Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 7; October, 2014 pp. 30-33

http://www.krishisanskriti.org/jbaer.html

Outlier Detection via online OSPCA in High Dimensional Space

P.G.K. Pabitha[1], K. Bhaskar Naik[2]

[1]*M.Tech, Sreevidyanikethan Engineering College*
[2]*Sreevidyanikethan Engineering College*

Abstract: ***Outlier detection is the process of identifying unusual behavior. It is widely used in data mining, for example, to identify customer behavioral change, fraud and manufacturing flaws. In recent years many researchers had proposed several concepts to obtain the optimal result in detecting the anomalies. But the process of PCA made it challenging due to its computations. In order to overcome the computational complexity, online oversampling PCA has been used. The algorithm enables quick Online updating of the principal directions for the effective computation and satisfying the online detecting demand and also oversampling will improve the impact of outliers which leads to accurate detection of outliers. Experimental results show that this method is effective in computation time and need less memory requirements also clustering technique is added to it for optimization.***

Keywords: ***online oversampling PCA, Online updating Technique, Outlier detection.***

1. INTRODUCTION

Outlier detection aims to identify a small group of instances which deviate remarkably from the existing data. A well-known definition of "outlier" is given in: "an observation which deviates so much from other observations" which gives the general idea of an outlier and motivates many anomaly detection methods.

The outlier detection technique finds application in credit card fraud, network intrusion detection, financial applications and marketing. This problem typically arises in the context of very high dimensional data sets. Much of recent work on finding outliers use methods which make implicit assumptions of relatively low dimensionality of the data. Thus, we discuss new techniques for outlier detection which finds the outliers by studying the behavior of projections from the data set. But in this real world applications limited amount of data is available because of which it is difficult to identify the anomaly of the unseen data.

Leave One Out (LOO) strategy can be use to calculate the principal direction of the data set without the target instance present and that of the original data set. Thus the anomaly can be determined by the variation of the resulting principal directions. The difference between these two eigenvectors indicate the anomaly of the target instance. By ranking the scores of all data points, it is easy to identify the outlier data by a predefined threshold or a predetermined portion of the data this can be considered as a decremental PCA (dPCA)-based approach for anomaly detection. It works well for applications with Small data set size, but it might not be significant when the size of the data set is large. It can produce the negligible difference in the eigenvectors hence it is not efficient to apply dPCA.

For addressing this practical problem, the "oversampling" strategy is used to duplicate the target instance, and to perform an oversampling PCA (osPCA) on such an oversampled data set. An outlier instance will be amplified due to its duplicates present in the PCA formulation due to this it becomes easier to detect outlier data.

2. OVERSAMPLING PCA FOR ANOMALY DETECTION

For practical anomaly detection problems, the size of the data set is typically large, and thus it might not be easy to observe the variation of principal directions caused by the presence of a single outlier. Furthermore, in the above PCA framework for anomaly detection, there need to perform n PCA analysis for a data set with n data instances in a p-dimensional space, which is not computationally feasible for large-scale and online problems. The proposed oversampling PCA (osPCA) together with an online updating strategy will address the above issues, as we now discuss. Here introduce osPCA, and discuss how and why this method is able to detect the presence of abnormal data instances according to the associated principal directions, even when the size of data is large. The well-known power method is applied to determine the principal direction without the need to solve each eigenvalue decomposition problem. While this power method alleviates the computation cost in determining the principal direction as verified in previous work will discuss its limitations and explain why the use of power method is not practical in online settings. This method presents a least squares approximation of our osPCA, followed

by the proposed online updating algorithm which is able to solve the online osPCA efficiently.

The proposed osPCA scheme will duplicate the target instance multiple times, and the idea is to amplify the effect of outlier rather than that of normal data. While it might not be sufficient to perform anomaly detection simply based on the most dominant eigenvector and ignore the remaining ones, our online osPCA method aims to efficiently determine the anomaly of each target instance without sacrificing computation and memory efficiency. More specifically, if the target instance is an outlier, this oversampling scheme allows us to overemphasize its effect on the most dominant eigenvector, and thus we can focus on extracting and approximating the dominant principal direction in an online fashion, instead of calculating multiple eigenvectors carefully. These existing approaches can be divided into three categories: distribution (statistical), distance and density-based methods. Statistical approaches assume that the data follows some standard or predetermined distributions, and this type of approach aims to find the outliers which deviate from such distributions.

