International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Anomaly Detection of Online Data using Oversampling Principal Component Analysis

Supriya A. Bagane[1], J. L. Chaudhari[2]

[1]ME, Department of Computer Engineering, Bhivarabai Sawant Institute of Technology and Research, Pune, India

[2]Assistant Professor, Department of Computer Engineering, Bhivarabai Sawant Institute of Technology and Research, Pune, India

Abstract: ***Anomaly detection is very important topic in data mining and machine learning. This technique is helpful in many real world applications such as intrusion or credit card fraud detection, fault detection in safety critical systems, and military surveillance for enemy activities. Anomaly detection is basically used to find the patterns in data that do not conform to their expected behavior. Such patterns are termed as anomalies, outliers, discordant observations, exceptions, aberrations etc in different application domains. From all these terms anomalies and outliers can be used interchangeably. Outlier detection methods can be used to deal with extremely unbalanced data distribution problems. Most of the anomaly detection methods are implemented in batch mode due to which they cannot be extended to large scale problems. If we extend them to large scale problems, they will result in sacrificing computation and memory requirements. To tackle this problem we proposed oversampling Principal Component Analysis (osPCA) scheme in this paper. This technique aims at detecting the presence of outliers from large amount of data. In previously proposed Principal Component Analysis methods, it is required to store entire data matrix or covariance matrix, but this is not the case with our osPCA approach. So it can be extended to large scale or online problems. Principal Component Analysis is used to find the principal direction of the data and oversampling technique will duplicate the target instance multiple times to amplify the effect of outliers. By oversampling the target instance and extracting the principal directions of the data the osPCA allows us to determine the anomaly in target instance according to the variations in the resulting dominant eigenvector. This online updating technique allows us to efficiently calculate dominant eigenvector without eigen analysis or storing entire covariance matrix. Compared with the other anomaly detection methods the required computational costs and memory requirements are significantly reduced.***

Keywords: *Anomaly detection, principal Component Analysis, outlier, oversampling*

1. Introduction

Anomaly detection is very important technique in data mining. Anomaly or outlier detection is basically used to find out the group of instances which deviate from original data. Anomaly detection is the primary step in many data mining applications. Various methods are available for outlier detection [1]. Numbers of applications are there, where anomaly detection is important such as homeland security, credit card fraud detection, intrusion and insider threat detection in cyber security, fault detection or malignant diagnosis.

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These patterns which do not conform to expected behavior are referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains [5].

The deviated data is not available in large amount but its presence may affect the solution model such as the distribution or the principal directions of the data [2]. For example, the calculation of data mean or the least square solution of the associated linear regression model is both sensitive to outliers.

Leave One out strategy can be used to calculate the principal directions of the data set without the presence of target instance and original data set. So it is helpful to determine the variation of resulting principal directions. Then by differentiating these two eigen vectors the anomaly of target instance is calculated. 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 decremental PCA (d-PCA) based approach for anomaly detection. This approach works well for applications with small data set size but it might not be significant when the size of the data set is large. In that case, it will produce negligible difference in eigen vectors hence it is not efficient to apply dPCA. To address this practical problem the oversampling strategy is used to duplicate target instance and to perform 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[3]. The effect of outlier instance will be amplified due to its duplicates present in Principal Component Analysis formulation and due to which it is easy to detect outliers or anomalies.

2. Related Work

The existing approaches can be divided into three categories: distribution or statistical, distance based and density based methods [1][2][3][4][5].

In statistical approach, it is assumed that the data follows some standard or predetermined distributions which aim to find outliers deviated from such distributions. Most distribution models are assumed univariate and thus the lack of robustness for multidimensional data is a concern.

In distance based method, basically the distances between each data point of interest and its neighbors are calculated. There is some predetermined threshold, if the result is above this threshold the target instance is considered as anomaly. Here there is no need of prior knowledge of data distribution. These approaches might encounter problems when data distribution is complex (e.g. multiclustered structure). In this type of cases the approach will result in determining improper neighbors and thus outliers cannot be correctly identified.

To deal with the problem faced in distance based methods, density based methods are proposed. In this method density based Local outlier factor (LOF) is used to measure the outlierness of each data instance. The LOF determines the degree of outlierness based on the local density of each data instance which provides suspicious ranking scores for all samples. The Local Outlier Factor is having the ability to estimate local data structure via density estimation which allows users to identify outliers sheltered under global data structure. It is important to note that the estimation of local data density for each instance is computationally very expensive, especially when size of data set is large.

Besides the above work, some anomaly detection approaches are recently proposed. Among them, the angle-based outlier detection (ABOD) method is very unique. Simply speaking, ABOD calculates the variation of the angles between each target instance and the remaining data points, since it is observed that an outlier will produce a smaller angle variance than the normal ones do. It is not surprising that the major concern of ABOD is the computation complexity due a huge amount of instance pairs to be considered. 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 [2].

