ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(32), DOI: 10.17485/ijst/2015/v8i32/87511, November 2015* ISSN (Online) : 0974-5645

Differentially Sampled Anomaly Detection System based on Outlier Identification Technique

N. Javed[1] and Kannan Subramanian[2*]

[1]Master of Computer Application, Jerusalem College of Engineering, Chennai – 600100, Tamil Nadu, India; javednjaved@gmail.com

[2]Master of Computer Application, Bharath University, Chennai – 600073, Tamil Nadu, India; kannan.mca@bharathuniv.ac.in

Abstract

Anomaly detection is the process of identifying unusual behavior and also a small group of instances that deviate remarkably from the existing data. The real world application of anomaly detection includes intrusion or credit card fraud detection that requires a most efficient framework for identifying the deviated data instances. The technique called Principal Component Analysis (PCA) which require large amount of computation memory requirements and therefore it is not suitable for large scale data like online applications. Therefore a new technique called online Oversampling Principal Component Analysis (osPCA) algorithm along with online updating technique is used for detecting the existence of outliers from a large number of data. When oversampling a data instance the online updating technique enables the osPCA to update the outlier identification effectively without solving the eigenvalue decomposition. The feasibility of osPCA provides more efficient and accurate results. The work extends by detecting outliers from high dimensional dataset using some clustering techniques with lesser time consumption.

Keywords: Anomaly Detection, Online Updating, Oversampling Principal Component Analysis

1. Introduction

Anomaly detection also called as outlier detection is a process of searching of an item that does not conform to an expected pattern. These patterns that are detected are generally called as anomalies. These kinds of anomalies are translated to critical and actionable information in several application domains. Practically anomalies are patterns in data that do not match a well defined notion of normal behavior. The major task in data mining is the difficulty of detecting the outliers from a large set of data objects and this motivates multiple number of anomaly detection techniques[1,2]. Anomaly detection finds its use in wide variety of applications that lead to illegal or abnormal behavior, such as credit card fraud detection, homeland security and network intrusion detection and insurance fraud, medical diagnosis, marketing or customer segmentation, intrusion detection for cyber-security[4]. These real world applications contain only a limited amount of labeled data, how to detect anomaly of unseen events draws attention from the researchers in machine learning and data mining areas.

The existences of anomalies are often determined from the amount of deviated data. The presence of such kind of deviated data immensely affects the distribution of data. The calculation of least squares and data mean associated with the linear regression model is both sensitive to outliers. In that event, anomaly detection needs to solve an unsupervised yet unbalanced data learning problem. The most frequently cited sentence for the meaning of outliers is "one person's noise is another person's signal". Many approaches are framed for detecting outliers from a large scale of data as well as high dimensional data yet most of the known methods are least theoretically applicable to high dimensional data.

* *Author for correspondence*

A well known framework called Principal Component Analysis (PCA) along with LOO (Leave One Out) strategy calculates effect of outlierness from the derived principal direction of the data set without the presence of target instance and that of the original data set. However in the real-world anomaly detection problems dealing with large amount of data, adding or removing a target data instance produces only a negligible difference in the resulting eigenvectors and it is not elementary to apply the (Decremental PCA (DPCA) technique for detecting the anomalies[3,5]. The LOO anomaly detection procedure with an oversampling strategy will strikingly increase the computational load as well as the storage capacity. Henceforth, the problem of detecting outliers in machine learning, data mining literature lose their algorithmic effectiveness for high dimensional data. Even though a well known power method is able to produce approximated PCA solutions; it cannot be extended easily to applications with streaming of data. For this reason an online updating technique along with osPCA is framed, which allows to efficiently calculate the approximated dominant eigen-vector without storing the entire covariance matrix. In order to increase the speed of the system a clustering technique called hierarchical clustering is used in accordance with osPCA online updating technique. Correlated to other anomaly detection methods like ABOD, PCA with LOO strategy, osPCA and power methods the proposed framework incomparably reduces the imperative computational costs and memory requirements. And thence the proposed method is eminently desirable in online, streaming data or large-scale problems.

The clustering technique along with the online updating osPCA is computationally preferable to precedent anomaly detection methodologies.

