ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Anomaly Detection on Data Streams with High Dimensional Data Environment

Mr. D. Gokul Prasath[1], Dr. R. Sivaraj, M.E, Ph.D., [2]

Department of CSE, Velalar College of Engineering & Technology, Erode[1]

Assistant professor, Department of CSE, Velalar College of Engineering & Technology, Erode[2]

Abstract: Classification consists of assigning a class label to a set of unclassified cases. Supervised and unsupervised classification methods are used to assign class labels. Classification is performed in two steps learning or training (model construction) and testing (model usage). Learning process is used to identify the class patterns from the labeled transactions. In training phase unlabeled transactions are assigned with the class values with reference to the learned class patterns. An outlier is an observation that deviates so much from other observations as to arouse suspicions. Distance based outlier detection methods are used to identify records that are different from the rest of the data set. The anomaly detection is referred as outlier detection process.

Batch mode based anomaly detection scheme is not suitable for large scale data values. Batch mode scheme requires high computational and memory resources. Principal component analysis (PCA) is a unsupervised dimension reduction method. PCA determines the principal directions of the data distribution. online oversampling principal component analysis (osPCA) algorithm is used to detect outliers from a large amount of data via online.

The over sampling based Principal Component Analysis (osPCA) method is enhanced to handle high dimensional data values. The learning process is improved to manage dimensionality differences. The system is tuned to handle data with multi cluster structure. The system is enhanced to perform anomaly detection on streaming data values.

I. INTRODUCTION

In data mining, anomaly detection (or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or finding errors in text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

In particular in the context of abuse and network intrusion detection, the interesting objects are often not *rare* objects, but unexpected *bursts* in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.

Three broad categories of anomaly detection techniques exist. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier. Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given *normal* training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

Anomaly detection is applicable in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Eco-system disturbances. It is often used in preprocessing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

II. RELATED WORK

In the past, many outlier detection methods have been proposed [4], [3], [5]. Typically, these existing approaches can be divided into three categories: distribution, 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 form 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. Nevertheless, the assumption or the prior knowledge of the data distribution is not easily determined for practical problems.

For distance-based methods [1], [6], 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. In such cases, this type of approach will result in determining improper neighbors, and thus outliers cannot be correctly identified.

To alleviate the aforementioned problem, density-based methods are proposed [3]. One of the representatives of this type of approach is to use a density-based local outlier factor (LOF) to measure the outlierness of each data instance. Based on the local density of each data instance, the LOF determines the degree of outlierness, which provides suspicious ranking scores for all samples. The most important property of the LOF is the ability to estimate local data structure via density estimation. This allows users to identify outliers which are sheltered under a global data structure. However, it is worth noting that the estimation of local data density for each instance is very computationally expensive, especially when the size of the data set is large.

Besides the above work, some anomaly detection approaches are recently proposed [4], [5]. Among them, the angle-based outlier detection (ABOD) method [4] 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 largescale problems, since the user will need to keep all data instances to calculate the required angle information.

It is worth noting that the above methods are typically implemented in batch mode, and thus they cannot be easily extended to anomaly detection problems with streaming data or online settings. While some online or incrementalbased anomaly detection methods have been recently proposed [7], [2], we found that their computational cost or memory requirements might not always satisfy online detection scenarios. For example, while the incremental LOF in [7] is able to update the LOFs when receiving a new target instance, this incremental method needs to maintain a preferred (or filtered) data subset. Thus, the memory requirement for the incremental LOF is O(np) [7], [2], where n and p are the size and dimensionality of the data subset of interest, respectively.

III. ANOMALY DETECTION PROCESS

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": "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, this LOO anomaly detection procedure with an oversampling strategy will markedly increase the computational

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

load. For each target instance, one always needs to create a dense covariance matrix and solves the associated PCA problem. This will prohibit the use of our proposed framework for real-world large-scale applications. Although the well known power method is able to produce approximated PCA solutions, it requires the storage of the covariance matrix and cannot be easily extended to applications with streaming data or online settings. Therefore, we present an online updating technique for our osPCA. This updating technique allows us to efficiently calculate the approximated dominant eigenvector without performing eigen analysis or storing the data covariance matrix. 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.

