ISSN:2320-0790

An International Journal of Advanced Computer Technology

Outlier Detection via Online Oversampling in High Dimensional space

Kavya M Menon[1], G.Sakthi[2]

[1]PG Scholar, Hindustan Institute of Technology
[2]Assistant Professor, Hindustan Institute of Technology

Abstract: Anomaly detection is an important topic in data mining. Many applications such as intrusion or credit card fraud detection require efficient method to identify deviated data instances, mostly anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. This paper proposes an online oversampling principal component analysis (osPCA) algorithm, and aim to detecting the presence of outliers from a large amount of data via an online updating technique. In prior principal component analysis (PCA)-based approaches, we store the entire data matrix or covariance matrix, and in proposed method there is no need to store the entire data matrix or covariance matrix thus this approach is especially of used in online or large-scale problems. Oversampling the target instance and extracting the principal direction of the data, the proposed osPCA allows determining the Anomaly or outlier of the target instance according to the variation of the resulting dominant eigenvector. In the proposed method osPCA need not perform Eigen analysis, so the proposed method is applicable for online applications Compared with the well-known power method for PCA and other popular anomaly detection algorithms, our experimental results verify proposed method is feasible in terms of both efficiency and accuracy.

Keywords: **Anomaly detection, online updating, oversampling, principal component analysis**

I. 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. Practically, anomaly detection can be found in applications such as homeland security, credit card fraud detection, intrusion and insider threat detection in cyber-security, fault detection, or malignant diagnosis. However, since only a limited amount of labeled data are available in the above real world applications, how to determine anomaly of unseen data (or events) draws attention from the researchers in data mining and machine learning communities.

Despite the rareness of the deviated data, its presence might enormously affect the solution model such as the distribution or principal directions of the data. For example, the calculation of data mean or the least squares solution of the associated linear regression model is both sensitive to outliers. As a result, anomaly detection needs to solve an unsupervised yet unbalanced data learning problem. Similarly, observe that removing (adding) an abnormal data instance will affect the principal direction of the resulting data than removing (or adding) a normal one does. Using the above "leave one out" (LOO) strategy, can calculate the principal direction of the data set without the target instance present and that of the original data set. Thus, the outlierness (or anomaly) of the data instance can be determined by the variation of the resulting principal directions. More precisely, the difference between these two eigenvectors will indicate the anomaly of the target instance. By ranking the difference scores of all data points, one can identify the outlier data by a predefined threshold or a predetermined portion of the data.

The above framework can be considered as a decremental PCA (dPCA)-based approach for anomaly detection. While it works well for applications with moderate data set size, the variation of principal directions might not be significant when the size of the data set is large. In real-world anomaly detection problems dealing with a large amount of data, adding or removing one target instance only produces negligible difference in the resulting eigenvectors, and one cannot simply apply the dPCA technique for anomaly detection. To address this practical problem, advance the "over sampling" strategy to duplicate the target instance, and perform an over sampling PCA (osPCA) on such an over sampled data set. It is obvious that the effect of an

outlier instance will be amplified due to its duplicates present in the principal component analysis (PCA) formulation, and this makes the detection of outlier data easier. However, this LOO anomaly detection procedure with an over sampling strategy will markedly increase the computational load. For each target instance, one always needs to create a dense covariance matrix and solves the associated PCA problem. This will prohibit the use of our proposed framework for real-world large-scale applications. Although the well known power method is able to produce approximated PCA solutions, it requires the storage of the covariance matrix and cannot be easily extended to applications with streaming data or online settings. Therefore, here present an online updating technique for our osPCA. This updating technique allows us to efficiently calculate the approximated dominant eigenvector without performing Eigen analysis or storing the data covariance matrix. Compared to the power method or other popular anomaly detection algorithms, the required computational costs and memory requirements are significantly reduced, and thus this method is especially preferable in online, streaming data, or large-scale problems.

II. 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.

For each target instance, one always needs to create a dense covariance matrix and solves the associated PCA problem. This will prohibit the use of our proposed framework for real-world large-scale applications. Although the well known power method is able to produce approximated PCA solutions, it requires the storage of the covariance matrix and cannot be easily extended to applications with streaming data or online settings. Therefore, we present an online updating technique for our osPCA. This updating technique allows us to efficiently calculate the approximated dominant eigenvector without performing Eigen analysis or storing the data covariance matrix. Compared to the power method or other popular anomaly detection algorithms, the required computational costs and memory requirements are significantly reduced, and thus our method is especially preferable in online, streaming data, or large-scale problems. Detailed derivations and discussions of the osPCA with our proposed online

updating technique. Besides the above work, some anomaly detection approaches are recently proposed.

Normal data with multiclustering structure, and data in a extremely high dimensional space. For the former case, it is typically not easy to use linear models such as PCA to estimate the data distribution if there exists multiple data clusters. Moreover, many learning algorithms encounter the "curse of dimensionality" problem in an extremely high-dimensional space. The proposed method can handle high-dimensional data since do not need to compute or to keep the covariance matrix, PCA might not be preferable in estimating the principal directions for such kind of data.

