International Journal of scientific research and management (IJSRM)
||Volume||2||Issue||4||Pages|| 763-767 ||2013||
Website: www.ijsrm.in ISSN (e): 2321-3418

DETECTING OUTLIERS USING osPCA

J.Gnana Sekaran[1], P.Saranya[2]

[1]PG Student,
Affiliated to Anna University Chennai, Dept. of Computer Science and Engineering
Gnanamani College of Engineering,
Namakkal, India.
Gsekar32@gmail.com

[2]Asst. Prof Dept. of Computer Science and Engineering,
Gnanamani College of Engineering,
Namakkal, India.
info@gce.org.in

Abstract:*Anomaly detection has been an important research topic in data mining and machine learning. The most anomaly detection methods are typically implemented in batch mode so that it cannot be easily extended to large-scale problems without sacrificing computation and memory specification. An online oversampling principal component analysis (osPCA) algorithm is implemented to address the problem, then aiming at detecting the presence of outliers from a large amount of data via an online updating technique and at the same time it sends notification message to user's mail id. The proposed framework is favored for online applications which have computation or memory restrictions. By checking with the well-known power method for PCA and other popular anomaly detection algorithms, the results verify the feasibility of our proposed method in terms of both accuracy and efficiency.*

Keywords: Anomaly detection, virtualization, multitier web application**.**

1. Introduction

ANOMALY (or outlier) detection aims to identify a small group of instances which deviate remarkably from the existing data. A well-known definition of "outlier" is given, "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism," which gives the general idea of an outlier and motivates many anomaly detection methods. Practically, 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. However, since only a limited amount of labeled data are available in the above real world applications, how to determine anomaly of unseen data (or events) draws attention from the researchers in data mining and machine learning communities.

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.

2. Related Works

Existing approaches can be divided into three categories: distribution (statistical), distance and density-based methods.

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

For distance-based methods 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.

The density based method to find the local data density. The estimation of local data density for each instance is very computationally expensive, especially when the size of the data set is large.

A 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.

3. OVERSAMPLING PCA FOR ANOMALY DETECTION

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 now discuss. While this power method alleviates the computation cost in determining the principal direction as verified in our previous work in, we will discuss its limitations and explain why the use of power method is not practical in online settings. The presented a least squares approximation of our osPCA, followed by the proposed online updating algorithm which is able to solve the online osPCA efficiently.

3.1 Oversampling Principal Components Analysis(osPCA)

As mentioned earlier, when the size of the data set is large, adding (or removing) a single outlier instance will not significantly affect the resulting principal direction of the data. Therefore, we advance the oversampling strategy and present an oversampling PCA (osPCA) algorithm for largescale anomaly detection problems. 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. 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 directioninanonlinefashion, instead of calculating multiple eigenvectors carefully.

We now give the detailed formulation of the osPCA. Suppose that we oversample the target instance n times, the associated PCA can be formulated as follows

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

In this osPCA framework, we will duplicate the target instance n times (e.g., 10 percent of the size of the original data set), and we will compute the score of outlierness s of that target instance, as defined in. If this score is above some predetermined threshold, we will consider this instance as an outlier. With this oversampling strategy, if the target instance is a normal data , we will observe negligible changes in the principal directions and the mean of the data. It is worth noting that the use of osPCA not only determines outliers from the existing data, it can be applied to anomaly detection problems with streaming data or those with online requirements, as we discuss later. Clearly, the major concern is the computation cost of calculating or updating the principal directions in largescale problems. We will discuss this issue and propose our solutions in the following sections.

3.2 Effects of the Oversampling Ratio on osPCA

Using the proposed osPCA for anomaly detection, the oversampling ratio r in (11) will be the parameter for the user to be determined. We note that, since there is no training or validation data for practical anomaly detection problems, one cannot perform cross-validation or similar strategies to determine this parameter in advance. When applying our osPCA to detect the presence of outliers, calculating the principal direction of the updated data matrix (with oversampled data introduced) can be considered as the task of eigenvalue decomposition of the perturbed covariance matrix. Theoretically, the degree of perturbation is dependent on the oversampling ratio r, and t the sensitivity of deriving the associated dominant eigenvector.

3.3 The Power Method for 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. Let Q=$\frac{AA^T}{n}$ be the outer product matrix and $x_t$be the target instance (to be oversampled), we use the following technique to update the mean μ and the covariance matrix $\sum_A$

$$\tilde{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu} + r \cdot \mathbf{x}_t}{1 + r}$$

and

$$\Sigma_{\tilde{\mathbf{A}}} = \frac{1}{1+r}\mathbf{Q} + \frac{r}{1+r}\mathbf{x}_t\mathbf{x}_t^\top - \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top$$

where r<1 is the parameter controlling the size when oversampling x . It is the one only needs to keep the matrix Q when calculating need to re-compute the entire covariance matrix in this LOO framework.To alleviate this computation load, we apply the wellknown power method, which is a simple iterative algorithm and does not compute matrix decomposition. This method starts with an initial normalized vector u , which could be an approximation of the dominant eigenvector or a nonzero random vector. Next, the new u (a better approximated version of the dominant eigenvector) is updated by converges under the assumption that the dominant eigenvalue of is markedly larger than others. It is clear that the power method only requires matrix multiplications, not decompositions; therefore, the use of the power method can alleviate the computation cost in calculating the dominant principal direction.

