ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 3; May-June-2017; Page No. 97-101

# An Efficient Novel Data Mining Approach to Calculate Outlier from Large Dataset Using Advance PCA.

**Priyanka R Patil[1], Laxmi R Adhav[2]**

[1] Pune University, Department of Computer Engineering, Sandip Foundation, Nashik, India

*pripatil310@gmail.com*

[2] Pune University, Department of Information Technology,

Sandip Foundation, Nashik, India

*Laxmi.adhav@gmail.com*

Received 04 April 2017; Accepted 10 May 2017

**ABSTRACT**

Outlier detection is an important term in the field of data mining. There are many detection techniques out of which most outlier detection methods are implemented in batch mode means they work on small datasets. Thus such methods cannot be easily used for large-scale problems without computation and memory requirements. Applications like intrusion or credit card fraud detection requires powerful and efficient framework to identify outlier data instances. In this paper, we propose an Advance Principal Component Analysis algorithm which aims at detecting the presence of outliers from a large amount of data via an online updating technique. Like The previous principal component analysis (PCA)-based approaches, we do not store the entire data matrix or covariance matrix, and thus this approach can be useful for the online or large-scale problems. By oversampling and extracting the principal direction of the data, the proposed advance PCA allows determining the anomaly of the target instance according to the variation of the resulting dominant eigenvector.

**Keywords:** Anomaly Detection, Oversampling, Online updating, PCA.

## INTRODUCTION

In the recent years many research is going on the anomaly detection because this work helps to mine the important data from the large data ware house. This outlier detection algorithm is also helpful to identify credit card fraud and intrusion detection. Anomaly detection is the process to identify the anomaly/outlier from the existing dataset. A well known definition of outlier can be defined as an observation which deviates so much from other observation such that it was generated by a different mechanism. Anomaly detection can be useful for the applications such as homeland security, credit card fraud detection, intrusion and insider threat detection or malignant diagnosis. But in this real world applications limited amount of data is available because of which it is difficult to identify the anomaly of the unseen data [1],[2],[3],[4],[5]

Leave one out (LOO) strategy can be use to calculate the principal direction of the data set without the target instance present and that of the original data set. Thus the anomaly can be determined by the variation of the resulting principal directions. The difference between these two eigenvectors indicates the anomaly of the target instance. 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 a decremental PCA (dPCA)-based approach for anomaly detection. It works well for applications with Small data set size, but it might not be significant when the size of the data set is large. It can produce the negligible difference in the eigenvectors hence it is not efficient to apply dPCA. For addressing this practical problem, the "oversampling" strategy is used to duplicate the target instance, and to perform an 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.

The outlier detection algorithms are divided into the three categories.

1. Distribution.
2. Distance.



**Corresponding author: Priyanka R Patil |*

3. Density-based.

In the first approaches [1] the data follows some standard or predetermined distributions, and this aims to find the outliers which deviate from such distributions.

In the second approach distance-based methods [7] the distances between each data point and its neighbors are calculated. If the obtained result is above some predetermined threshold, it will be considered as an outlier. It is not required to have prior knowledge on data distribution. The approach might encounter some problems if the data distribution is complex. In some cases the approach may result in determining improper neighbors, and thus outliers cannot be correctly identified.

In the density-based methods, It uses a density-based local outlier factor (LOF) to measure the outlierness of each data instance [2]. Based on the local density of each data, the LOF determines degree of outlierness, which provides suspicious ranking scores for all samples. The important property of LOF is to estimate local data structure via density estimation. This allows users to identify outliers which are under a global data structure.

The existing system has some disadvantages such as it does not support the large dataset, it requires large iteration and it is not capable of handling high dimensional data. To overcome the problems of the existing system the new advanced PCA technique is designed. It is the data mining schema to identify anomaly using generic analysis method to detect miss behavior of data in large data ware house. I propose an advanced principal component analysis technique to address this problem, and I aim at detecting the presence of outliers from a large amount of data via an online updating technique. By oversampling the target instance and extracting the principal direction of the data, the proposed osPCA allows us to determine the anomaly of the target instance according to the variation of the resulting dominant eigenvector. For the analysis purpose I Compared power method for PCA and other popular anomaly detection algorithms.

## II.LITERATURE SURVEY

D.M.Hawkins [1] proposes several methods for outlier detection in 1980 and the difference between uni and multivariate techniques and parametric and nonparametric procedures. one of the drawback of the Statistical parametric methods is that it is not suitable for high-dimensional data sets and for arbitrary data sets.