2. Principle Component Analysis

Previously many number of outlier detection techniques have been proposed. One among those is the well known PCA method. This PCA is a renowned unsupervised dimension reduction method to determine

The principal directions of the data distribution. In LOO strategy the principle direction of the data set can be calculated without the presence of target instance and that of the original data set by adding or removing the abnormal (or normal) data instance. It is examined that removing (or adding) an abnormal data will produce a greater difference in the principal direction when compared to same operation performed with the normal data[3,5]. In such a way, the outlierness can be discovered by the variation of the resulting principal directions.

Though PCA is a well known dimension reduction method, it is not suitable for real-world an anomaly detection problem which deals with a large amount of a data. Therefore, adding or removing a target instance will create only a negligible difference in the resulting eigenvectors. The PCA method retains characteristics of the data set which contributes most of its variance by handling with the lower-order principal components. This method is sensitive to outliers and the main data structure is represented with the help of few principal components[3]. In these techniques, the presence of an outlier data instance is determined by many processes, but one of those processes is to determine the corresponding principal direction of the data set. Thus, whenever an outlier data instance is introduced, accordingly the principal direction will create corresponding angle changes. That is, adding or removing an outlier data or a deviated data instance will cause larger effect on these principal directions.

$$S_t = 1 - \left| \frac{\left(\bar{u}_t, u\right)}{\left\|\bar{u}_t\right\| \|u\|} \right|$$

Consequently, the variation of principle directions while removing or adding an instance is discovered. Accordingly, the first principal direction is greatly affected whenever an outlier or a deviated data instance is added or removed from the existing data set. In such a case the first principal direction is affected badly. Therefore, the initial principal direction is varied and that forms a large angle between itself and the old one. Therefore, in such kind of situations the initial principal direction will not get affected and it only form an extremely smaller angle between the old initial principal direction and the new one if the normal data instance is removed from its original data set. By means of this observation the LOO procedure is used in each and every individual point by adding or removing (with or without) effect. The same procedure is used with the LOO technique but with the incremental strategy. This methodology enables adding a data instance to determine the possible variations of the principal directions from the above mentioned strategies it was found that the principal directions are significantly affected with the removal of an outlier data instance

while this variation of the principal direction will be much smaller with the removal of a normal data instance. Conversely adding an outlier data instance will also cause significantly larger influence on the principal directions while the variation of the principal direction will be notably smaller with adding a normal data instance. However, it is not applicable for detecting the presence of outliers from a large scale data, since in case of a large scale data or streaming data removing a outlier will not create such a difference in the resulting principal direction. Equation 1 shows the method of finding the score of outlierness, where u_t is the eigen vector value for the target instance.

To address the above problem another technique called oversampling PCA (osPCA) is proposed[3]. This is because, the effect of "with or without" a specified data instance might be diminished when the size of the data is large. Thence to overcome the problem, this oversampling strategy employs duplicating the target instance on such an oversampled data and this makes detecting of outliers to be much easier.

In the PCA scheme for anomaly detection, n PCA analysis for a data set with n data instances in a p-dimensional space has to be performed, despite it is computationally infeasible to compute. In the practical anomaly detection problems like real world anomaly detection problems this may not be so easy to discover and monitor the variation of the principal directions caused by the presence of a single outlier. Since in a real-world data set, the size of the data is typically large and thence detecting the presence of outlier from such a large amount of data is really a tedious task.

The power method for osPCA is a simple iterative algorithm which does not need to compute matrix decomposition. This method requires only the matrix multiplications and it does not bother about decomposition of matrix that is formed. Due to these reasons the computation cost can be mitigated in calculating the computational cost. But this method is computationally expensive and also it

cannot be applied for large-scale anomaly detection techniques. In order to reduce the computational cost and the memory requirements of the anomaly detection system a new methodology called osPCA with online updating technique was proposed. Though the cost of computation is not so expensive as well it does not require large amount of memory the time taken to train the entire set of real world data is tremendous. This training delays the other activities of the system and therefore a new framework has to be proposed in order to reduce the system response time and the performance time. However the computational complexity and the memory requirements for the online osPCA method is very low when compared to other methods including osPC power method.

