ISSN:0975-9646

Holy Grail of Outlier Detection Technique: A Macro Level Take on the State of the Art

Safal V Bhosale
CSE, MIT, Aurangabad

Abstract— **A phenomenal interest in big data among research community has emerged. Outlier detection analysis in big data leads to new discoveries in databases. Outlier can be a noise or anomaly. Outlier detection indicates identifying deviations from normal behavior in the data set. Cause of Outliers maybe error prone human inputs. It is desirable to identify outliers at the initial stages to avoid ballooning effect. Present focus of the state of the art is on time series, temporal outliers and categorical outliers. Outlier detection and removal is a form of filtering data from impurities. This paper is a survey of state of the art outlier detection and analysis mechanisms. This survey will give a comparative overview of contemporary outlier detection techniques.**

Keywords: Outlier Detection, noise, anomaly, time series, temporal, Categorical data

INTRODUCTION

Data Mining is defined as “the non-trivial extraction of implicit, formerly unidentified and potentially constructive information from data in databases” [1], [2]. Data mining is being used in various domains. Data mining can accurately predict market basket analysis, fraud detection, direct marketing, market segmentation, trend analysis and so on [3, 4, 5, 7]. Outlier detection, a subdivision of data mining differs in the sense it detects data showing different behavior from the rest of the data [8]. As per Hawkins [9]: ‘‘An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism’’. As per Thakran[10]: “In many applications outliers are more interesting than normal cases for example network intrusion detection, fault diagnosis in machines (motors, space shuttles, etc.), credit card fraud detection, marketing, detecting outlying cases in wireless sensor network data”. As per Grubbs (Grubbs, 1969): “An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs” [12]. As per Hodge [13] et al: “Outlier detection is a critical task in many safety critical environments as the outlier indicates abnormal running conditions from which significant performance degradation may well result, such as an aircraft engine rotation defect or a flow problem in a pipeline. An outlier can denote an anomalous object in an image such as a land mine. An outlier may pinpoint an intruder inside a system with malicious intentions so rapid detection is essential. Outlier detection can detect a fault on a factory production line by constantly monitoring specific features of the products and comparing the real-time data with either the features of normal products or those for faults”.

As per Chandola [8] et al: “The importance of outlier detection is due to the fact that outliers in data translate to significant (and often critical) information in a wide variety of application

domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. In public health data, outlier detection techniques are widely used to detect anomalous patterns in patient medical records which could be symptoms of a new disease. Similarly, outliers in credit card transaction data could indicate credit card theft or misuse. Outliers can also translate to critical entities such as in military surveillance, where the presence of an unusual region in a satellite image of enemy area could indicate enemy troop movement. Or anomalous readings from a space craft would signify a fault in some component of the craft. Outlier detection has been found to be directly applicable in a large number of domains”.

In this paper, we are giving a comprehensive overview of Outlier detection techniques. The remainder of this paper is organized as follows. Section 2 provides a review of related works. In section 3 explains basic concept of application areas. In section 4 focus on various categories of Outlier detection techniques, we conclude in section 6.

REVIEW OF RELATED WORKS

Literature presents several techniques for outlier detection. Here, we review some of the techniques presented for literature. Barnett & Lewis (Barnett and Lewis, 1994) [14] and Rousseeuw & Leroy (Rousseeuw and Leroy, 1996) [15] describe and analyse a broad range of statistical outlier techniques and Marsland (Marsland, 2001) [16] analyses a wide range of neural methods. As per [13] outlier detection methods are derived from three fields of computing: statistics (proximity-based, parametric, non-parametric and semi-parametric), neural networks (supervised and unsupervised) and machine learning. The methods applied include distance-based, set-based, density-based, depth-based, model-based and graph-based algorithms.

Samples inside each data set are grouped such that the dependencies among groups of dissimilar sets incarcerate as much of pair wise dependencies between the samples as feasible. In a new probabilistic way they have formalized this problem, as optimization of a Bayes factor. The technique is used to expose commonalities and exceptions in gene expression among organisms and to propose regulatory interactions in the form of dependencies between gene expression profiles and regulator binding patterns. Outlier detection algorithms started with Statistical approaches and most of them are applicable only for single dimensional data sets.

The most well known single dimensional method as quoted in [12] is Grubbs' method (Extreme Studentized Deviate) (Grubbs, 1969) which calculates a Z value as the difference between the mean value for the attribute and the query value divided by the standard deviation for the attribute where the mean and standard deviation are calculated from all attribute values including the query value. The Z value for the query is compared with a 1% or 5% significance level. The technique requires no user parameters as all parameters are derived directly from data. However, the technique is susceptible to the number of exemplars in the data set. The higher the number of records the more statistically representative the sample is likely to be.

