IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. II (Jan. 2014), PP 86-90
www.iosrjournals.org

Anomaly Detection using multidimensional reduction Principal Component Analysis

Krushna S.Telangre, Prof.S.B.Sarkar
Department Of Computer Engineering Stes's Skn-Sinhgad Institute Of Technology Lonavala,India
Department Of Computer Engineering Stes's Skn-Sinhgad Institute Of Technology Lonavala,India

Abstract: *Anomaly detection has been an important research topic in data mining and machine learning. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. However, most 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. In this paper, we propose multidimensional reduction principal component analysis (MdrPCA) algorithm to address this problem, and we aim at detecting the presence of outliers from a large amount of data via an online updating technique. Unlike prior principal component analysis (PCA)-based approaches, we do not store the entire data matrix or covariance matrix, and thus our approach is especially of interest in online or large-scale problems. By using multidimensional reduction PCA the target instance and extracting the principal direction of the data, the proposed MdrPCA allows us to determine the anomaly of the target instance according to the variation of the resulting dominant eigenvector. Since our MdrPCA need not perform eigen analysis explicitly, the proposed framework is favored for online applications which have computation or memory limitations. Compared with the well-known power method for PCA and other popular anomaly detection algorithms*

Index Terms: *Anomaly detection, online updating, least squares, oversampling, principal component analysis*

I. INTRODUCTION

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes interchangeably. Anomaly detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities.

The importance of anomaly detection is due to the fact that anomalies in data translate to significant (and often critical) actionable 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.

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, we observe that removing (or 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, we 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.

We note that 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, we advance the "oversampling" strategy to duplicate the target instance, and we perform an oversampling PCA (osPCA) on such an oversampled 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 oversampling 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.

An online updating technique for 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. Moreover, many learning algorithms encounter the "curse of dimensionality" problem in a extremely high-dimensional space. In osPCA method, although we are able to handle high-dimensional data since we 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.

So as proposed to solution to this problem is MdrPCA i.e. multi dimensional reduction PCA. We will MdrPCA as it not easy to use linear models such as PCA to estimate the data distribution if there exists multiple data clusters. Using MdrPCA we can handle multiple cluster, high dimensions, curse of dimensionality problem.

II. LITERATURE SURVEY

In past few years many anomaly detection algorithms has been proposed. These approaches are broadly divided into three categories:-

- **Statistical Approach:-**

Statistical approaches [1], [3] assume that the data follows some standard or predetermined distributions, and this type of approach aims to find the outliers which deviate from such disributions. 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.

- **Distance- Based Approach:-**

For distance-based methods [4], 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., multiclustered structure). In such cases, this type of approach will result in determining improper neighbors, and thus outliers cannot be correctly identified.

- **Density- Based Approach:-**

To alleviate the aforementioned problem, density-based methods are proposed [5]. 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. The most important property of the LOF is the ability to estimate local data structure via density estimation. This allows users to identify outliers which are sheltered under a global data structure. However, it is worth noting that the estimation of local data density for each instance is very computationally expensive, especially when the size of the data set is large. Beside this some recently proposed approaches are given below:-

- **Angle Based Outlier Detection (ABOD) Method:-**

A novel approach named ABOD (Angle-Based Outlier Detection) [6] and some variants assessing the variance in the angles between the difference vectors of a point to the other points. This way, the effects of the "curse of dimensionality" are alleviated compared to purely distance based approaches. A main advantage of this approach is that ABOD method does not rely on any parameter selection influencing the quality of the achieved ranking. As compare with well established distance-based method LOF performance of very well especially on high dimensional data. Consequently, a fast ABOD algorithm is proposed to generate an approximation of the original ABOD solution. The difference between the standard and the fast ABOD approaches is that the latter only considers the variance of the angles between the target instance and its k nearest neighbors. However, the search of the nearest neighbors still prohibits its extension to large scale problems (batch or online modes), since the user will need to keep all data instances to calculate the required angle information.