It is undesirable that the above described methods are typically implemented in batch mode. Therefore it is not so elementary to perform the anomaly detection problems with online data (large-scale data) or streaming data.

3. Proposed Work

3.1 Training Dataset

The data training set is a set of data used in different areas of information science. The main objective of training the data set is to conceive a potentially predicated relationship. While considering a real time KDD (Knowledge D Discovery in Databases) which consists of a large amount of data instances. This data training methodology includes training the undefined and unstructured data set in such a way that these data sets are arranged accordingly in the database. The KDD data set or any real time can be used for this purpose. The most dominant application which requires detection in online updating strategies is credit card fraud detection, fault detection, detection in cyber security. Many of the anomaly detection techniques need a set of completely pure normal data set to train the model. However, these kind of anomaly detection techniques implicitly assumes that these anomalies can also be treated as patterns that are not observed before.

By this reason, an outlier may be defined as the data point which is highly unique from the rest of the data instances, based on some predefined measure. And therefore, several outlier detection schemes are proposed in order to verify how efficiently the problem of anomaly detection can be pledged. The data from the majority class is considered as the normal data and randomly 1 percent of the data instances are selected from the minority classes as the samples.

3.2. Cleaning Data

The training data set can be selected based on the user assumption in case of handling a real-world data

instances. The main goal of data cleaning phase is to identify the suspicious outliers[2]. A set of data instances from the original data set after the data training process is taken as the predefined input. These data sets may be contaminated due to noise, incorrect data labeling etc., but they become an error free data due to training of data. Using LOO strategy the absolute value of cosine similarity is determined to measure the difference of the principle direction and obtain the suspicious outlier scores. When the suspicious outlier score is higher, it implies the higher probability of being an outlier data instance. The ranking for all the data instances can be made after calculating the suspicious outlier scores for each data instances. The over-sampling principal component analysis outlier detection algorithm is used for cleaning the data. The data cleaning process aims to filter out the most deviated data using osPCA. This process is done offline before performing the online anomaly detection process. The percentage of training normal data instance to be disregarded can be determined by the user[3]. While performing the data cleaning process the smallest score of outlierness is used which of the remaining training data instances as the threshold for outlier detection. The trained data extracted from the data cleaning process is completely recycled by the online detection process in order to detect each arriving target instance.

3.3 Data Clustering

The process of training the data is done based on the user assumption and hence there may be a problem that even an outlier data can be considered as a normal data instance. This is because of training the data based on the assumption made by the user.

Clustering method is the best solution for the above mentioned problem. The type of clustering used here is the hierarchical clustering. This kind of clustering technique creates an easily understandable hierarchical cluster model for the data instance[4]. Thereby it tends to increase the speed of the anomaly detecting system and also the amount of time consumption of the remaining processes tend to get reduced accordingly. The clusters are usually formed for the input data instances, where the outlier calculation is applied to each cluster for detecting the exact outlier data instance. The score of outlierness is updated accordingly from the process of data cleaning.

3.4 Online Detection

As soon as the data cleaning process is completed, the suspicious points are filtered. As a result a pure normal data is obtained to which the online anomaly technique is to be applied. The main objective of the online anomaly detection method is to identify the newly arriving abnormal data instance. Similarly the oversampling process is also applied for the newly arriving instance[5]. The idea behind this method is to determine the mean and standard scores that are computed from all normal data points. Once when the mean and the standard deviation is calculated, a newly arriving data instances will be marked as an outlier if it's suspicious score than the previously calculated values.

For a KDD dataset there are four categories of attacks which are to be considered as an outlier data instance.

- DoS (Denial of Service), example, ping-of-death, teardrop, etc.
- U2R, it is the unauthorized access to local super user privileges by a local privileged user, example. and various buffer overflow attacks.
- PROBING, surveillance and probing, example, port-scan, ping –sweep, etc.
- R2L, unauthorized access from a remote machine, example, guessing password.

