Detection of Anomalies using Online Oversampling PCA

Miss Supriya A. Bagane, Prof. Sonali Patil

***Abstract*— Anomaly detection is the process of identifying unexpected behavior and it is an important research topic in data mining and machine learning. Many real-world applications require an effective and efficient framework to identify such data instances. So it is required to propose the outlier detection methods. The previously proposed anomaly detection methods are implemented in batch mode, so they cannot be extended to large scale or online problems. In earlier Principal Component Analysis (PCA) formulation it is required to store entire data matrix or covariance so the computational cost and memory requirements will get increased. To handle this issue the method proposed here is useful that is oversampling Principal Component Analysis (osPCA). PCA is used to perform dimension reduction which helps to get principal directions of data, based on that anomaly detection is performed. Oversampling effect will duplicate the target instance multiple times to amplify the effect of outliers. The method can also be applied for normal data with multiclustering structure. Here in osPCA online updating technique is used to handle online, large scale, or streaming data problems.**

***Index Terms*— Anomaly detection, principal Component Analysis, outlier, oversampling**

I. INTRODUCTION

Anomalies are the patterns in data that do not conform to their expected behavior and Anomaly detection refers to the problem of finding those patterns in data. The patterns which are different from their expected behavior are referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains.

Anomaly detection is very important in data mining. Anomaly or outlier detection is basically used to find the group of instances which deviate from original data [1]. Anomaly detection is the first step in number of data mining applications. Numbers of methods are available for outlier detection and 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.

Anomalies are also referred to as outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains [5]. The terms anomalies and outliers can be used interchangeably.

Anomalous 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 [1]. Consider the example of calculation of data mean or the least square solution of the associated linear regression model both are sensitive to outliers. They may get affected by 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. This thing is helpful in determining the variation of resulting principal directions. After this by differentiating these two eigen vectors the anomaly of target instance is calculated. To address this problem which is there in existing system we use the oversampling strategy to duplicate target instance and to perform oversampling PCA (osPCA). The effect of 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 method proposed here works on dimensionality reduction provided by PCA and the calculation of threshold value. Using this threshold value as a parameter the outlier detection is performed. So it is not required to store entire data or covariance matrix. Due to which computation and memory requirements will get reduced.

II. EXISTING SYSTEM

The available approaches are divided into three categories: distribution or statistical, distance based and density based methods [1][2][3][4][5]. One more approach Angle Based Outlier detection is also popular [1].

Statistical approach assumes that the data follows some standard or predetermined distributions which aim to find outliers deviated from those standard distributions. Most distribution models are assumed here are univariate and so there is lack of robustness for multidimensional data.

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

Manuscript received July 2015.

***Miss. Supriya A. Bagane**, Computer Engineering, Bhivarabai Sawant Institute of Technology and Research., Pune, India, 9049720859*

***Prof. Sonali Patil**, Computer Engineering, Bhivarabai Sawant Institute of Technology and Research, Pune, India, 8308885446,*

ISSN: 2278 – 7798

The problem faced in distance based methods we have density based methods. This method uses density based Local outlier factor (LOF) to measure the outlierness of each data instance. The LOF finds the degree of outlierness by using the local density of each data instance which provides suspicious ranking scores for all samples. The Local Outlier Factor estimates local data structure using density estimation which allows users to identify outliers which are there in global data structure. The estimation of local data density for each instance is computationally very expensive, when size of data set is large.

Along with these three approaches one more approach is quite popular and unique that is Angle Based Outlier Detection (ABOD). Simply speaking, ABOD calculates the variation of the angles between each target instance and the remaining data points, because it is observed that an outlier will produce a smaller angle variance than the normal ones do. The main point to note regarding ABOD is the computation complexity due to a huge amount of instance pairs to be considered. So, a fast ABOD algorithm is proposed to generate an approximation of the original ABOD solution. However, the search of the nearest neighbours 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 [1][10].

There are some problems with the methods specified here such as the anomaly detection methods specified here are typically implemented in batch mode so it is not possible to extend this anomaly detection to large scale or online data. 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; due to that their solution models might suffer from the noise present in the data.