- **Incremental Local Outlier Detection for Data Streams:-**

Incremental LOF (Local Outlier Factor) [7] algorithm provides equivalent detection performance as the iterated static LOF algorithm (applied after insertion of each data record), while requiring significantly less computational time. In addition, the incremental LOF algorithm also dynamically updates the profiles of data points. This is a very important property, since data profiles may change over time. Incremental LOF provides theoretical evidence that insertion of a new data point as well as deletion of an old data point influence only limited number of their closest neighbors and thus the number of updates per such insertion/deletion does not depend on the total number of points *N* in the data set. It is found that their computational cost or memory requirements might not always satisfy online detection scenarios. For example, while the incremental LOF in [7] is able to update the LOFs when receiving a new target instance, this incremental method needs to maintain a preferred (or filtered) data subset. Thus, the memory requirement for the incremental LOF is O (np), where n and p are the size and dimensionality of the data subset of interest, respectively.

- **Online Anomaly Detection using KDE:-**

Large backbone networks are regularly affected by a range of anomalies. Online anomaly detection algorithm based on Kernel Density Estimates [8]. Algorithm sequentially and adaptively learns the definition of normality in the given application, assumes no prior knowledge regarding the underlying distributions, and then detects anomalies subject to a userset tolerance level for false alarms. Comparison with the existing methods of Geometric Entropy Minimization, Principal Component Analysis and One-Class Neighbor Machine demonstrates that the proposed method achieves superior performance with lower complexity. But an online kernel density estimation for anomaly detection algorithm requires at least $O(np^2+p^2)$ for computation complexity [8]. In online settings or large-scale data problems, the abovementioned method might not meet the online requirement, in which both computation complexity and memory requirement are as low as possible.

- **Oversampling PCA (osPCA):-**

The proposed osPCA [9] 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.

III. PROBLEM STATEMENT

However, most anomaly detection methods are typically implemented in batch mode, and thus can not be easily extended to large-scale problems without sacrificing computation and memory requirements. In base paper, they propose an online oversampling principal component analysis (osPCA) algorithm to address this problem, and they aim at detecting the presence of outliers from a large amount of data via an online updating technique. Unlike prior principal component analysis (PCA)-based approaches, we do not store the entire data matrix or covariance matrix, and thus our approach is especially of interest in online or large-scale problems. By oversampling the target instance and extracting the principal direction of the data, osPCA allows us to determine the anomaly of the target instance according to the variation of the resulting dominant eigenvector. Since osPCA need not perform eigen analysis explicitly, the proposed framework is favored for online applications which have computation or memory limitations.

- **Limitations of Existing Methods**

1. Normal data with multi clustering 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.
2. Moreover, many learning algorithms encounter the “curse of dimensionality” problem in a extremely high-dimensional space.
3. Although we are able to handle high-dimensional data since we 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.

IV. OBJECTIVE

In this project we have main aim is to present the extended method for Anomaly Detection using multi dimensional reduction Principal Component Analysis with improved reliability and performance:

To present literature review different methods Principal Component Analysis.

To present the present new framework and methods.
To present the practical simulation of proposed algorithms and evaluate its performances.
To present the comparative analysis of existing and proposed algorithms in order to claim the efficiency.

V. SCOPE OF PROPOSED SYSTEM

A major problem is *the curse of dimensionality.*

If the data x lies in high dimensional space, then an enormous amount of data is required to learn distributions or decision rules. Example: 50 dimensions. Each dimension has 20 levels. This gives a total of cells. But the no. of data samples will be far less. There will not be enough data samples to learn. One way to deal with dimensionality is to assume that we know the form of the probability distribution. For example, a Gaussian model in N dimensions has N + N (N-1)/2 parameters to estimate. Requires data to learn reliably. This may be practical. One way to avoid the curse of dimensionality is by projecting the data onto a lower-dimensional space.

Multidimensional reduction (Mdr) methods are projection techniques that tend to preserve, as much as possible, the distances among data. Therefore data that are close in the original data set should be projected in such a way that their projections, in the new space (output space), are still close. This depends only on the distances between data. When the rank order of the distances in the output space is the same as the rank order of the distances in the original data space, stress is zero. Stress is minimized by iteratively moving the data in the output space from their initially randomly chosen positions according to a gradient-descent algorithm. The intrinsic dimensionality is determined in the following way. The minimum stress for projections of different dimensionalities is computed. Then a plot of the minimum stress versus dimensionality of the output space is performed. ID is the dimensionality value for which there is a knee or a flattening of the curve.