There are some problems with this existing system such as the anomaly detection methods specified here in this existing system are typically implemented in batch mode so it is not possible to extend this anomaly detection to large scale or online data. Most distribution models are assumed univariate in existing system due to which there is lack of robustness when high dimensional data is concerned. Moreover these models are typically implemented in the original data space directly; their solution models might suffer from the noise present in the data.

3. Anomaly Detection using Principal Component Analysis

The problems which are there with existing system do not allow anomaly detection with large scale or online data. Also with large data there exist memory and computational complexity. To overcome this problem we can use oversampling technique with Principal Component Analysis (PCA).

3.1 Principal Component Analysis

Principal Component Analysis is a dimension reduction method which determines the principal directions of data. To calculate these principal directions, it is required to construct data covariance matrix and calculate its dominant eigen vectors. These eigenvectors are very informative among the vectors which are there in original data space. So they are considered as principal directions [2]. Principal component components are the linear combinations of p random variables X_1, X_2,…X_p with three important properties; the principal components are uncorrelated, the first principal component has the highest variance, the second principal component has second highest variance and so on, and the total variation in all the principal components combined is equal to the total variation in the original variables X_1, X_2, .., X_p. They are easily obtained from an eigen analysis of covariance matrix or the correlation matrix of X_1, X_2, …, X_p [4].

Principal components from the covariance matrix and the correlation matrix are usually not the same. In addition they are not simple functions of the others. When some variables are in a much bigger magnitude than others, they will receive heavy weights in the leading principal components. For this reason, if the variables are measured on scales with widely different ranges or if the units of measurement are not commensurate, it is better to perform PCA on the correlation matrix [4].

3.2 The use of PCA for Anomaly Detection

Here we will study the variation of principal directions when we add or remove a data instance and also how we use this property to determine the outlierness of target data point. To see the illustration of this check fig. 1.

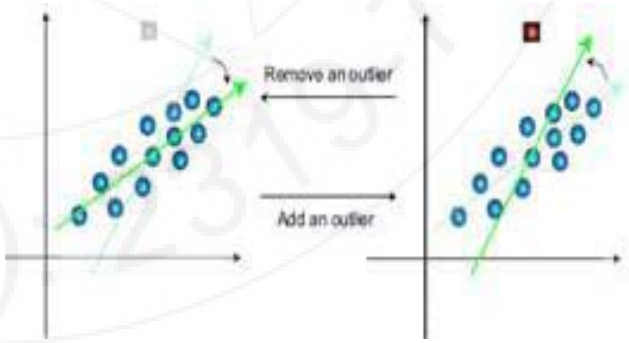

Figure 1: The effect of adding or removing outlier on the principal directions

We note that the clustered blue circles in Fig. 1 represent normal data instances, the red square denotes an outlier, and the green arrow is the dominant principal direction. From Fig. 1, we see that the principal direction is deviated when an outlier instance is added. More specifically, the presence of such an outlier instance produces a large angle between the

resulting and the original principal directions. On the other hand, this angle will be small when a normal data point is added. Therefore, we will use this property to determine the outlierness of the target data point using the LOO strategy [2].

3.3 Oversampling PCA

When the data set is large, the adding or removing of a single outlier instance will not significantly affect the resulting principal direction of the data. So this oversampling strategy is used and presents an oversampling Principal Component Analysis (osPCA) algorithm for large scale anomaly detection problems.

The osPCA scheme duplicates the target instance multiple times [2]. Also this technique will not store entire covariance matrix like previously available methods. By oversampling the target instance, osPCA allows to determine the anomaly of target instance with the help of variations in the dominant eigenvector [3].

Oversampling is basically used to amplify the outlierness of each data point. For identifying those outliers by using LOO strategy it is required to duplicate target instances instead of removing it. That means we can duplicate the target instance many times and observe how much variation is there in the principal direction. With this oversampling scheme the principal directions and mean of the data will only be affected slightly if the target instance is a normal data point shown in fig. 2(a). On the contrary, the variations will be enlarged if we duplicate an outlier shown in fig. 2(b).

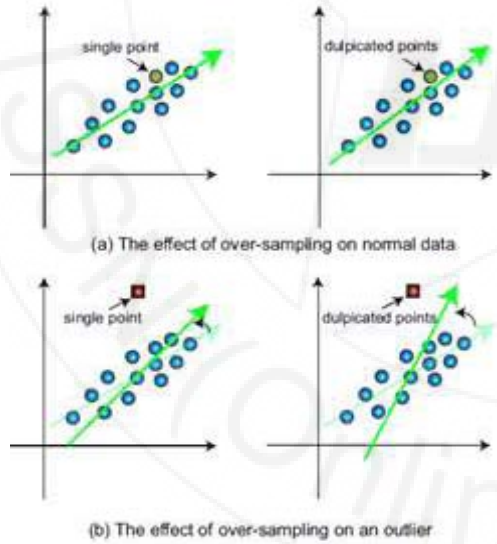

Figure 2: The effect of oversampling on an outlier and normal instances

On the other hand we can also apply oversampling scheme in the LOO strategy. The main idea is to enlarge the effect between the normal data point and an outlier. This is possible with the help of oversampling PCA [6]. This oversampling PCA produces number of instances of target data. 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, 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. Our proposed oversampling PCA (osPCA) together with an online updating strategy will address the above issues, as we discussed [2].