Figure 1 depicts the architecture of the proposed anomaly detection system. The input is the KDD data set which is obtained from the KDD website[6]. It consists of thousands of data sets including both the normal data instance and the outlier data instances. This KDD data set is considered as the anomaly data set and is processed based on the proposed methodology.

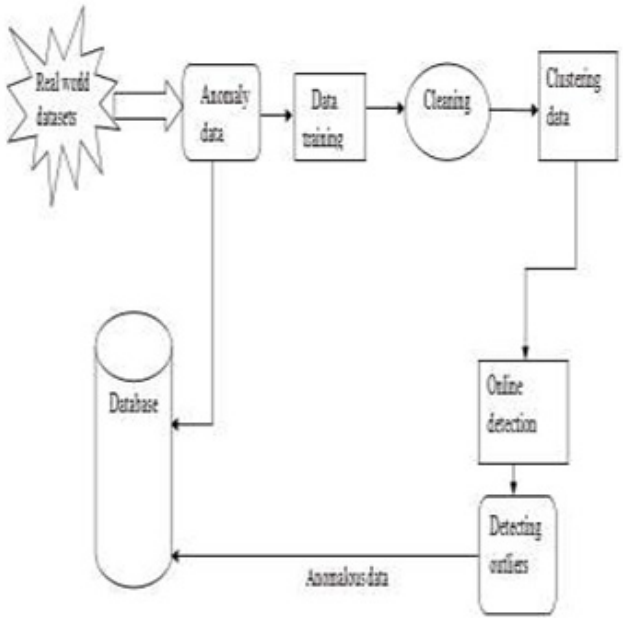

Figure 1. Proposed framework.

Any real world data sets excluding KDD data set can also be given as the input anomaly data. These data sets are trained in the data training process. After storing these data sets into the database it is cleaned in the data cleaning process. These processes include filtering of the most deviated or outlier data[7].

The cleaning block includes extracting the normal data from the contaminated data set. The data cleaning process calculates the score of outlierness of each receiving test input data instance. The data cleaning determines the threshold and report the correct and incorrect rates over the receiving data instances. The clustering of data instance includes priority wise or range wise clustering of data which makes searching and cleaning of data faster which in turn makes the implementation becomes much easier[8]. The result of clustering or the data cleaning is fed as input to the online detection process. It aims at detecting the newly arriving abnormal instance. As a result of the online anomaly detection process, the possible suspicious data instances, say abnormal data instances are detected and these sets of data are stored separately stored in the database and therefore whenever a new data set is taken for processing the suspicious data sets can be detected easily and ignored in order to obtain the system security.

Therefore any unauthorized access from a remote machine, unauthorized access to a local super user machine or any probing attacks can be easily detected. Even though these kinds of data instances resembles same as the normal data instances with a background noise that characterize as a normal data. The attacked data files can be listed out as a suspicious data instances by comparing its features with the some predefined list files[9]. Many of the connections are being used by the DOS and probing attacks and therefore the proposed framework must able to capture even those kinds of data that represents the same characteristics. However the connection feature of the data content was the important characteristics for detecting R2L (the unauthorized access from a remote machine) and U2R (the unauthorized access to a local super user privileges) attack types, while the time-based and the connection-based features were the most important features for detecting and determining the DoS (Denial-of-Service) attack types.

3.5 Outlier Detection

Since it is impossible for a human analyst to investigate practically all the outlier detection scenarios, the clustering technique along with the online osPCA is proposed which involves faster detection of outlier data instances and storing it in a separate field on the database[10]. So, whenever a real time data like credit card data has to be detected through online then the previously investigated results can be used in order to compare it with the newly arrived data and henceforth the suspicious data instances can be detected and blocked. These kinds of anomaly detection techniques can be most probably used in many application domains which consist of real world data sets. The online osPCA with clustering technique achieves slightly a larger false positive rate than the other anomaly detection methods. It also achieves a comparable performance significantly less computation time along with less memory requirement; this is because the proposed framework does not require storing the entire data matrix during the online detection process[11].