Identifying Density-Based Local Outliers [2] is proposed by M. Breunig et al. This paper proposes a LOF technique which assigns a degree to each object which helps to find density based outlier. It is necessary to improve the performance of the outlier which is the difficult task to achieve.

Anomaly Detection Survey [3] is proposed by V. Chandola et al .The contribution of this survey is the classification of the existing research into three distinct categories, based on the problem formulation which are as follows

1. Identifying anomalous sequences according to the normal database.
2. Identifying an anomalous subsequence for long sequence.
3. Identifying a pattern in a sequence which has the frequency of occurrence anomalous.

Adapting existing solutions to other related problems, such as online anomaly detection and handling multivariate sequences is an important direction for future research.

Angle-Based Outlier Detection in High-Dimensional Data [4] is proposed by H.-P. Kriegel et al in 2008. It focuses on processing a angle based outlier detection method and some variants assessing the variance in the angle between the difference vectors of point to the other point. It is not suitable for unsupervised data.

Conditional Anomaly Detection [5] is proposed by X. Song et al in the year 2007. This paper describes method called conditional anomaly detection for taking differences among attributes.It also proposes three expectation-maximization algorithms to learn the model used in conditional anomaly detection. This paper considered very simple type of domain knowledge to boost the accuracy of anomaly detection. Scalability during the construction of the model is not properly addressed.

Singular Value Decomposition for a Fast Intrusion Detection System [6] is proposed by S. Rawat et al in 2008. It proposes a Singular Value Decomposition as a processing step to reduce the dimensionality and the computational time of the data. It is necessary to analyze the decomposition of incidence matrix to gain some insight about the association among system calls under normal and abnormal execution. The study of this paper shows that by applying LSI technique, dimensionality can be reduced without loosing its performance. This method of reduction works better in case of 'per-application' data set. It is important to further analyze the decomposition of incidence matrix under normal and abnormal execution to better understand the process profiling.

Distance-Based Detection and Prediction of Outliers [7] is proposed by F. Angiulli in 2006 .It proposes the state of distance-based outlier detection research, by giving the first proposal of distance based outlier prediction method. The method introduced is based

on the notion of outlier detection solving set, a subset of the data set that can be used to predict if new unseen objects are outliers. Dataset does not guarantee about the computational cost and accuracy.

Ranking Outliers Using Symmetric Neighborhood Relationship [8] is proposed by W. Jin in 2006. It proposes a simple but effective measure called INFLOW on local outliers based on a symmetric neighborhood relationship. It considers both neighbors and reverse neighbors of an object when estimating its density distribution. As a result, outliers discovered are more meaningful. LOCI [17] address the difficulty of choosing values for MinPts in the LOF technique by using statistical values derived from the data itself.

Robust Outlier Detection Using and Eigenspace Embedding [9] is proposed by N.L.D. Khoa and S. Chawla in 2010. This paper presents the network intrusion detection problem and propose the use of commute distance, a random walk metric, to discover anomalies in network traffic data .CDOF is less sensitive to perturbations than other measures.

local Incremental Outlier Detection for Data Streams [10] is proposed by D. Pokrajac, A. Lazarevic, and L. Latecki Proc.The proposed incremental LOF algorithm provides equivalent detection performance as the iterated static LOF algorithm (applied after insertion of each data record), while requiring significantly less computational time. Its main drawback is its computational time.

Online Anomaly Detection using KDE [11] is proposed by T. Ahmed in 2009.This paper presents an algorithm based on Kernel Density Estimates. The algorithm adaptively learns the definition of normality in the given application, assumes no prior knowledge regarding the underlying distributions, and then detects anomalies for false alarms. Future work will revolve around investigating other means of bootstrapping KEAD to achieve a user-specified error tolerance limit.

Projection Subspace Tracking [12] is proposed by B. Yang in jan 1995. The Projection Approximation Subspace Tracking with Deflation (PASTD) algorithm originally developed for subspace tracking.

Streaming Pattern Discovery in Multiple Time-Series [13] is proposed by S. Papadimitriou et al in 2005. This introduces SPIRIT (Streaming Pattern discovery in multiple Time series).