IV. TECHNIQUES FOR ANOMALY DETECTION

Principal Component Analysis

PCA is a well known unsupervised dimension reduction method, which determines the principal directions of the data distribution. To obtain these principal directions, one needs to construct the data covariance matrix and calculate its dominant eigenvectors. These eigenvectors will be the most informative among the vectors in the original data space and are thus considered as the principal directions. Let $A = [x_1^T, x_2^T, \ldots, x_n^T] \in IR^{n \times p}$, where each row x_i represents a data instance in a p dimensional space, and n is the number of the instances. Typically, PCA is formulated as the following optimization problem

$$\max_{U \in \mathbb{R}^{p \times k}, \|U\|=I} \sum_{i=1}^{n} U^T (x_i - \hat{\mu}) (x_i - \hat{\mu})^T U;$$

U is a matrix consisting of k dominant eigenvectors. From this formulation, one can see that the standard PCA can be viewed as a task of determining a subspace where the projected data has the largest variation.

Anomaly Detection Using PCA

In this section, we study the variation of principal directions when we remove or add a data instance, and how we utilize this property to determine the outlierness of the target data point. We note that the clustered blue circles represent normal data instances, the red square denotes an outlier, and the green arrow is the dominant principal direction. 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. Given a data set A with n data instances, we first extract the dominant principal direction u from it. If the target instance is x_t, we next compute the leading principal direction $\tilde{u}_t$ without x_t present.

After deriving the principal direction $\tilde{u}_t$ of $\tilde{A}$, calculate the score s_t, and the outlierness of that target instance can be determined accordingly. This strategy is also preferable for online anomaly detection applications, in which we need to determine whether a newly received data instance is an outlier.

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.

We introduce our osPCA and discuss how and why we are able to detect the presence of abnormal data instances according to the associated principal directions, even when the size of data is large. The well-known power method is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

applied to determine the principal direction without the need to solve each eigenvalue decomposition problem. While this power method alleviates the computation cost in determining the principal direction as verified in our previous work in [9], we will discuss its limitations and explain why the use of power method is not practical in online settings. We present a least squares approximation of our osPCA, followed by the proposed online updating algorithm which is able to solve the online osPCA efficiently.

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.

Issues on osPCA based Anomaly Detection

Batch mode based anomaly detection scheme is not suitable for large scale data values. Batch mode scheme requires high computational and memory resources. Principal component analysis (PCA) is an unsupervised dimension reduction method. PCA determines the principal directions of the data distribution. Anomaly detection on high dimensional data is not supported

- Multi cluster structure data model is not handled
- Data distribution estimation is not optimized
- Detection latency is high

Cluster Based Anomaly Detection Scheme

Clustering is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

An outlier is an object that differs from most other objects significantly. Therefore it can be considered as an anomaly. For outlier detection, only the distance to the appropriate centroid of the normal cluster is calculated. If the distance between an object and the centroid is larger than a predefined threshold d_{max}, the object is treated as an outlier and anomaly. In contrast to the classification method, outlier detection does not make use of the anomalous cluster centroid, i.e. it may be less accurate in detecting known kinds of anomalies.

Anomaly Detection on High Dimensional Data Streams

The Principal Component Analysis (PCA) method is enhanced to handle high dimensional data values. The learning process is improved to manage dimensionality differences. The system is tuned to handle data with multi cluster structure. The system is enhanced to perform anomaly detection on streaming data values. The system is designed to analyze data collected from streams. Data values are collected from different sources. Census data values are analyzed by the system.

Data Receiver

Data receiver listens in TCP port to collect data from remote nodes. Received data values are updated in to the database. Missing data elements are assigned using aggregation functions. Different attribute types are used in the data collection process.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