3.4 Online Anomaly Detection for Practical Scenario

For Online Anomaly detection applications like spam mail filtering, it is desired to design an initial classifier using the training normal data, and this classifier is updated by newly received normal or outlier data accordingly.

In practical scenario the training normal data collected in advance can be contaminated by noise of incorrect data labeling. To build a simple and effective online detection model it is required to disregard these potentially deviated data instances from the training set of normal data.

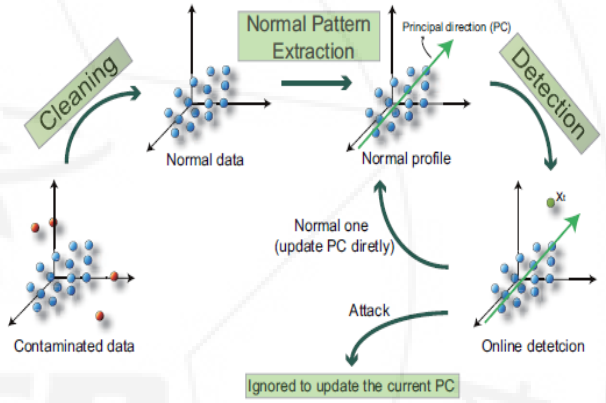

Figure 3: Online Anomaly Detection Framework

The flow chart of this scheme is shown in fig. 3. Two main phases are there which are needed to implement this technique: Data cleaning and online detection. In the data cleaning phase, the goal is to filter out the most deviated data using osPCA before performing online anomaly detection. This first phase is offline and the percentage of the training normal data to be disregarded can be determined by the user. In the second phase of online detection, this threshold is used to determine the anomaly of each received data point. In this phase the dominant principal direction of the filtered training normal data which extracted in data cleaning phase is used to detect each arriving target instance [2].

3.5 The Power Method of osPCA

Typically, the solution to PCA is determined by solving an eigenvalue decomposition problem. In the LOO scenario, one will need to solve the PCA and to calculate the principal directions n times for a data set with n instances. This is very computationally expensive, and prohibits the practical use of such a framework for anomaly detection.

It can be observed that, in the PCA formulation with the LOO setting, it is not necessary to recompute the covariance matrices for each PCA. This is because when we duplicate a data point of interest, the difference between the updated covariance matrix and the original one can be easily determined. 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.

4. Conclusion

The proposed method in this paper is an online anomaly detection based on oversampling technique. This osPCA with LOO strategy will amplify 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. Although there are various methods for anomaly detection, the method proposed here 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. Thus our osPCA is preferable for online large scale or streaming data problems.

5. Future Work

Future research will be directed to the following anomaly detection scenarios: normal data with multi clustering structure. This proposed work is to cluster the high dimensional data using enhanced k-means clustering process based on the categorical and mixed data types in efficient manner and apply Online OS PCA to detect outlier. The goal is to use detect outliers on high dimensional categorical data that works well. In addition, also the study the quick updating of the principal directions for the effective computation and satisfying the on-line detecting demand will be made. 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.

6. Acknowledgement

I would like to acknowledge all the people who helped and assisted me throughout my work. First of all I would like to thank respected Principal Dr. D. M. Yadav sir and respected H.O.D. Prof. G. M. Bhandari Mam and my guide Prof. J. L. Chaudhari Mam and all the professors in our department for their constant support.

References

[1] Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining
[2] 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
[3] Priyanka R. Patil, R. D. Kadu, “ A Novel data Mining approach to Calculate Outlier from Large Data set using Advance Principal component Analysis”, Proceedings of IRF International Conference, 5th & 6th February 2014, Pune India. ISBN: 978-93-82702-56-6
[4] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, LiWu chang, “A Novel Anomaly Detection Scheme Based on Principal Component Classifier”, Naval Research Laboratory, Center for High Assurance Computer Systems
[5] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 15:1-15:58, 2009.
[6] Y. Srilaxmi, D. Ratna Kishore, “Online Anomaly Detection under Oversamplig PCA”, International Journal of Science and Research(IJSR), ISSN(Online):2319-7064, Volume 3 Issue 9, September 2014
[7] Y.-R. Yeh, Z.-Y. Lee, and Y.-J. Lee, “Anomaly Detection via Oversampling Principal Component Analysis,” Proc. First KES Int’l Symp. Intelligent Decision Technologies, pp. 449-458, 2009
[8] T. Ahmed, “Online Anomaly Detection using KDE,” Proc. IEEE Conf. Global Telecomm., 2009.
[9] D.M. Hawkins, “Identification of Outliers”. Chapman and Hall, 1980.