Laurikkala et al. (Laurikkala et al., 2000) [17] use informal box plots to pinpoint outliers in both univariate and multivariate data sets.This produces a graphical representation and allows a human auditor to visually pinpoint the outlying points. Their approach can handle real-valued, ordinal and categorical (no order) attributes. Box plots plot the lower extreme, lower quartile, median, upper quartile and upper extreme points. For multivariate data sets there are no unambiguous total orderings but it is recommended using the reduced sub-ordering based on the generalised distance metric using the Mahalanobis distance measure. The Mahalanobis distance measure includes the inter-attribute dependencies so the system can compare attribute combinations. It was found that the approach most accurate for multivariate data where a panel of experts agreed with the outliers detected by the system.

$$\boxed{\sqrt{(\mathbf{x}-\mu)^T\mathbf{C}^{-1}(\mathbf{x}-\mu)}|}$$

Equation for Mahalanobis Distance extracted from [12]

$$\boxed{\sqrt{\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{y}_i)^2}}$$

Equation for Euclidean Distance extracted from [12]

Ramaswamy et al. (Ramaswamy et al., 2000) [18] introduce an optimised k-NN to produce a ranked list of potential outliers. A point p is an outlier if no more than n − 1 other points in the data set have a higher Dm (distance to mth neighbour) where m is a user-specified parameter. In figure 3, V is most isolated followed by X, W, Y then Z so the outlier rank would be V, X, W, Y, Z. This approach is susceptible to the computational growth as the entire distance matrix must be calculated for all points (ALL k-NN) so Ramaswamy et al. include techniques for speeding the k-NN algorithm such as partitioning the data into cells.. It was next clustered with a slightly improved density-based clustering algorithm. Their experimental result illustrated that their planned algorithm outperformed in high quality there are various flavours of k-Nearest Neighbour (k-NN) algorithm for outlier detection but all calculate the nearest neighbours of a record using a suitable distance calculation metric such as Euclidean distance or Mahalanobis distance.

Application Areas Of Outlier Detection

An erratic pattern in a network indicates a hacked computer is sending out sensitive data to an unauthorized destination. Similarly, outliers in credit card transaction data could indicate credit card theft or misuse. Efficient detection of such outliers could reduce the risk of making poor decisions based on erroneous data, and aids in identifying and preventing the effects of malicious or faulty behavior [19]. Defaulters of credit card, insurance, tax fraud detection in financial system, intrusion detection in computer networks, fault detection in safety critical systems, military surveillance for enemy activities are few of the areas which can be exploited well by outlier detection. Outlier detection schemes have been in early detection of *Insider Trading*. Insider trading is a phenomenon found in stock markets, where people make illegal profits by acting on (or leaking) inside information before the information is made public. The inside information can be of different forms [Donoho 2004] [20]. It could refer to the knowledge of a pending merger/acquisition, a terrorist attack affecting a particular industry, a pending legislation affecting a particular industry or any information which would the stock prices in a particular industry. Early detection of insider trading done based on this information is required to prevent people/organizations from making illegal profits. Outlier detection in the medical and public health domains typically work with patient records. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy. The data typically consists of records which may have several different types of features such as patient age, blood group, weight. The data might also have temporal as well as spatial aspect to it.

Categories Of Outlier Detection Techniques

Outlier detection techniques can be divided into three categories:

A Supervised:

Such techniques assume the availability of a training data set which has labeled instances for normal as well as outlier class. Typical approach in such case is to build predictive models for both normal and outlier classes. Any unseen data instance is compared against the two models to determine which class it belongs to. Supervised outlier detection techniques have an explicit notion of the normal and outlier behavior and hence accurate models can be built.

B Semi-supervised:

Such techniques assume the availability of labeled instances for only one class. it is often difficult to collect labels for other class. For example, in space craft fault detection, an outlier scenario would signify an accident, which is not easy to model. The typical approach of such techniques is to model only the available class and declare any test instance which does not this model to belong to the other class.

C. Unsupervised:

The third category of techniques does not make any assumption about the availability of labeled training data. Thus these techniques are most widely applicable. The techniques in this category make other assumptions about the data. For example, parametric statistical techniques, assume a parametric distribution of one or both classes of instances. Similarly, several techniques make the basic assumption that normal instances are far more frequent than outliers. Thus a frequently occurring pattern is typically considered normal while a rare occurrence is an outlier. The unsupervised techniques typically suffer from higher false alarm rate, because often times the underlying assumptions do not hold true.

