Detection of Outliers and Change points in a Data Stream of Bio Informatics Data

M. L. Prasanthi [1], A.Krishna Chaitanya [2], Dr. N. Sambasiva Rao [3]
Computer Science and Engineering [1], Department of Information Technology [2,3]
Vardhman College of Engineering, Hyderabad [1,2,3]

Abstract

From the past decade outlier detection has been in use. Detection of outliers is an emerging topic and is having robust applications in medical sciences and geological sciences. Outlier detection is used to detect anomalous behaviour of data. Typical problems in Bioinformatics can be addressed by outlier detection. We are introducing PCA based MVE (Minimum Volume Ellipsoidal model) algorithm to detect outliers and change points in bioinformatics data samples, and the detected outliers can be represented using statistical analysis charts. This paper can have several applications in Medical sciences for example diagnosing cancer cells, identifying anomalous blood samples to diagnose malaria, detecting the unwanted growth of cells, detection of lever expansion.

Keywords: *Outlier detection, change points, data mining, bioinformatics, data streams, Principle Component Analysis.*

1. Introduction

Identification of Outliers in a Bioinformatics is an emerging topic in data mining. This proposal is generated on the basis of the concept identification of outliers in a data stream. It can lead to the discovery of unexpected and interesting knowledge. On the other hand, the issue of detecting change points in time-series data has extensively been addressed in statistics and has become one of the issues receiving vast attention in data mining, which is recognized as event change detection and closely related to activity monitoring. Here, by the term change point, we mean a time point at which the data properties suddenly change.

Outlier detection in univariate samples is a common practice and can be carried out straightforwardly by visual inspection of the data or by statistical tests using order statistics. Outlier detection is less straightforward in two-dimensional spaces, because visual inspection is less effective and the order statistics are lacking. In higher-dimensional spaces, the problem is even more difficult. Traditional multivariate outlier-detection methods are based on the calculation of the generalized squared (Mahalanobis) distances for each data point. Unfortunately, outliers greatly inflate the covariance matrix and can therefore effectively mask their own existence. To counter this masking problem, ***Rousseeuw*** introduced the robust minimum volume ellipsoid (MVE) method for detection of outliers in multidimensional data subsets to find the subset that minimizes the volume occupied by the data. The best subset (smallest volume) is then used to calculate the covariance matrix and the Mahalanobis distances to all the data points. An appropriate cut-off value is then estimated, and the observations with distances that exceed that cut-off are declared to be outliers. A serious problem is that both the traditional multivariate and the MVE approach require inversion of the covariance matrix. Therefore, neither method can be applied to samples with singular covariance matrices. We encountered this problem frequently.

Here, we describe implementation of the PCA based MVE method, which deals with several of these issues. We provide a straightforward method for dealing with singular covariance matrices and facilitate re-examination of the original samples by sorting the outliers by their Mahalanobis distances

A more exhaustive list of applications that utilize outlier detection is:
- Fraud detection - detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones.
- Loan application processing - to detect fraudulent applications or potentially problematical customers.
- Intrusion detection - detecting unauthorized access in computer networks.
- Activity monitoring - detecting mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
- Network performance - monitoring the performance of computer networks, for example to detect network bottlenecks.

- Fault diagnosis - monitoring processes to detect faults in motors, generators, pipelines or space instruments on space shuttles for example.
- Structural defect detection - monitoring manufacturing lines to detect faulty production runs for example cracked beams.
- Satellite image analysis - identifying novel features or misclassified features.
- Detecting novelties in images - for robot neotaxis or surveillance systems.
- Motion segmentation - detecting image features moving independently of the background.
- Time-series monitoring - monitoring safety critical applications such as drilling or high-speed milling.
- Medical condition monitoring - such as heart-rate monitors.
- Pharmaceutical research - identifying novel molecular structures.
- Detecting novelty in text - to detect the onset of news stories, for topic detection and tracking or for traders to pinpoint equity, commodities, FX trading stories, outperforming or underperforming commodities.
- Detecting unexpected entries in databases - for data mining to detect errors, frauds or valid but unexpected entries.
- Detecting mislabeled data in a training data set.