A security analyst is interested in detecting "illegal" user sessions on a computer belonging to a corporate network. An illegal user session is caused when an unauthorized person uses the computer with malicious intent. To detect such intrusions, we present an online updating technique for our osPCA. This updating technique allows us to effectively detect the anomaly and at the same time it holds the account misused by attackers and then sends notification message to user's mail id. Compared to the 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.

3.2 Advantages

- The required computational costs and memory requirements are significantly reduced.

- Our method is especially preferable in online, streaming data, or large scale problems.

4.Authenticated User Module

Authorization is the process of giving user permission to do or have something. In multi-user computer systems, a system administrator defines for the system which users are allowed access to the system and what privileges of use (such as access to which file directories, time access, maintain history, and so forth). Assuming that someone has logged in to a computer operating system or application, the system or application may want to identify what resources the user can be given during this session. Thus, authorization is sometimes seen as both the preliminary setting up of permissions by a system administrator and the actual checking of the permission values that have been set up when a user is getting access. Authentication is the process of determining whether someone or something is, in fact, who or what it is declared to be. In private and public computer networks (including the Internet), authentication is commonly done through the use of logon passwords. Knowledge of the password is assumed to guarantee that the user is authentic. Each user registers initially (or is registered by someone else), using an assigned or self-declared password. On each subsequent use, the user must know and use the previously declared password.

4.1 Online Anomaly Detection

Online anomaly detection has the advantage that it can allow analysts to undertake preventive or corrective measures as soon as the anomaly is manifested in the sequence data. A technique that detects anomalous events within a sequence might not be directly applicable to detecting anomalies that are caused by a subsequence of events occurring together. Anomaly detection is a proven approach to defending against the array of threats facing online banking. This anomaly detection has been so successful at stopping online fraud.

4.2 Personal Identification Number

A personal identification number is a secret numeric password shared between a user and a system that can be used to authenticate the user to the system. This PIN number will be send to user's mail id based on user's account number. Typically, the user is required to provide a non-confidential user identifier or token (the user ID) and a confidential PIN to gain access to the system. Upon receiving the user ID and PIN, the system looks up the PIN based upon the user ID and compares the looked-up PIN with the received PIN. The user is granted access only when the number entered matches with the number stored in the system.

4.3 Money Transfer

Money transfer refers to any system, mechanism, or network of people that receives money for the purpose of making the funds or an equivalent value payable to a third party in another geographic location, whether or not in the same form.

An Email Money Transfer resembles an e-check in many respects. The money is not actually transferred by e-mail. Only the instructions to retrieve the funds are

- The sender opens an online banking session and chooses the recipient, the amount to send, as well as a security question and answer. The funds are debited instantly, usually for a surcharge.

- An e-mail is then sent to the recipient, with instructions on how to retrieve the funds and answer the question, via a secure website.

4.4 Withdrawal

Savings accounts of deposit are bank accounts that pay interest. Account holders can withdraw funds from savings accounts by making withdrawals at the teller, online or at automated teller machines.