As soon as the anomalous data sets are detected they are separately stored in the database and whenever a newly arriving data instance that possess the same characteristics of the already detected anomalous data then the system is given an alert message or an acknowledgement that the newly arrived data is the suspicious data. This helps the system user to protect the original data sets from the attacker sets of data and also ignore these kinds of data automatically. Thus the proposed work involves obtaining much faster and more efficient results when compared to other methods.

4. Conclusion

An anomaly detection methodology for detecting intrusions and suspicious data sets are proposed in this paper. To support the applicability of anomaly detection schemes, several clustering techniques are used accordingly along with the online osPCA methodology by determining the outlier scores for each data instances. This outlier scores are framed in order to rank the outliers in an efficient manner. Moreover the proposed method does not require storing the entire data matrix during the online detection process. Therefore, when compared to other anomaly detection methodologies the proposed framework produces a better result. On the other hand, it can be used for detecting anomalies in large scale data including online data stream or any unbalanced data distribution (includes network security problems).

The future work can be extended in order to have a high detection rate and also to increase the speed of the anomaly detection system which automatically deletes the anomalous data from the original data set.

5. References

1. Breunig M, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. Proc ACM SIGMOD Int'l Conf Management of Data; 2000. p. 93–104.
2. Beula Devamalar PM, Thulasi Bai V, Srivatsa SK. Design and architecture of real time web-centric tele health diabetes diagnosis expert system. International Journal of Medical Engineering and Informatics. 2009; 1(3):307–17. ISSN: 1755-0661.
3. Kriegel HP, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data. Proc 14th ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining; 2008. p. 444–52.
4. Jebaraj S, Iniyan S. Renewable energy programmes in India. International Journal of Global Energy. 2006; 26(3-4):232–57. ISSN: 0954-7118.
5. Yeh YR, Lee ZY, Lee YJ. Anomaly detection via oversampling principal component analysis. Proc First KES Int'l Symp. Intelligent Decision Technologies; 2009. p. 449–58.
6. Wang W, Guan X, Zhang X. A novel intrusion detection method based on principal component analysis in computer security. Proc Int'l Symp Neural Networks; 2004. p. 657–62.
7. Sharmila S, Jeyanthi Rebecca L, Saduzzaman M. Biodegradation of domestic effluent using different solvent extracts of Murraya koenigii. Journal of Chemical and Pharmaceutical Research. 2013 Jan; 5(2):279–82. ISSN: 0975-7384.
8. Lee YJ, Yeh YR, Wang YF. Anomaly detection via online oversampling principal component analysis. IEEE Transactions on Knowledge and Data Engineering. 2013 Jul; 25(7):1460–70.
9. Jayalakshmi T, Krishnamoorthy P, Ramesh Kumar G, Sivamani P. Optimization of culture conditions for keratinase production in Streptomyces sp. JRS19 for chick feather wastes degradation. Journal of Chemical and Pharmaceutical Research. 2011; 3(4):498–503. ISSN: 0975-7384.
10. Pokrajac D, Lazarevic A, Latecki L. Incremental local outlier detection for data streams. Proc IEEE Symp Computational Intelligence and Data Mining; Honolulu HI. 2007. p. 504–15.
11. Gopalakrishnan K, Prem Jeya Kumar M, Sundeep Aanand J, Udayakumar R. Thermal properties of doped azopolyester and its application. Indian Journal of Science and Technology. 2013 Jun; 6(S6):4722–5. ISSN: 0974-6846.
12. Kimio T, Natarajan G, Hideki A, Taichi K, Nanao K. Higher involvement of subtelomere regions for chromosome rearrangements in leukemia and lymphoma and in irradiated leukemic cell line. Indian Journal of Science and Technology. 2012 Apr; 5(1):1801–11.
13. Cunningham CH. A laboratory guide in virology. 6th ed. Minnesota: Burgess Publication Company; 1973.
14. Sathishkumar E, Varatharajan M. Microbiology of Indian desert. In: Ecology and vegetation of Indian desert. In: Sen DN, editor. India: Agro Botanical Publ; 1990. p. 83–105.
15. Varatharajan M, Rao BS, Anjaria KB, Unny VKP, Thyagarajan S. Radiotoxicity of sulfur-35. Proceedings of 10th NSRP; India. 1993. p. 257–8.