However, most distribution models are assumed univariate, and thus the lack of robustness for multidimensional data is a concern. Moreover, since these methods are typically implemented in the original data space directly, their solution models might suffer from the noise present in the data. Nevertheless, the assumption or the prior knowledge of the data distribution is not easily determined for practical problems. For distance-based methods the distances between each data point of interest and its neighbors are calculated. If the result is above some predetermined threshold, the target instance will be considered as an outlier. While no prior knowledge on data distribution is needed, these approaches might encounter problems when the data distribution is complex (e.g., multi-clustered structure). In such cases, this type of approach will result in determining improper neighbors, and thus outliers cannot be correctly identified. To alleviate the aforementioned problem, density-based methods are proposed one of the representatives of this type of approach is to use a density-based local outlier factor (LOF) to measure the outlierness of each data instance. Based on the local density of each data instance, the LOF determines the degree of outlierness, which provides suspicious ranking scores for all samples.

3. IMPLEMENTATION DETAILS

1. Proposed System:

Traditional outlier detection or anomaly detection algorithm was not able to work with large amount of data due to memory and computational complexity, when we work with large size of input dataset, LOO technique will not significantly affect the resulting principal direction of the data. Therefore, we extend traditional PCA to the online oversampling strategy and present an online oversampling PCA (oosPCA) algorithm for largescale anomaly detection problems. The proposed oosPCA scheme will replicate the target instance many times, and the idea is to expand the effect of anomaly other than normal data. Due to the introduction of online over sampling, the computation and memory efficiency will not be compromised and we can achieve noteworthy results as compared with the over sampling PCA. Instead of calculating single eigenvectors for single outlier, we are going to used the over sampling technique to maximize the anomalies so that we can calculate the principal direction as an average of over sampled data. In our method we use online services to calculate eigenvectors value. Therefore the memory and computational complexity will not be limited to detect the outliers.

2.BlockDiagram:

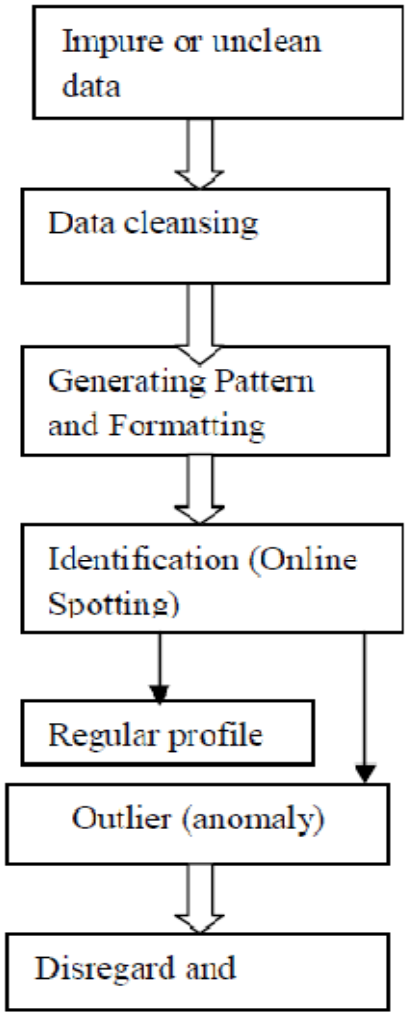

Fig: 1 Block diagram showing the overall control flow of outlier detection.

4. METHODOLOGY

- 1.PCA (Principal Component Analysis).
- 2.Oversampling Principal Components Analysis (osPCA).

1. PCA is a method of dimension reduction which identifies the principal directions of the data distribution. To obtain these principal directions, it is important to construct the data covariance matrix and calculate its dominant eigenvectors. These eigenvectors provides more information among the vectors in the data space therefore they are consider as the principal directions.

2. osPCA.is the method in which it is not necessary to store the entire covariance matrix, Stherefore it is used in online or large-scale problems. By oversampling and extracting the target instant, the osPCA allows to determine the anomaly of the target instance according to the variation of the dominant eigenvector. In our osPCA framework, we will duplicate the target instance multiple times over web services and we will compute the score of anomalies of that output instance. If this score or the obtained result is above predefined threshold, we will consider this instance as an outlier. osPCA not only determines outliers from the existing data, it can be applied to anomaly detection execution end rather it will be dependent on web service performance. Problems with streaming data or those with online requirements.

Quality Attributes

Usability: The application seem to user friendly since the GUI is interactive.

Maintainability: This application is maintained for long period of time since it will be implemented under java platform.

Reusability: The application can be reusable by expanding it to the new modules.

Portability: The application is purely a portable mobile application since it can only be operated on an android operating system.

5. RESULTS

In this result section we present the experimental results in terms of the selected features and the time required to obtain the result which is carried on the KDD Intrusion Detection Dataset. In the KDD dataset the result is obtained on the categories of the dataset. This result is generated for two methodologies the first one is the osP CA and the second one is the oosPCA. The result generated is carried out on the platform Java using JDK 1.6 and experiments were performed on a PC with an Intel Core 2 Duo, 2GB of main memory.

1. DATASET

Data sets use in my framework is the KDD intrusion detection, splice, pnedigits, adult pima, code-rna etc .

