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. 3, Issue 8, August 2015

Anomaly Detection via Online Oversampling Principal Component Analysis

M.Saraswathi[1], N.Kowsalya[2]

Research Scholar, Dept. of CS, Vivekanandha College of Arts and Sciences for Women (Autonomous), Namakkal, India[1]

Associate Professor, Department of CS & Application, Vivekanandha College of Arts and Sciences for Women (Autonomous), Namakkal, India[2]

ABSTRACT: Anomaly detection has been an important research topic in data mining and machine learning. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. However, most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. In this paper, we propose an online oversampling principal component analysis (osPCA) algorithm to address this problem, and we aim at detecting the presence of outliers from a large amount of data via an online updating technique. Unlike prior principal component analysis (PCA)-based approaches, we do not store the entire data matrix or covariance matrix, and thus our approach is especially of interest in online or large-scale problems. 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. Since our osPCA need not perform eigen analysis explicitly, the proposed framework is favored for online applications which have computation or memory limitations. Compared with the well-known power method for PCA and other popular anomaly detection algorithms, our experimental results verify the feasibility of our proposed method in terms of both accuracy and efficiency.

KEYWORDS: Anomaly detection, online updating, least squares, oversampling, principal component analysis

I. 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 in : "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism," 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 . 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 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.. 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.

We note that the above 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.To address this practical problem, we advance the "oversampling" strategy to duplicate the target instance, and we perform an over-sampling

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. 3, Issue 8, August 2015

PCA (osPCA) on such an oversampled data set. one always needs to create a dense covariance matrix and solves the associated PCA problem. 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. This updating technique allows us to efficiently calculate the approximated dominant eigen-vector without performing eigen analysis or storing the data covariance matrix. and thus our method is especially preferable in online, streaming data, or large-scale problems.

II. RELATED WORK

In the past, many outlier detection methods have been proposed. Typically, these 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 predeter-mined distributions, and this type of approach aims to find the outliers which deviate form such distributions. How-ever, 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 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. To alleviate the aforementioned problem, density-based methods are proposed. 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.

Besides the above work, some anomaly detection approaches are recently proposed. Among them, the angle-based outlier detection (ABOD) method 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 k nearest neighbors.

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. We found that their computational cost or memory requirements might not always satisfy online detection scenarios. For example, while the incremental LOF in [17] is able to update the LOFs when receiving a new target instance, this incremental method needs to maintain a preferred (or filtered) data subset. but the proposed algorithm requires at In online settings or large-scale data problems, the aforementioned methods might not meet the online requirement, in which both computation complexity and memory requirement are as low as possible.

III. PROPOSED WORK

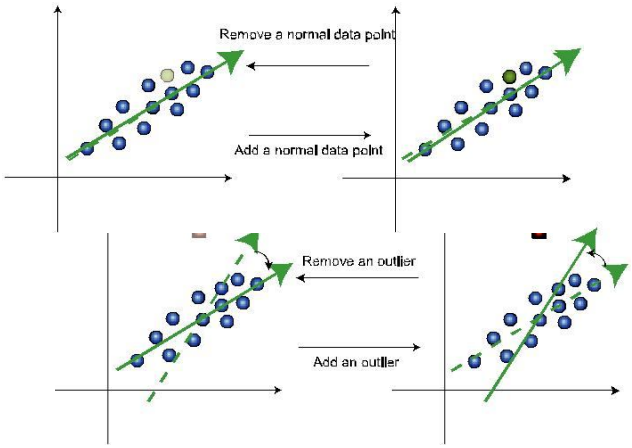
PCA is a well known unsupervised dimension reduction method, which determines the principal directions of the data distribution. To obtain these principal directions, one needs to construct the data covariance matrix and calculate its dominant eigenvectors. These eigenvectors will be the most informative among the vectors in the original data space, and are thus considered as the principal directions Typically, PCA is formulated as the following optimization problem

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. 3, Issue 8, August 2015



where 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.