An Adaptive Filter Theory for outlier detection [14] is proposed by S. Haykin in 1991.This paper considers the signal extraction problem for time-discrete data when only a priori assumptions regarding the distributions of signal and noise are possible.

The Use of the Area under the Roc Curve in the Evaluation of Machine Learning Algorithms [15] is proposed by A.P. Bradley in 1997. This paper described the use of AUC to assess the quality of rules extracted from a SVM. It shows that ROC curves and AUC provide a more reliable measure for assessing the quality of the extracted rules than the commonly used measures of accuracy and fidelity.

## III. IMPLEMENTATION DETAILS

1. Proposed System

Traditional outlier detection or anomaly detection algorithm was not able to work with large amount of data due to memory and computational complexity, when we work with large size of input dataset, LOO technique will not significantly affect the resulting principal direction of the data. Therefore, we extend traditional PCA to the online oversampling strategy and present an online oversampling PCA (oosPCA) algorithm for largescale anomaly detection problems.

The proposed oosPCA scheme will replicate the target instance many times, and the idea is to expand the effect of anomaly other than normal data. Due to the introduction of online over sampling, the computation and memory efficiency will not be compromised and we can achieve noteworthy results as compared with the over sampling PCA. Instead of calculating single eigenvectors for single outlier, we are going to used the over sampling technique to maximize the anomalies so that we can calculate the principal direction as an average of over sampled data. In our method we use online services to calculate eigenvectors value. Therefore the memory and computational complexity will not be limited to execution end rather it will be dependent on web service performance.

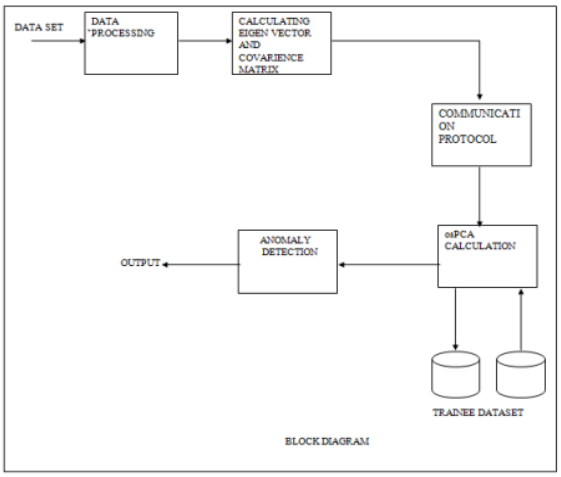


**Figure1: Block diagram showing the overall control flow of outlier detection.**

Dataset will be given as the input to the Data

Processing unit. In this unit the data is filtered and all the redundant data will be removed. Eigen values are stored online by using the soap communication protocol.Comparison will be performed on the oversampled data and the trainee dataset. If some difference is found in the comparison then it will be treated as the anomalous data.

## IV. METHODOLOGY

### *1. PCA(Principal Component Analysis).*

PCA is a method of dimension reduction which identifies the principal directions of the data distribution. To obtain these principal directions, it is important to construct the data covariance matrix and calculate its dominant eigenvectors. These eigenvectors provides more information among the vectors in the data space therefore they are considering as the principal directions.

### *2. Oversampling Principal Components Analysis (osPCA).*

osPCA.is the method in which it is not necessary to store the entire covariance matrix, therefore it is used in online or large-scale problems. By oversampling and extracting the target instant, the osPCA allows to determine the anomaly of the target instance according to the variation of the dominant eigenvector.

In our oosPCA framework, we will duplicate the target instance multiple times over web services and we will compute the score of anomalies of that output instance. If this score or the obtained result is above predefined threshold, we will consider this instance as an outlier. oosPCA not only determines outliers from the existing data, it can be applied to anomaly detection problems with streaming data or those with online requirements.

We can calculate PCA [16] by using following formula.

$$\max_{U \in IR^{p*k}, \|U\|=1} \sum_{i=1}^{n} \mathrm{U}^T (X_i - \mu)\ (\mathrm{X}_i - \mu)^T U \quad (1)$$

where U is a matrix consisting of k dominant eigenvectors. From this formulation, we can found that the standard PCA can be viewed as a task of determining a subspace.