III. SYSTEM DESIGN

The system design can be divided into three levels and the three levels are

A. Filtration

In this detection application such as spam mail filtering, one typically designs an initial classifier using the training normal data, and this classifier is updated by the newly received normal or outlier data accordingly in practical scenarios, even the training normal data collected in advance can be contaminated by noise or incorrect data labelling. To construct a simple yet effective model for online detection, one should disregard these potentially deviated data instances from the training set of normal data

B. Clustering

The training data will be selected only by our assumption. So there is a possibility that some outlier data may be considered as normal data in the previous method due to our training data. So the clustering method is used to solve this problem. The clusters are formed for input data instances and then the outlier calculation is applied for each cluster to find the outlier exactly. Rigorous security definition and proved the security of the proposed scheme under the provided definition to ensure the confidentiality. In this clusters are formed for input data instance and then the outlier calculation is applied for each cluster to find outlier exactly

.

C. Anomaly Detection

This is for detecting the outlierness of the user input. When the user giving the input to the system, the system calculate the S_t Value for the new input. And then compare that new S_t Value with the threshold value which is calculated in earlier.

If the S_t Value of the new data instance is above the threshold value, then that input data is identified as an outlier and that value will be discarded by the system. Otherwise it is considered as a normal data instance, and the PCA value of that particular data instance is updated accordingly.

There are two phases required in this framework: Data cleaning and online detection. In the data cleaning phase, goal is to filter out the most deviated data using osPCA before performing online anomaly detection. This data cleaning phase is done offline, and the percentage of the training normal data to be disregarded can be determined by the user. In implementation, choose to disregard 5 percent of the training normal data after this data cleaning process, and use the smallest score of outlierness (i.e., st) of the remaining training data instances as the threshold for outlier detection. More specifically, in the second phase of online detection, use this threshold to determine the anomaly of each received data point. If st of a newly received data instance is above the threshold, it will be identified as an outlier; otherwise, it will be considered as a normal data point, and we will update our osPCA model accordingly.

In the online detection phase, use the dominant principal direction of the filtered training normal data extracted in the data cleaning phase to detect each arriving target instance. Note that, during the entire online detection phase, only need to keep this p-dimensional eigenvector. To determine the outlierness of a newly received instance, we apply the osPCA with the proposed online updating technique to evaluate the variation of the updated principal direction. If the resulting st is above the threshold determined previously, the target instance will be detected as an outlier; otherwise, will consider this input as a normal data instance and update the principal direction accordingly

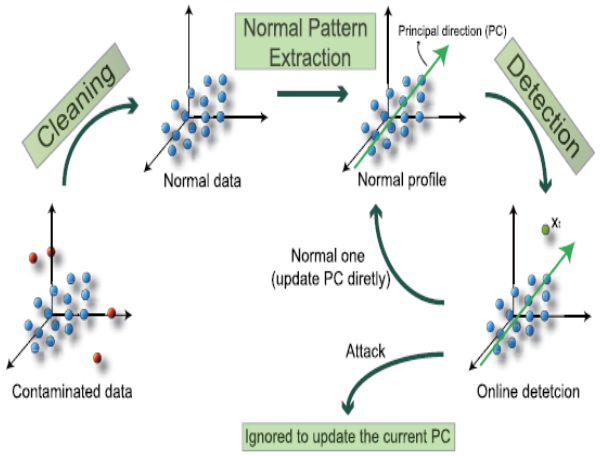

Fig 1: Online Anomaly Detection

	TP Rate	*FP Rate*	*Time(sec.)*
OsPCA	0.9183+_0.0223	0.0427+_0.0054	=1.0E-1
Online OsPCA	0.9133+_0.0327	0.0697+_0.0188	<1.0E-4

Table 1: Online Anomaly Detection Results on the Intrusion Detection Data Set

Note that TP and FP indicate true and false positive rates

IV. CONCLUSION

This work proposed an online anomaly detection method based on over sample PCA. Here shows that the osPCA with LOO strategy will amplify the effect of outliers, and thus can successfully use the variation of the dominant principal direction to identify the presence of rare but abnormal data. When over sampling a data instance, proposed online updating technique enables the osPCA to efficiently update the principal direction without solving eigenvalue decomposition problems. Furthermore, this method does not need to keep the entire covariance or data matrices during the online detection process. Therefore, compared with other anomaly detection methods, this approach is able to achieve satisfactory results while significantly reducing computational costs and memory requirements. Thus, online osPCA is preferable for online large-scale or streaming data problems.

The proposed system can be used many real-world applications such as intrusion or credit card fraud detection to identify deviated data instances and can be extended in various web application and large scale problem for data flow detection.

V. REFERENCES

.

[1] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, "Anomaly Detection Via Online Oversampling Principle Component Analysis",vol.25,no 7.2013

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

[3] D.M. Hawkins, Identification of Outliers. Chapman and Hall, 1980.

[4] 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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] . 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.

[11] E.M. Knox and R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. Int'l Conf. Very Large Data Bases,1998.

[12] H.-P. Kriegel, P. Kro¨ger, E. Schubert, and A. Zimek, "Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data, "Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2009.

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

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