Alternatively, one can approach the PCA problem as minimizing the data reconstruction error, i.e.

$$\min_{U2IR^{p\_k};kUk¼I} J(U) = \sum_{i=1}^{n} \|(x_i - \bar{x}) - UU^{>}(x_i - \bar{x})\|_2^2$$

where $U^{>}\delta x_i \_ \_ = X$

While PCA requires the calculation of global mean and data covariance matrix, we found that both of them are sensitive to the presence of outliers. present in the data, dominant eigenvectors produced by PCA will be remarkably affected by them, and thus this will produce a significant variation of the resulting principal directions.

Use of PCA for Anomaly Detection

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. Fig. 1. The effects of adding/removing an outlier or a normal data instance 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.

We now present the idea of combining PCA and the LOO strategy for anomaly detection. Given a data set A with n data instances, we first extract the dominant principal direction u from it. threshold, we then identify this instance as an outlier. We refer to this process as a decremental PCA with LOO scheme for anomaly detection.

In contrast with decremental PCA with the LOO strategy, we also consider the use of adding/duplicating a data instance of interest when applying PCA for outlier detection. This setting is especially practical for streaming data anomaly detection problems. To be more precise, when receiving a new target instance x_t, we solve the following PCA problem and 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

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. 3, Issue 8, August 2015

received data instance (viewed as a target instance) is an outlier. If the recently received data points are normal ones, adding such instances will not significantly affect the principal directions (and vice versa).

IV. OVERSAMPLING PCA FOR ANOMALY DETECTION

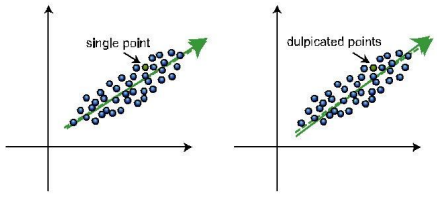
For practical anomaly detection problems, the size of the data set is typically large, and thus it might not be easy to observe the variation of principal directions caused by the presence of a single outlier. Furthermore, in the above PCA framework for anomaly detection, we need to perform n PCA analysis 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 over-sampling PCA (osPCA) together with an online updating strategy will address the above issues, as we now discuss.

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 large-scale anomaly detection problems.

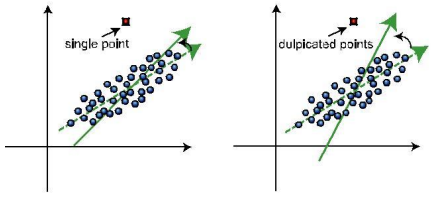
The proposed osPCA scheme will duplicate the target instance multiple times, and the idea is to amplify the effect of outlier rather than that of normal data. we can focus on extracting and approximating the dominant principal direction in an online fashion, instead of calculating multiple eigenvectors carefully.

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.



(a) Over-sampling a normal data point



(b) Over-sampling an outlier

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 the sensitivity of deriving the associated dominant eigen-vector can be studied as follows:

The above theoretical analysis supports our use of the variation of the dominant eigenvector for anomaly detec-tion. Using (12), while we can theoretically estimate the perturbed eigenvector u_ with a residual for an over-sampled target instance, such an estimation is associated with the residual term Oð_^{2}Þ, and _ is a function of n~ (and thus a function of the oversampling ratio r). Based on (12), while a larger r values will more significantly affect the resulting principal direction, the presence of the residual term prevents us from performing further theoretical

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. 3, Issue 8, August 2015

evaluation or comparisons (e.g., threshold determination).. No matter how larger the oversampling ratio r is, the presence of outlier data will affect more on the dominant eigenvector than a normal instance does.

V. EXPERIMENTAL RESULTS

Anomaly Detection on Synthetic and Real-World Data

Two-Dimensional Synthetic Data Set

To verify the feasibility of our proposed algorithm, we conduct experiments on both synthetic and real data sets. We first generate a two-dimensional synthetic data, which consists of 190 normal instances (shown in blue dots in Fig. 3a) and 10 deviated instances (red stars in Fig. 3a). The normal data points are sampled from the following multivariate normal distribution
We note that each deviated data point is sampled from a different multivariate normal distribution

TABLE 1
Description of Data Sets

Data set	Size	Attributes	Classes
pima	768	8	2
splice	1000	60	2
pnedigits	7494	16	10
adult	48842	123	2
cod-rna	59535	8	2
kdd_tcp	190065	38	5

