Conditional Anomaly Detection by Principal Component Analysis

[1] P.ROHINI BAI, [2] P.VASANTHI
[1]M.Tech Student, Dept. of CSE, Gates Institute of Technology , Affiliated to JNTUA , Andhra Pradesh, India
[2]Assistant Professor in Dept. of CSE, Gates Institute of Technology, Affiliated to JNTUA, Andhra Pradesh, India

***Abstract*— Most of the existing anomaly detection methods in data mining are typically implemented in batch mode and hence cannot be scaled to large databases. Many critical applications such as credit card fraud detection needs an efficient method to identify the deviated data instances. In the PCA methods designed, the study of outliers is based on the derived directions. But, in PCA methods, the addition or removal of a data instance deviates the target from the principal directions. In this paper, We propose an Online Oversampling PCA where the adding or removing a single outlier instance will not affect the resulting principal direction of the data. 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. The proposed scheme focus on extracting and approximating the dominant principal direction in an online fashion, instead of calculating multiple eigenvectors.**

***Keywords:* osPCA, online oversampling, outlier detection, anomaly .**

I. INTRODUCTION

A well-known definition of "outlier" is given in: "an observation which deviates so much from other observations". Outlier detection aims to identify a small group of instances which deviate remarkably from the existing data. 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. how to determine anomaly of unseen data (or events) draws attention from the researchers in data mining and machine learning communities. anomaly detection needs to solve an unsupervised yet unbalanced data learning problem.

We propose an Over sampling method using online updating technique which 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.

II. EXISTING SYSTEM

The existing outlier detection strategies are classified into statistical, distance and density based methods. using statistical approaches, the data is made to follow some standard and predetermined distributions, and the instance which deviate from this distributions are treated as outliers. For distance-based methods, the distances between each data point of interest and its neighbours are calculated. If the result exceeds above some predetermined threshold, the target instance will be considered as an outlier. In the density-based methods a local outlier factor is derived which measure the outlierness of each data instance. The LOF has the ability to estimate the local data structure via density estimation.

The another anomaly detection approaches such as ABOD calculates the variation of the angles between each target instance and remaining data points. The outliers produce a smaller angle variance than the normal ones. A Fast ABOD algorithm is also proposed to generate an approximation to the original ABOD. The Fast ABOD algorithm considers the variance of the angles between the target instance and its k-nearest neighbours. The above discussed existing methods are implemented in batch mode, and cannot be extended easily to anomaly detection problems with streaming data or online settings. The computational cost or memory requirements might not always satisfy online detection scenarios. An online kernel density algorithm is proposed but requires at least $O(np+p^2)$ for computational complexity. The proposed online detection techniques require only $O(n)$ for both computation and memory costs.

The basic Principal Component Analysis(PCA) derives the outliers based on the principal directions. PCA is an unsupervised dimension reduction method, which determines the principal directions of the data distribution. In order to obtain these principal directions, one needs to construct the data covariance matrix and calculate its dominant eigenvectors. These eigenvectors will be the most informative among the vectors in the original data space, and are thus considered as the principal directions. PCA is formulated as the following optimization problem.

$$\max_{\mathbf{U}\in\mathbb{R}^{p\times k},\|\mathbf{U}\|=\mathbf{I}} \sum_{i=1}^{n} \mathbf{U}^{\top}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^{\top}\mathbf{U},$$

Using the derived principal dimensions, the calculation of data mean or the least squares solution of the associated linear regression model is both sensitive to outliers. removing (or adding) an abnormal data instance will affect the principal direction of the resulting data than removing (or adding) a normal one does. This strategy is termed as "Leave One Out" , used to calculate the principal direction of the data set without the target instance present and that of the original data set. 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 said approach works well with moderate data set size, and is not preferable when the data set size is large. Hence, to address this practical problem, we advance the "oversampling" strategy to duplicate the target instance, and we perform an oversampling PCA (osPCA) on such an

oversampled data set. 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.