This Framework provides first generation of the dataset, which was followed by the application of algorithm which is PCA based MVE model. Algorithm calculates the statistical data such as mean, median, covariance. This data is used to generate the reports called statistical analysis charts, using this charts we can easily detect the outliers.

The broad view of this proposal has beneficial application in Bioinformatics' technology. This proposal cooperates with the medical sciences to have a better idea about diseases in inner parts of the human body, so that they can be cured easily in preliminary stages such as detecting unwanted growth of cells in diagnosing cancer and detecting the points where the lever was expanded.

2. Problem Definition:

Our procedure for detecting outliers in multivariate data sets by the MVE method effectively eliminates the problems associated with singular variance-covariance matrices. In addition, the return of a ranked list of outliers facilitates error checking and helps to emphasize that outlier detection is a somewhat arbitrary process. The validity of the outliers cannot always be checked against the raw data, but our implementation always indicates which cases should be viewed with the most caution. In MVE the best ellipsoid could be missed because of the random sampling of the data set, so some outliers might be missed or some valid points labeled as outliers. In practice, the number of sub samples examined ensures that the best subset is always close to the actual best MVE, so the discrepancies will be small and only individuals very close to the optimal cut-off value will be missed. The implementation of MVE in SAS masks this random effect by seeding the pseudo-random number generator with the same seed every time. The random aspect in the method will always remain a point of concern, and outliers close to the edge of the ellipsoid should be treated cautiously. A more serious issue is the assumption of multivariate normality, which may often be false. Slight departures from normality may either increase or decrease the proportion of observations declared outliers.

Outlier detection is related fraud detection, rare event discovery whereas change point detection relates to event/trend change, activity monitoring.

Existing system uses statistical models such as Gaussian mixture model for continuous variables and histogram density model. For discrete variables outliers and change points are detected independently from stationary data stream. For these
statistical models an algorithm called online discount learning model was in use. It can track time varying data sources.

3. Design Methodology:

Outliers in the data are detected with the MVE module. The loss of degrees of freedom due to the alignment of landmarks means that the variance-covariance matrices of even the largest data sets will be singular. We therefore perform a principal components analysis before outlier detection and score the observations on the eigenvectors with positive eigenvalues. These variates are then subjected to the MVE algorithm. The Mahalanobis distance of each detected outliers is retained, and these are sorted from largest to smallest.

The outlier observations are then inspected by a human observer and corrected if necessary with the digitizing program. Checking proceeds from the largest outliers to the smallest. Commonly a large group of observations are correctly partitioned but fall slightly farther from the distribution than expected under normality. Once the observer checking finds that the overwhelming majority of remaining images are in this category, checking is suspended, so additional time is saved in large data sets. Observations that remain anomalous after errors are removed can either be removed from the data set or retained at the discretion of the user. Design methodology for this proposal provides a five step process.

3.1 Procedure:

Step1: Proposal uses a Bioinformatics tool called "TEA_O" to generate or create different types of data sets.

Step 2: After creating datasets apply a PCA based MVE (Minimum Volume Ellipsoidal) Model algorithm to selected samples of data.
Step 3: Algorithm takes minimum five samples of data as input and generates appropriate statistical data for the selected dataset.
Step 4: Apply input filters to the appropriate dataset
Step 5: Reports are used to visualize the outliers in the data set by using statistical Analysis charts.

3.2 Statistical charts representing outliers:

Statistical models are generally suited to quantitative real-valued data sets or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical (numerical) processing. This limits their applicability and increases the processing time if complex data transformations are necessary.

3.2.1 Box plots:

One of the simplest statistical outlier detection techniques described here, uses informal box plots to pinpoint outliers in both univariate and multivariate data sets. Fig.1 produces a graphical representation and allows a human auditor to visually pinpoint the outlying points.