III. ANOMALY DETECTION USING OVERSAMPLING PCA

The problems faced with existing system do not allow anomaly detection with large scale or online data. With large data there exist the problem of high memory and computational requirements. To overcome this problem we can use oversampling technique with Principal Component Analysis (PCA).

Outlier detection is the process of identifying unusual behavior which is different from normal one. 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 somewhat different due to its computations. In order to overcome the computational complexity, online oversampling PCA has been used. The algorithm uses Online updating technique of the principal directions for the effective computation and satisfying the online detecting demand and also oversampling will improve the effect 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 purpose.

A. *Principal Component Analysis*

Principal component analysis is appropriate when you have obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses.

Principal Component Analysis uses dimension reduction and also determines the principal directions of data. We have used this method to calculate principal directions of data. To generate principal directions, we have to construct data covariance matrix and calculate its dominant eigen vectors. These eigenvectors are very informative among all the vectors which are there in original data space. So they are considered as principal directions [1].

By using the dimensionality reduction, the dimensionality of the data set is reduced to get the reduced items. These items are considered as attributes. By calculating the threshold value of all these attributes anomaly detection is performed. Minimum threshold is chosen to perform this task.

B. *Use of PCA for Anomaly Detection*

By using PCA we can find principal directions of data instances. Here we will study the variation of principal directions to perform detection. When we add or remove a data instance there is variation in principal direction. When a normal instance is added there is a slight change in the direction. When an outlier instance will get added there is drastic change in principal direction. Based on this detection is performed. The illustration of this is given fig. 1.

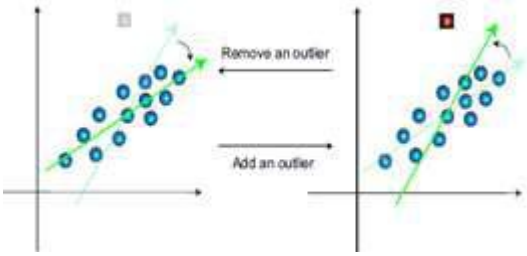

Fig. 1 The effect of adding or removing outlier on the principal directions

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, this is used to determine the outlierness of the target data point using the LOO strategy [1].

C. *Oversampling PCA*

When we have large data set at that time the adding or removing of a single outlier instance will not significantly

ISSN: 2278 – 7798

affect the resulting principal direction of the data. Due to which oversampling strategy is used and presents an oversampling Principal Component Analysis (osPCA) algorithm for online or large scale anomaly detection problems.

The osPCA scheme duplicates the target instance multiple times [1]. 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].

The osPCA will duplicate the target instances n no of times from data set and computes the score of outlier. If the score is above the threshold it is determined to be an outlier. The solving of PCA to calculate the principal directions n times using n instances and it is very costlier and it prohibits the online data in anomaly detection.

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).

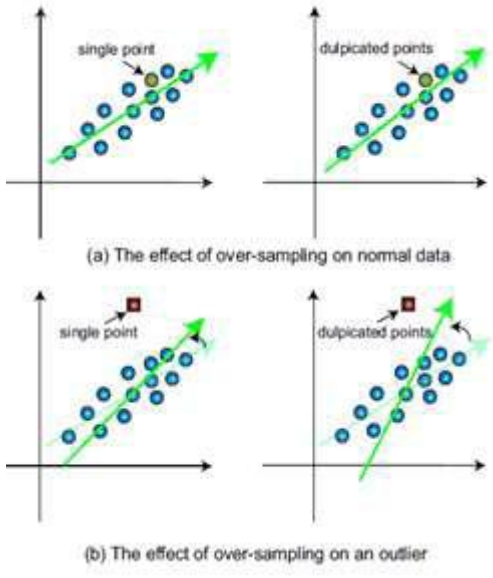

Fig. 2: The effect of oversampling on an outlier and normal instances

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].

IV. ARCHITECTURE OF PROPOSED SYSTEM

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.