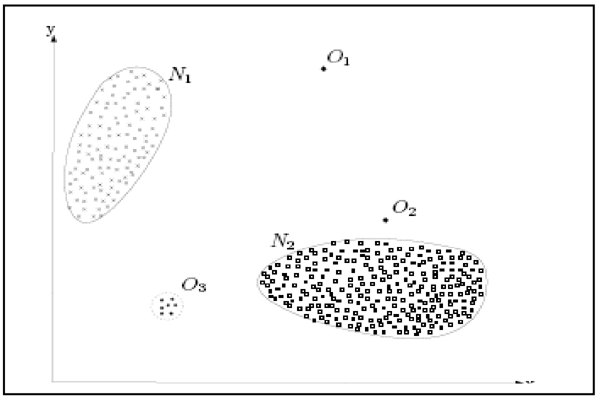

Figure 1: Outliers in Two Dimensional Data extracted from Chandola et al

CONCLUSION

In this article, we present a comprehensive overview of outlier detection techniques. Outlier detection technique is yet to be explored completely for a specific domain. Different techniques for different problems may have to be dealt with. This survey is an effort to formulate research questions for a new problem domain and come out with a novel algorithm to solve the problem.

In this survey basics of outlier detection technique, related works, application areas and categories of outlier are discussed.

REFERENCES

[1] Osmar R. Z., “Introduction to Data Mining”, In: Principles of Knowledge Discovery in Databases. CMPUT690, University of Alberta, Canada, 1999.

[2] Kantardzic, Mehmed. “Data Mining: Concepts, Models, Methods, and Algorithms”, John Wiley and Sons, 2003.

[3] E. Wainright Martin, Carol V. Brown, Daniel W. DeHayes, Jeffrey A. Hoffer and William C. Perkins, “Managing information technology”, Pearson Prentice-Hall 2005.

[4] Andrew Kusiak and Matthew Smith, “Data mining in design of products and production systems”, in proceedings of Annual Reviews in control, vol. 31, no. 1, pp. 147- 156, 2007.

[5] Mahesh Motwani, J.L. Rana and R.C Jain, "Use of Domain Knowledge for Fast Mining of Association Rules", in Proceedings of the International Multi-Conference of Engineers and Computer Scientists, 2009.

[6]. Souptik Datta Kanishka Bhaduri Chris Giannella Ran Wolff Hillol Kargupta "Distributed Data Mining in Peer-to-Peer Networks", Journal of internet computing, vol.10, no.4, pp.18-26. 2006.

[7] Ron Wehrens and Lutgarde M.C. Buydens, “Model-Based Clustering for Image Segmentation and Large Datasets via Sampling”, Journal of Classification, Vol. 21, pp.231-253, 2004.

[8] Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Comput Surv 41(3):Article 15

[9] Hawkins D (1980) Identification of outliers. Chapman and Hall,London

[10] Thakran. Y, Toshniwal .D, “ Unsupervised outlier detection in streaming data using weighted clustering”, Intelligent Systems Design and Applications (ISDA), 2012.

[11] Grubbs, F. E. (1969). Procedures for detecting outlying observations, Technomctrics, 11, 1- 21

[12] Inderjit S. Dhillon and Dharmendra S. Modha, “A Data-Clustering Algorithm On Distributed Memory Multiprocessors”, Proceedings of KDD Workshop High Performance Knowledge Discovery, pp. 245-260, 1999.

[13] Victoria J. Hodge and Jim Austin, A Survey of Outlier Detection Methodologies, Kluwer Academic Publishers, 2004

[14] Barnett, V. and Lewis, T.: 1994, *Outliers in Statistical Data*. John Wiley & Sons. 3 edition.

[15] Rousseeuw, P. and Leroy, A.: 1996, *Robust Regression and Outlier Detection*. John Wiley & Sons., 3 edition.

[16] Marsland, S.: 2001, ‘On-Line Novelty Detection Through Self-Organisation, with Application to Inspection Robotics’. Ph.D. thesis, Faculty of Science and Engineering, University of Manchester, UK.

[17] Laurikkala, J., Juhola, M., and Kentala, E.: 2000, ‘Informal Identification of Outliers in Medical Data’. In: Fifth International Workshop on Intelligent Data Analysis
in Medicine and Pharmacology IDAMAP-2000 Berlin, 22 August. Organized as a workshop of the 14th European Conference on Artificial Intelligence ECAI-2000.

[18] Ramaswamy, S., Rastogi, R., and Shim, K.: 2000, ‘Efficient Algorithms for Mining Outliers from Large Data Sets’. In: Proceedings of the ACM SIGMOD Conference on Management of Data. Dallas, TX, pp.427–438.

[19] Kumar V (2005) Parallel and distributed computing for cybersecurity. IEEE Distrib Syst Online 6(10). doi:10.1109/MDSO. 2005.53.

[20] Donoho, S. 2004. Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 420, 429.