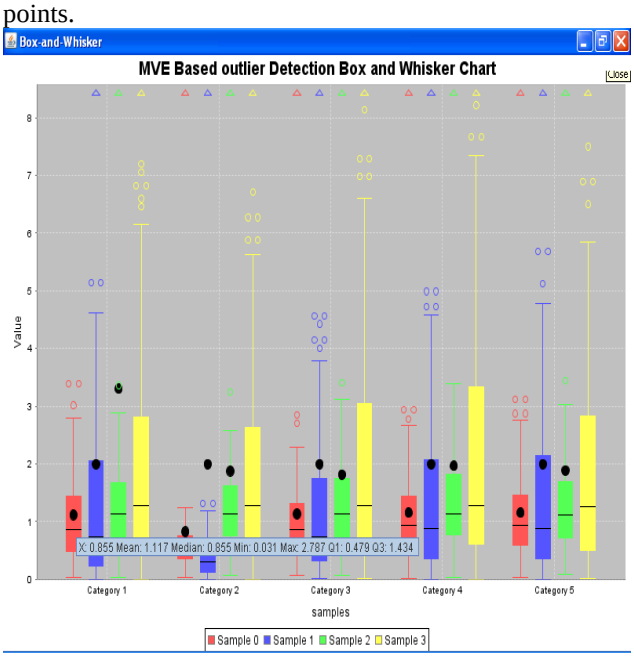

Fig1. MVE Based Box and Whisker Chart

3.2.2 Linear regression model:

Linear regression is the most widely used of all statistical techniques. It is the study of linear (i.e., straight-line) relationships between variables, usually under an assumption of normally distributed errors. The following Fig.2 shows the Linear Regression Statistical Analysis Model for the given Data Samples. In this model human can directly observe outliers.

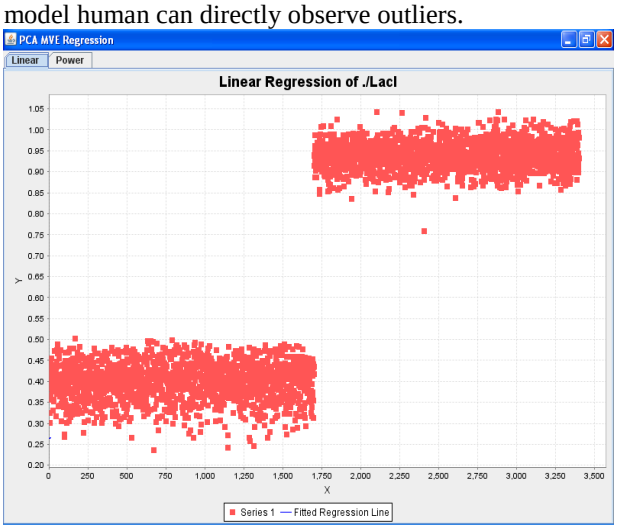

Fig.2 Linear Regression Analysis Model

4. Conclusion:

The proposal for detecting outliers in multivariate data sets by the MVE method effectively eliminates the problems associated with singular variance-covariance matrices. In addition, the return of a ranked list of outliers facilitates error checking and helps to emphasize that outlier detection is a somewhat arbitrary process. When the underlying data are still available, as in our case, the data can themselves be checked and an informed decision made about the disposition of each flagged observation.

The validity of the outliers cannot always be checked against the raw data, but our implementation always indicates which cases should be viewed with the most caution. In MVE the best ellipsoid could be missed because of the random sampling of the data set, so some outliers might be missed or some valid points labeled as outliers. In practice, the number of subsamples examined ensures that the best subset is always close to the actual best MVE, so the discrepancies will be small and only individuals very close to the optimal cut-off value will be missed.

5. Bibliography:

[1] V. Barnett and T. Lewis 1994 “Outliers in Statistical Data,” John Wiley & Sons.
[2] P. Burge and J. Shaw-Taylor 1997 “Detecting Cellular Fraud Using Adaptive Prototypes,” Proc. AI Approaches to Fraud Detection and Risk Management, pp. 9-13.
[3] T. Cover and J.A. Thomas 1991 “Elements of Information Theory,” Wiley-International.
[4] T. Fawcett and F. Provost 1999 “Activity Monitoring: Noticing Interesting Changes in Behavior,”

Proc. ACM-SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 53-62.