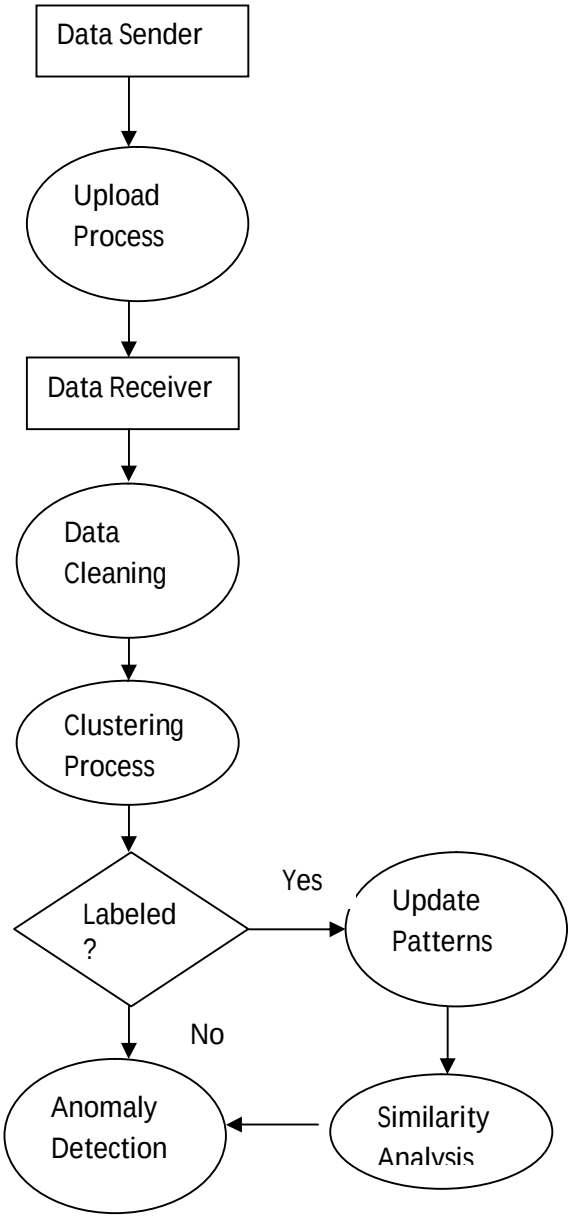
Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Data Sender

The data sender application is used to upload the client data values. Census data values are used in the system. TCP streams are used to transfer the data values. Data upload process is verified with acknowledgement messages.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Cluster Process

Clustering process is applied to partition the high dimensional data values. Clustering is applied on data streams. Data values are partitioned into multi cluster structures. Similarity analysis mechanism is used for clustering process.

Learning Process

Learning process is performed to identify the class patterns. Labeled transactions are used in the learning process. Learning process is initiated on clustered data values. Multi cluster structure is adapted for learning process.

Anomaly Detection

The anomaly detection is applied on streaming data values. Learned patterns are used in the anomaly detection process. Dimensionality variations are considered in the anomaly detection process. Patterns are compared using similarity measures.

V. CONCLUSION

Anomaly detection methods are used to detect anomalyous data values in a data collection. Principal Component Analysis model is used to handle dimensionality reduction process. The system enhances over sampling based Principal Component Analysis (osPCA) model to support data with multi cluster structure. The system is enhanced to discover anomalies in high dimensional data values from data streams. High dimensional data classification model is supported by the system. The system performs the classification on data streams. False positive and false negative errors are reduced by the system. Classification accuracy is improved by the enhancement of over sampling based principal component analysis method.

REFERENCES

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

[2]Ahmed T., "Online Anomaly Detection using KDE," Proc. IEEE Conf. Global Telecomm., 2009.

[3] Jin W., Tung A. K. H., Han J., and Wang W., "Ranking Outliers Using Symmetric Neighborhood Relationship," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2006.

[4] Kriegel H. P., Schubert M., and Zimek A., "Angle-Based Outlier Detection in High-Dimensional Data," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and data Mining, 2008.

[5] Kriegel H.-P.,Kro¨ger P. , Schubert E., an Zimek A., "Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2009.

[6] Khoa N. L. D. and Chawla S., "Robust Outlier Detection Using Commute Time and Eigenspace Embedding," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2010.

[7] Rawat S., Pujari A. k., and Gulati V. P., "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.

[8] Pokrajac D., Lazarevic A., and L. Latecki, "Incremental Local Outlier Detection for Data Streams," Proc. IEEE Symp. Computational Intelligence and Data Mining, 2007.

[9]Yeh y.R., Lee , Z.-Y. and Y.-J. Lee, "Anomaly Detection via Oversampling Principal Component Analysis," Proc. First KES Int'l Symp. Intelligent Decision Technologies, pp. 449-458, 2009.