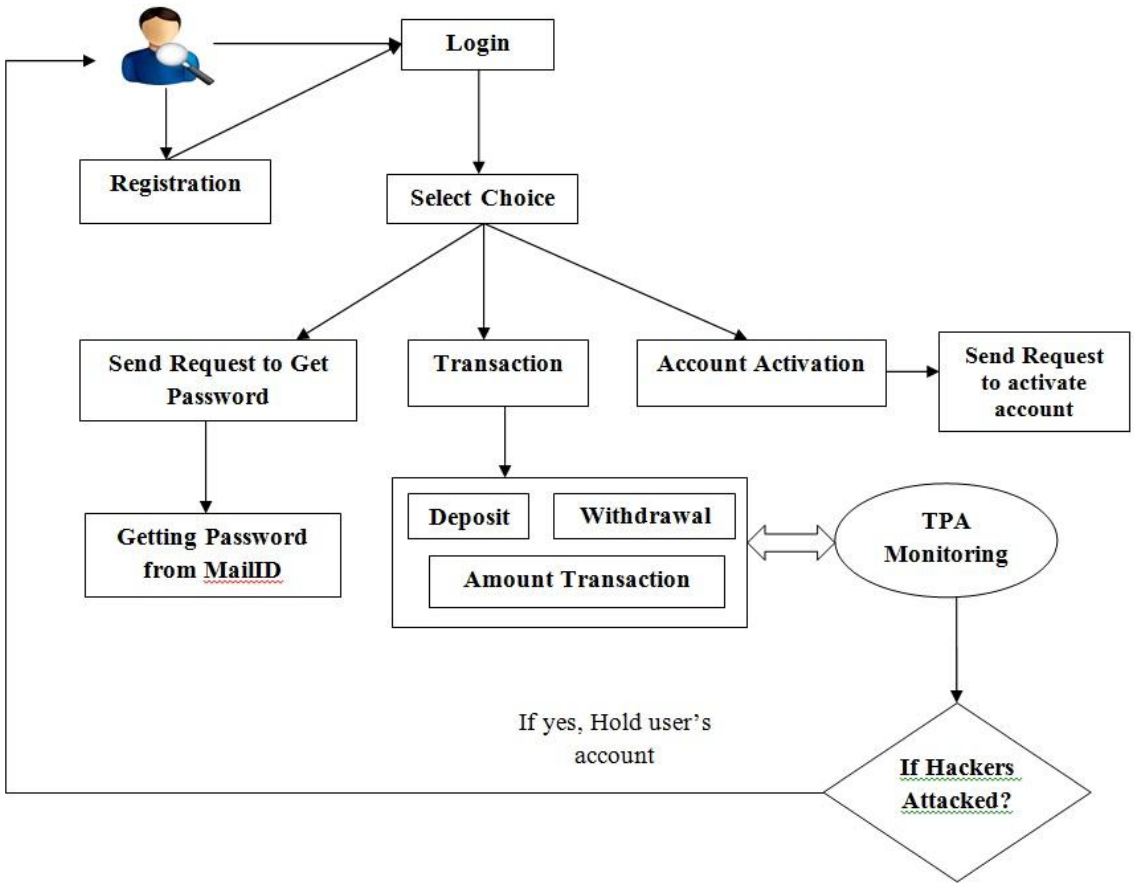

Fig 1: Authenticated User Module

This module is used to retrieve amount from user's account. The withdrawal of cash from the bank accounts by the customer. To withdraw, click on 'withdrawal' tab at the top of the page after login, and then click on 'withdrawal request'.

4.5 Intrusion Detection Module

A security analyst is interested in determining if the frequency with which a user executed a particular sequence of commands is higher (or lower) than an expected frequency. The sequence login, password, login, password corresponds to a failed login attempt followed by a successful login attempt.

Occurrence of this sequence in a user's daily profile is normal if it occurs occasionally, but is anomalous if it occurs very frequently, since it could correspond to an unauthorized user surreptitiously attempting an entry into the user's computer by trying multiple passwords. To detect such intrusions, the analyst can use the third formulation, in which the sequence of commands is the query pattern, and the frequency of the query pattern in the user sequence for the given day is compared against the expected frequency of the query pattern in the daily sequences for the user in the past, to detect anomalous behavior.

4.6 Auditing

The general definition of an auditing is an evaluation of a person, organization, system, process, enterprise, project or product. The term most commonly refers to audits in monitoring the process of organization. This is one method using that we can audit the data without giving the data. Whenever the hackers enters, it stores the detail of its hackers and informs to admin

4.7 Automated Response Module

Response to anomalies is automated or performed by staff. Proactive response stops criminals in their tracks AND builds trust with account holders. The staffs immediately hold their particular persons account and stop payments. They give alert notification immediately through mail or mobile that some intrusion is going to happen. Detects the widest range of

malware and non-malware fraud attacks. Automatically monitors all clients on all devices. Monitors every online and mobile banking session for fraudulent login, reconnaissance, fraud setup, and anomalous transactions.

5. Conclusion

In this paper, proposed online oversampling principal component analysis (osPCA) algorithm to detect intrusions, we present an online updating technique for our osPCA. This updating technique allows us to effectively detect the anomaly and at the same time it holds the account misused by attackers and then sends notification message to user's mail id. Compared to the other popular anomaly detection algorithms, the required computational costs and memory requirements are significantly reduced. Future research will be directed to the following anomaly detection scenarios: normal data with multiclustering structure, and data in an 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. Moreover, many learning algorithms encounter the "curse of dimensionality" problem in an extremely high-dimensional space.

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] 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.

[10] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, 2006.

Author Profile

J.Gnana Sekaran received the B.E degree in Computer Science and Engineering from Surya Engineering College, Perundurai in 2012. During 2012-2014, he stayed in Master Degree (Computer Science and Engineering) at Gnanamani College of Engineering, Namakkal.

P.Saranya received the M.E degree in Computer Science and Engineering from Paavai College of Technology, Affiliated to Anna University, Chennai. Received B.E degree in Computer Science and Engineering from Kongu College of Engineering, Affiliated to Anna University, Chennai in 2008. Now working as Assistant Professor in Gnanamani College of Engineering, Affiliated to Anna University, Chennai Since June 2012. Her research interest includes Data Mining.