For the KDD intrusion detection data set, there are four categories of attacks to be considered as outliers:

- DOS: denial-of-service.
- R2L: unauthorized access from a remote machine.
- U2R: unauthorized access to local superuser (root) privileges.
- Probe: surveillance probing and other probing.

The Method starts with collecting a benchmark dataset to start the analysis.The data sets are from the UCI repository of machine learning data archive. The data sets were meant for binary classification.In that the majority class as normal data and randomly select one percent data instances from the minority class as outlier samples.The Unformatted data such as noisy, incomplete datas are filtered.

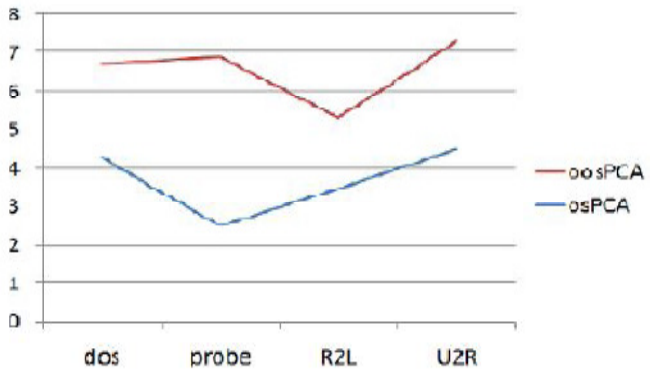

Fig 2.graph showing the result for accuracy.

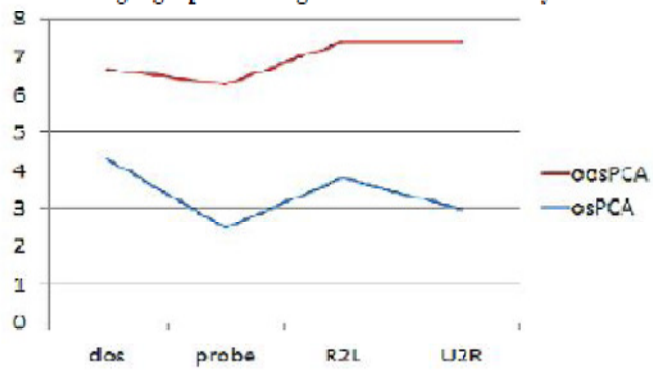

Fig 3.graph showing the result for performance.

6. CONCLUSION

An online anomaly detection method based on oversample PCA, and thus can successfully use the variation of the dominant principal direction to identify the presence of rare but abnormal data, online osPCA is preferable for online large-scale or streaming data problems, compared with other anomaly detection methods, this approach is able to achieve satisfactory results while significantly reducing computational costs and memory requirements. This paper provides a solution to the curse of dimensionality problem in the pair wise scoring techniques. The future work can be carried on the normal data with the multiclustering structure. It is also required to focus on the extremely high dimensional data so as to solve the problem of curse of dimensionality.

REFERENCES

[1] M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, "*LOF: Identifying Density-Based Local Outliers*, " Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000**.**

[2] F. Angiulli, S. Basta, and C. Pizzuti, "*Distance-Based Detection and Prediction of Outliers*, " IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, 2006.

[3] N.L.D. Khoa and S. Chawla, "*Robust Outlier Detection Using Commute Time and Eigen space Embedding*, " Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2010.

[4] V. Barnett and T. Lewis, "*Outliers in Statistical Data*", John Wiley Sons, 2006.

[5] D.M. Hawkins, "*Identification of Outliers*". Chapman and Hall, 1980.

[6] W. Jin, A.K.H. Tung, J. Han, and W. Wang, "*Ranking Outliers Using Symmetric Neighborhood Relationship*, " Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2006.

[7] H.-P. Kriegel, M. Schubert, and A. Zimek, "*Angle-Based Outlier Detection in High- Dimensional Data*, " Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and data Mining, 2008.

[8] C.C. Aggarwal and P.S. Yu, "*Outlier Detection for High Dimensional Data*," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2001.

[9] T. Ahmed, "*Online Anomaly Detection using KDE*, " Proc. IEEE Conf. Global Telecomm., 2009.

[10] X. Song, M. Wu, and C.J., and S. Ranka, "*Conditional Anomaly Detection*, "IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 631-645, May 2007.

[11] L. Huang, X. Nguyen, M. Garofalakis, M. Jordan, A.D. Joseph, and N. Taft, "*In-Network Pca and Anomaly Detection*, " Proc. Advances in Neural Information Processing Systems 19, 2007.

[12] V. Chandola, A. Banerjee, and V. Kumar, "*Anomaly Detection: A Survey*, " ACM Computing Surveys, vol. 41, no. 3, pp. 15:1-15:58, 2009.

[13] W. Wang, X. Guan, and X. Zhang, "*A Novel Intrusion Detection Method Based on Principal Component Analysis in Computer Security*, " Proc. Int'l sym. Neural Networks, 2004.