The projected data has the largest variation. We can minimize the data reconstruction error using following formula.

$$\min_{U \varepsilon IR^{p*k}, \|U\|=1} J(\mathrm{U}) = \sum_{i=1}^{n} \| (\mathrm{x}_i - \mu) - UU^T (\mathrm{x}_i - \mu) \|^2$$

(2)

Where $\sum_A$ determines the optimal coefficients to weight each principal directions when reconstructing the approximated version of $(X_i - \mu)$. Generally, the problem in either (1) or (2) can be solved by deriving an eigenvalue decomposition problem of the covariance data matrix, i.e.

$$\sum_A U = U \wedge \quad (3)$$

Where

$$\sum\nolimits_A = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{x}_i - \mu)(\mathrm{x}_i - \mu)^T \quad (4)$$

is the covariance matrix, $\mu$ is the global mean. Each column of U represents an eigenvector o $\sum_A$ , and the corresponding diagonal entry in ^ is the associated eigenvalues. For the purpose of dimension reduction, the last few eigenvectors will be discarded due to their negligible contribution to the data distribution.

## V.MODULES

1. Cleaning Data
2. Detecting Outliers
3. Clustering

**Module 1 Cleaning Data:**

In this module the data is cleaned by removing the redundancies.

For this purpose the target instance will be duplicated multiple times this will amplify the effect of outlier rather than that of normal data. After this the Leave One Out (LOO) strategy is applied which identifies the angle difference. If we add or remove one data instance, the angle direction will be changed. The angle difference will be smaller for the normal data but it might be larger for the anomalous data.

A set of data instances from the original data set is taken as predefined input. This data may be contaminated by noise or it may be the incorrect data or This data might be error free as it is treated as the training data. Here the cleaning part is done using Over-Sampling Principal Component Analysis (osPCA) method. And then the score of outliernessst is calculated for each data instances. The smallest value st is taken as the threshold value.

**Module 2 Detection:**

This module is used for detecting the outlierness of the user input. When the user gives the input to the system, the system calculates the St value for the new input. And then compare that new St value with the threshold value calculated earlier. If the St value of the new data is above the threshold then that input data is identified as an outlier and that value will be discarded by the system. Otherwise it will be considered as a normal data instance and the PCA value of that data instance is updated.

**Module 3 Clustering :**

The training data will consider as the assumption So there are the chances that some outlier data may be considered as normal data. To slove 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.

**VI.CONCLUSION**

The advanced PCA is the efficient technique as it reduces the memory requirement and the computational cost. This technique can be able to handle the high dimensional data without the large iterations. Communication protocol helps in reducing the computational cost. Due to the online storage the memory requirement is also reduced.

**IX. FUTURE SCOPE:**

The future work can be carried on the normal data with the multiclustering structure. It is also required to focus on the extremely high dimensional data so as to solve the problem of curse of dimensionality.

**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. 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.
5. 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.
6. 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.
17. 
7. F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," .Knowledge and Data Eng.vol.18,no.2,pp145-160,2006.
8. W. Jin, A.K.H. Tung, J. Han, and W. Wang, "Ranking Outliers Using Symmetric Neighborhood Relationship," Proc. Pacific-AsiaConf.KnowledgeDiscoveryandDataMining,2006.
9. N.L.D. Khoa and S. Chawla, "Robust Outlier Detection Using Commute Time and Eigenspace Embedding," Proc. Pacific-AsiaConf.KnowledgeDiscoveryandDataMining,2010.
10. D. Pokrajac, A. Lazarevic, and L. Latecki, "Incremental Local Outlier Detection for Data Streams," Proc. IEEE Symp. Computational Intelligence and Data Mining, 2007.
11. T. Ahmed, "Online Anomaly Detection using KDE," Proc. IEEE Conf. Global Telecomm., 2009.
12. B. Yang, "Projection Approximation Subspace Tracking," IEEE Trans. Signal Processing, vol. 43, no. 1, pp.95-107, Jan.1995.
13. S. Papadimitriou, J. Sun, and C. Faloutsos, "Streaming Pattern Discovery in Multiple Time-Series," Proc. 31st Int'l Conf. Very Large Data Bases, 2005.
14. S. Haykin, Adaptive Filter Theory. Prentice Hall, 1991.
15. A.P. Bradley, "The Use of the Area under the Roc Curve in the Evaluation of Machine LearningAlgorithms,"PatternRecognition,vol.30,pp.114
16. Yuh-JyeLee,Yeh-RenYeh and Yu Chiang Frank Wang "Anomaly Detection via Online Oversampling Principal Component Analysis"vol 25,no 7 july 2013.