VI. METHODOLOGY

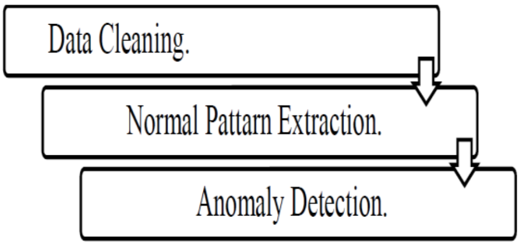

Figure 1.Proposed Methodology

Data Cleaning:-

In the data cleaning phase, our goal is to filter out the most deviated data using our 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. It can be anything as specified by user may be 5% or 10% as per the requirement of application. And that much amount of instances is removed from the original dataset.

Normal Pattern Extraction:-

In this phase normal pattern of remaining instances of previous phase is derived and Principal Direction is computed by using Principal Component Analysis.

Online Anomaly Detection:-

After ignoring of specific amount of the training normal data after this data cleaning process, here we use the smallest score of outlierness of the remaining training data instances as the threshold for outlier detection. More specifically, in this phase of online detection, we use this threshold to determine the anomaly of each received data point. If smallest score 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 MdrPCA model accordingly.

.

VII. Proposed System

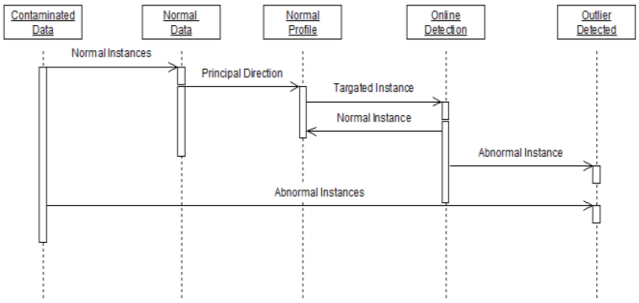

Figure 2. Sequence Diagram of Proposed System

Steps explanation-

Step 1:- Data cleaning is done on contaminated data. In this the instances that are deflecting the principal direction in large amount is removed by using Oversampling Principal Component Analysis. And normal data is secured.

Step 2:- Once we get a normal data, we instances i.e the principal direction is computed.

Step 3:- Analyzing threshold value for detecting the outliers.

Step 4:- Smallest score of outlierness for new instance is measured.

Step 5:- Smallest score of outlierness of new instance is compared with the threshold value, and if the smallest score of outlierness is outlier and ignored to update the current PC.

Step 6:- If the instance is normal i.e the value of smallest score of outlier is smaller than defined threshold value, it is detected as normal instance and will extract the normal pattern for the remaining greater than threshold value then the instance is detected as updated directly to the PC.

VIII. CONCLUSION

This paper proposed multidimensional reduction principal component analysis (MdrPCA).that is quiet different than osPCA. Performance of MdrPCA is better than osPCA because we have found that the osPCA strategy amplifies the effect of outliers, and thus we can successfully use the variation of the dominant principal direction to identify the presence of rare but abnormal data. But we are extending the same strategy of osPCA for the detection of outlier in data with high dimensional space using online updating technique with the help of MdrPCA.

REFERENCES

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

[2] 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.

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

[4] 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.

[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] A. Lazarevic, L. Erto¨ z, V. Kumar, A. Ozgur, and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," Proc. Third SIAM Int'l Conf. Data Mining, 2003.

[7] 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.

[8] S. Rawat, A.K. Pujari, and V.P. Gulati, "On the Use of Singular Value Decomposition for a Fast Intrusion Detection System," Electronic Notes in Theoretical Computer Science, vol. 142, no. 3, pp. 215-228, 2006.

[9] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang," Anomaly Detection via Online Oversampling Principal Component Analysis" IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 7, July 2013