III. PROPOSED SYSTEM

For practical anomaly detection problems, the size of the data set is typically large, and thus it is not easy to observe the variation of principal directions caused by the presence of a single outlier. For the n-PCA algorithm, we 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. We introduce a Online Oversampling PCA(osPCA) which addresses the above issues and also able to detect the presence of abnormal instances according to the associated principal directions even in the case of large data sets. The applied principal direction method does not require any need to derive the Eigen vectors.

System Architecture:

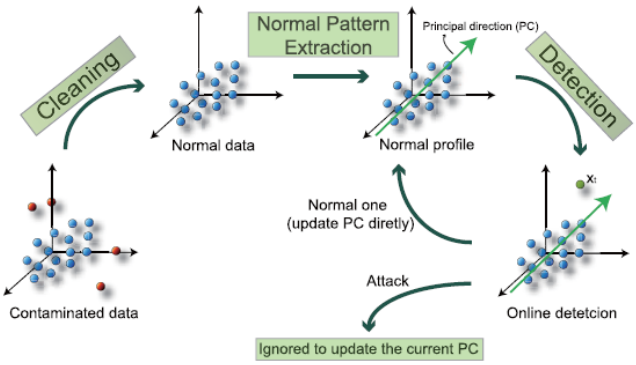

Fig 1: Architecture for Online Anomaly Detection

In contrast to the existing PCA algorithm, the osPCA algorithm duplicate the target instance multiple times, so that the effect of the outlier is amplified rather than the normal data. The detection of the anomaly for each target instance can be done without sacrificing the computational and memory requirements.

If we oversample the target instance *n* times, the associated PCA can be given as follows:

$$\Sigma_{\tilde{\mathbf{A}}}\tilde{\mathbf{u}}_t = \lambda\tilde{\mathbf{u}}_t,$$

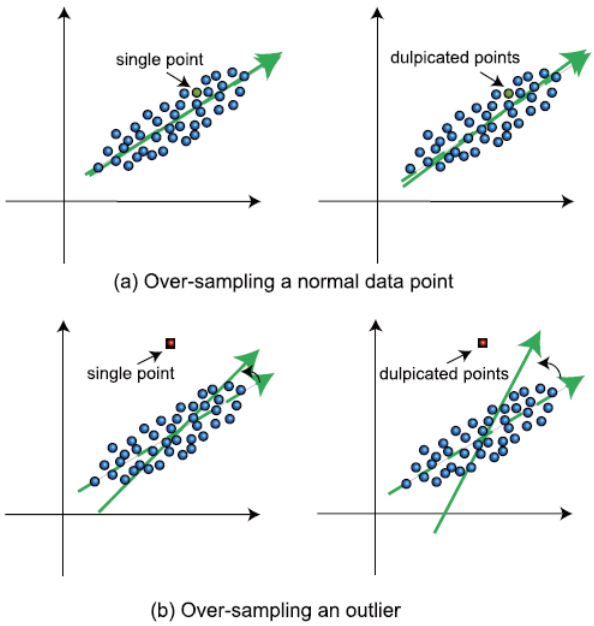

Fig 2: The effect of oversampled normal data or an outlier

For the osPCA, the oversampling ratio *r*, should be determined by the user. This parameter cannot be defined in advance as one cannot cross perform or compare with similar approaches. Although the well known power method is able to produce approximated PCA results, it needs the storage of the covariance matrix and cannot be easily extended to applications with streaming data or online settings.

IV. SYSTEM DESIGN

The proposed system can be divided into 3 altitudes namely Filtering, Clustering, Anomaly detection.

Filtering

In practical scenarios, an initial classifier is updated by the newly received normal or outlier data . Very recent data is considered, as the training data that is collected in advance can be contaminated by noise or incorrect data labelling.

Clustering

The clusters are formed for input data instances and then the outlier calculation is applied for each cluster to find the outlier exactly.

Anomaly detection

At this level, the outliers are detected based on the user input. The system calculates the S_T value for the user input data. Then compares this S_T with the threshold value. If the St 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.

V. CONCLUSION

The proposed over sampling data instance, uses online updating technique that enables the osPCA to efficiently update the principal direction without solving Eigen value

decomposition problems. 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. This method does not need to keep the entire covariance or data matrices during the online detection process.

The proposed method can be extended to various web applications and large scale problem for data flow detection.

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] 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.

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

[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] 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.