Θ- Notation bounds a function to within constant factors. we write $f(n) = \Theta(g(n))$ if there exist positive constant n_0, c_1 and c_2 such that o the right of n_0, the value of $f(n)$ always lies between $c_1\, g(n)$ and $c_2\, g(n)$ inclusive.

3.2. O notation

The Θ-notation asymptotically bounds a function from above and below. When we have only an asymptotically upper bound, we use O- notation. For a given function g(n), we denote by O(g(n)) the set of function.

$O(g(n)) = \{ f(n)$: there exist positive constant n_0 and c such that $O \leq f(n) \leq c\, g(n)$ for all $n \geq n_0 \}$

3.3. Ω notation

All papers must be Just as O- notation provides asymptotically upper bounds on a function. Ω-notation provide an asymptotically lower bound, one use O- notation. For a given function g(n), we denote by O(g(n)) the set of function

$O(g(n)) = \{ f(n)$: there exist positive constant n_0 and c such that $O \leq c\, g(n) \leq f(n)$ for all $n \geq n_0 \}$ fellow academic professionals, to enhance communication, exchange and cooperation of most recent research, education and application on relevant fields. It will bring you new like-minded researchers, fresh idea. It also provides a friendly platform for academic and application professionals from crossing fields to communication together.

4. Statistical Inference of Scale, Shape and Location Parameters

Different types outliers can be discerned scale,shape and location model as categories of outliers can be considered (1) normal distributions points which are observation isolated from the major part of the observation in the data. (2) Exponential distribution points which addition to being isolated from the major part of X deviated strongly from the robust estimation model defined the range other observation.(3) outliers that are not leverage points of the Poisson distribution residuals in collaborations and therefore referred to as high residual outliers. In robust analysis the good leverage points are usually denoted as outliers as there are not detrimental to Poisson distribution model but merely reflect unfortunate design these types of outliers can occur both during model fitting an prediction. Outlier is the major problem of the every situation into analysis the data. In this paper a study taking for several distributions is considered.

Statistical inference concerns the problem of inferring properties of an unknown distribution from data generated by that distribution. The most common type of inference involves approximating the unknown distribution by choosing a distribution from restricted family of distributions. Generally the restricted family of distributions is specified parametrically. For example, $N(\mu, \sigma^2)$ represents the family of normal distributions where μ corresponds to the mean of a normal distribution and ranges over the real numbers (R) and σ^2 corresponds to the variance of a normal distribution and ranges over the positive real numbers (R^+). μ and σ^2 are called parameters with R and R^+ their corresponding parameter spaces. The two-dimensional space R cross R^+ is the parameter space for the family of normal distributions. By fixing $\mu = 0$ one can restrict attention to the family corresponding to the set of zero-mean normal distributions. In the following,

one use θ to denote a real-valued parameter or vector of real-valued parameter $\theta = \{\theta_1, \theta_{2,} \ldots\ldots \theta_k\}$ with parameter space Ω and θ.

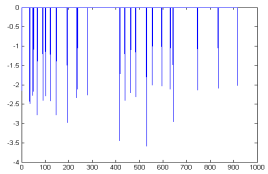

Fig. 1: Normal distribution

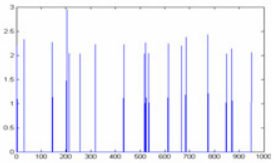

Fig. 2: Exponential distribution

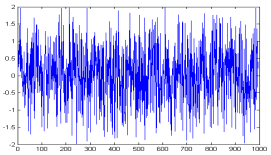

Fig.3: Poisson distribution

5. Conclusion

The method identified outliers in data based on the analysis of robust estimation rejection on direction corresponds to extremes for the distribution function. It is observed that there is often little difference in performance among different data analysis method when applied to normal, exponential and poisson distribution. In this paper an attempt is made to compute the scale, shape and location parameter of normal, exponential and poisson distribution using MATLAB software detecting the outliers and compared with non detecting outliers. The outlier has been to state and test process which served to circumscribe and show to enhancement, the properly which can described Normal, Exponential and Poison distribution.

6. Acknowledgements

The authors wish to thank Mr.N.Senthilkumaran and Mr. K. Mahesh. The authors are grateful to the referees for their support and helpful comments.

7. References

[1] Ying YANG. Asymptotic of M-estimation in Non-linear regression, Acta Mathematica Sinica, English series, 2004, 20, No.4, pp. 749-760.

[2] A. Arning, R. Agrawal, and P. Raghavan. A linear method for deviation detection in large databases. In *Proc. 1996 Int. Conf.DataMining and KnowledgeDiscovery (KDD'96)*, Aug. 1996, pp. 164–169, Portland, Oregon.

[3] H. V. Jagadish, N. Koudas, and S. Muthukrishnan:. Mining deviants in a time series database. In *Proc. 1999 Int. Conf. Very LargeData Bases (VLDB'99)*, Sept. 1999, pp. 102–113, Edinburgh, UK.

[4] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, University of Illinois, Urbana-Champaign, Morgan Kaufmann publications, 2006.

[5] E. Knorr and R. Ng. Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Trans. Knowledge and Data Engineering*, 1996,(8), pp. 884–897,.

[6] V. Barnett and T. Lewis. *Outliers in Statistical Data*. JohnWiley & Sons, 1994.

[7] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, May 2000, pp. 93–104, Dallas, TX.

[8] C. Aggarwal, J. Han, J.Wang, and P. S. Yu. On demand classification of data streams. In *Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04),* Aug. 2004, pp. 503–508, Seattle, WA.

[9] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98,)* Mar. 1998, pp 168–182, Valencia, Spain.