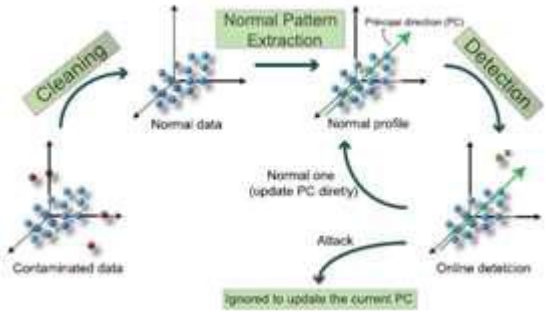

Fig. 3 Online Anomaly Detection Framework

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 filters out the most deviated data using osPCA. 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, 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 [1]

.

V. IMPLEMENTATION DETAILS

This online Anomaly detection can be carried out in three phases: Data Cleaning, Detection and Clustering.

A. Data Cleaning

Data preprocessing is the first step in data mining which involves transformation of raw data into an understandable format. Real world data is incomplete, inconsistent and/or lacking in certain behaviors and may contain errors.

A set of data instances from original data set is selected as predefined input. The PCA formulation is applied on the data set selected. The PCA formulation helps to perform reduction. Only the reduced attributes are used for anomaly detection, instead of using entire data matrix or covariance matrix. This is shown in Fig.4.

ISSN: 2278 – 7798

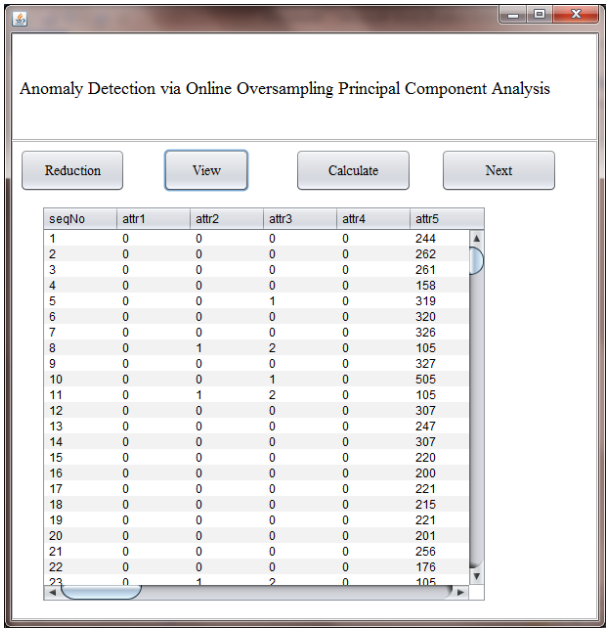

Fig. 4 Dimensionality Reduction

B. *Detection*

This phase is for detecting the outlierness of the user input. After reduction of data set totally reduced attributes are calculated. Next the threshold is calculated for each attribute. Smallest value is selected as threshold for anomaly detection. When the user giving the input to the system, the system calculate the threshold value of that input, then compares that new value with the threshold value which is previously calculated.

If the St value of the new data instance is greater than threshold value, then 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 data instance is updated accordingly. Based on these threshold values predicted outliers are calculated first as shown in Fig. 5.