Fig. 3. Outlier detection results with the two-dimensional synthetic data. Range ½_5; 5&. We apply our online osPCA algorithm on the entire data set, and rank the score of outlierness (i.e., s_t in Section 3.2) accordingly. We aim to identify the top 5 percent of the data as deviated data, since this number is consistent with the number of outliers we generated. The scores of outlierness of all 200 data points are shown in Fig. 3b. It is clear that the scores of the deviated data (shown in red) are clearly different from those of normal data, and thus all outliers are detected by setting a proper threshold to filter the top 5 percent of the data. Note that we mark the filtered data points with black circles in Fig. 3a. This initial result on a simple synthetic data set shows the effectiveness of our proposed algorithm.

Real-World Data Sets

Next, we evaluate the proposed method on six real data sets. The detailed information for each data set is presented in Table 1. The pendigits , pima, and adult data sets are from the UCI repository of machine learning data archive [25]. The splice and cod-rna are available at http:// www.csie.ntu.edu.tw/cjlin/libsvmtools/data sets/, and the KDD intrusion detection data set is available at http:// kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

To compare the anomaly detection results of our proposed method with other methods For the pen digits data set, we consider the digit “0” as the normal data instances (a total of 780 instances) and use other digits “1” to “9” (20 data samples randomly chosen for each category) as the outliers to be detected. For other data sets for binary classification. We consider the data from the majority class as normal data and randomly select 1 percent data instances from the minority class as outlier samples. In all our experiments, we repeat the procedure with 5 random trials.

From these three tables, we observe that our proposed online osPCA consistently achieved better or comparable results, while ours is the most computationally efficient one among the methods considered. By comparing the first and the second (or third) columns in Tables 3 and 4, it is interesting to note that the AUC score of osPCA is significantly

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. 3, Issue 8, August 2015

better than that of dPCA (without oversam-pling strategy).Comparing the second and the third columns, we note that the performance of our proposed online osPCA is comparable to that of osPCA with power method. This observation is consistent with our discussion in that using the proposed online approximation technique, our online osPCA is able to produce the approximated version of the principal direction without sacrificing computation and memory requirements.

For the KDD intrusion detection data set, there are four categories of attacks to be considered as outliers:

DOS: denial-of-service.

R2L: unauthorized access from a remote machine.

U2R: unauthorized access to local super user (root) privileges.

Probe: surveillance and other probing.

We use binary and continuous features (38 features) and focus on the 10 percent training subset under the tcp protocol. The size of normal data is 76,813. In this experiment, data points from four different attacks are considered as outliers.

However, in practical scenarios, even the training normal data collected in advance can be contaminated by noise or incorrect data labeling. Data cleaning and online detection. In the data cleaning phase, our goal is to filter out the most deviated data using our osPCA before performing online anomaly detection. In our implementation, we choose to disregard 5 percent of the training normal data after this data cleaning process, and we use the smallest score of outlierness (i.e., s_t) of the remaining training data instances as the threshold for outlier detection.

We now our proposed osPCA for online anomaly detection using the KDD data set. We first extract 2,000 normal instances points from the data set for training. In the data cleaning phase, we filter the top 5 percent (100 points) to avoid noisy training data or those with incorrect class labels. Next, we extract the dominant principal direction using our online osPCA, and we use this principal direction to calculate the score of outlierness of each receiving test input.

VI. CONCLUSION

In this paper, we proposed an online anomaly detection method based on oversample PCA. We showed that the osPCA with LOO strategy will amplify the effect of outliers, and thus When oversampling a data instance, our proposed online updating technique enables the osPCA to efficiently update the principal direction without solving eigenvalue decomposition problems. Furthermore, our method does not need to keep the entire covariance or data matrices during the online detection process.

Future research will be directed to the following anomaly detection scenarios normal data with multiclustering structure, and data in a extremely high dimensional space. The “curse of dimensionality” problem in a extremely high-dimensional space. In our proposed method, although we are able to handle high-dimensional data since we do not need to compute or to keep the covariance matrix

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. 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.

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. 3, Issue 8, August 2015

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.
11. V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley&Sons, 1994.
12. W. Jin, A.K.H. Tung, J. Han, and W. Wang, "Ranking Outliers Using Symmetric Neighborhood Relationship," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2006.