[5] S.B. Guthery, "Partition Regression 1974," J. Am. Statistical Assoc., vol. 69, no. 348, pp. 945-947.

[6] D.M. Hawkins 1976 "Point Estimation of Parameters of Piecewise Regression Models," J Royal Statistical Soc. Series C, vol. 25, no. 1, pp. 51-57.

[7] M. Huskova, 1993 "Nonparametric Procedures for Detecting a Change in Simple Linear Regression Models," Applied Change Point Problems in Statistics.

[8] G. Kitagawa and W. Gersch 1996 "Smoothness Priors Analysis of Time Series," Lecture Notes in Statistics, vol. 116, Springer-Verlag.

[9] E.M. Knorr and R.T. Ng 1998 "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. 24th Very Large Data Bases Conf.,pp. 392-403.

[10] U. Murad and G. Pinkas 1999 "Unsupervised Profiling for Identifying Superimposed Fraud," Proc. Third European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 251-261.

[11] R.M. Neal and G.E. Hinton 1993 "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants".

[12] T. Ozaki and G. Kitagawa 1995 "A Method for Time Series Analysis," Asakura Shoten, (in Japanese).

[13] J. Rissanen 1996 "Fisher Information and Stochastic Complexity," IEEE Trans. Information Theory, vol. 42, no. 1, pp. 40-47.

[14] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne May 2004 "Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms," Data Mining and Knowledge Discovery J., vol. 8, no. 3, pp. 275-300.

[15] K. Yamanishi and J. Takeuchi 2001 "Discovering Outlier Filtering Rules from Unlabeled Data," Proc. Fourth Workshop Knowledge Discovery and Data Mining, pp. 389-394.

[16] K. Yamanishi and J. Takeuchi 2002 "A Unifying Approach to Detecting Outliers and Change-Points from Nonstationary Data," Proc of the Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.

[17] B.K. Yi, N.D. Sidiropoulos, T. Johnson, H.V. Jagadish, C. Faloutsos, and A. Biliris 2000 "Online Data Mining for Co-Evolving Time Sequences" Proc. 16th Int'l Conf. Data Eng.

Lakshmi Prasanthi Malyala received the Bachelor of Technology degree with Information Technology from the Jawaharlal Nehru Technological University, Hyderabad, in 2005 and the Master of Technology degree in Computer Science and Engineering-Software Engineering from Jawaharlal Nehru Technological University, Hyderabad, in 2008. She is currently working as Associate Professor in the Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad. His research interests include Data Mining, Network Security and Cryptography, Software Engineering, Cloud Computing, and Data Structures.

Krishna Chaitanya Atmakuri received the Bachelor of Technology degree with Information Technology from the University of Madras, Chennai, in 2004 and the Master of Technology degree in Computer Science and Engineering-Software Engineering from Jawaharlal Nehru Technological University, Hyderabad, in 2008. He is currently working as Associate Professor in the Department of Information Technology, Vardhaman College of Engineering, Hyderabad. His research interests include Data Mining, Network Security and Cryptography, Information Retrieval Systems, Database Systems, and Cloud Computing.

Dr. N Sambasiva Rao received the Bachelor of Technology degree with majors in Electrical and Electronics Engineering from Regional Engineering College, Warangal, he got double Master of Technology from Computer Science and Engineering and Power Systems Engineering from Regional Engineering College, Warangal and Motilal Nehru Regional College of Engineering, and the Ph.D degree in computer science from Anna University, chennai in 2008. He is currently working as a Professor & Head in the Department of Information Technology and Principal of Vardhaman College of Engineering. In his 15 years of Research, he published more than 15 papers on referred Journals and Conferences and Guiding more than four Research Scholars. His research interests include Software Engineering and Formal Methods, Data Mining, Computer Networks and Security, and Power Electronics.