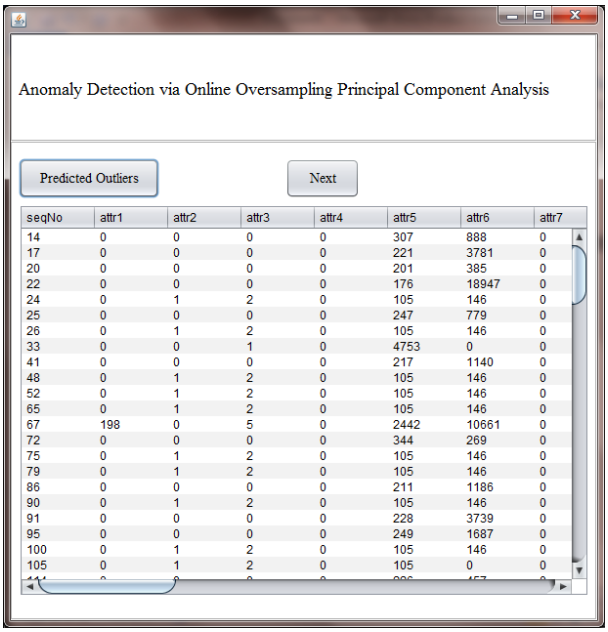

Fig. 5 Predicted Outliers

C. *Clustering*

The training data will be selected by assumption. So it may happen that some outlier data may be considered as normal data in the previous method due to training data. So the clustering method is used to tackle this problem. The clusters are created for input data instances and then the outlier calculation is done for each cluster to find the outlier exactly. By using simple clustering algorithm the data is divided into multiple clusters and then detection is performed based on threshold values. Fig. 6 shows this where outliers from each cluster are calculated.

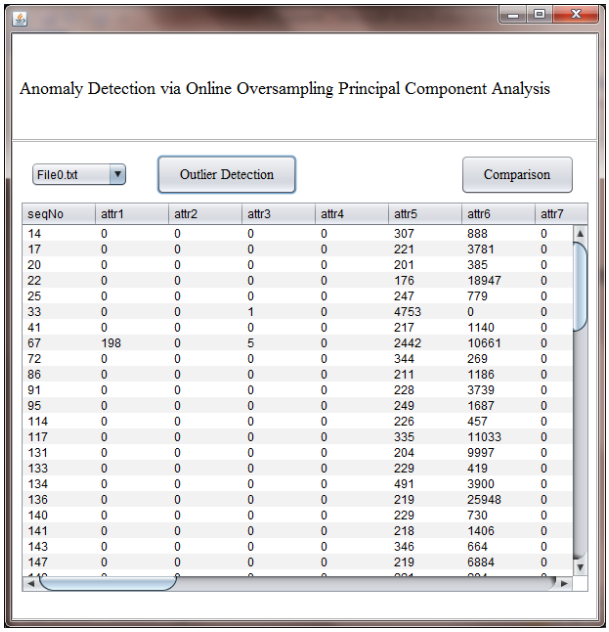

Fig. 6 Proposed System

VI. RESULTS

In existing system, the methods used are implemented in batch mode so they cannot be extended to large scale problems. If they are extended to large scale problems, they will result in sacrificing computation and memory requirements. The proposed system will overcome this limitation. By using oversampling PCA, to perform large scale or online anomaly detection, memory requirements will get reduced.

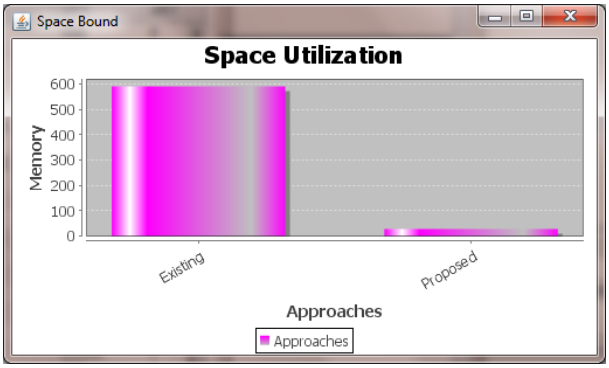

Fig. 7 Graph comparing Existing and Proposed System

ISSN: 2278 – 7798

VII. Conclusion

The method proposed here is an online updating technique with oversampling PCA. Oversampling PCA with LOO strategy will amplify the effect of outliers. 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. Further in future work this will extend the same strategy of oversampling Principal Component Analysis for the detection of outlier in data with high dimensional space using online updating technique with the help of multidimensional reduction techniques.

Acknowledgment

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. T. K. Nagaraj sir and respected H.O.D. Prof. G. M. Bhandari Mam and my guide Prof. Sonali Patil Mam and all the professors in our department for their constant support.

REFERENCES

[1] 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

[2] Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining

[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.

[10] 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.

ISSN: 2278 – 7798