Anomaly Detection via Online Oversampling Principal Component Analysis

MUTHOJU DIVYA SRI
Department of Computer Science & Engineering, (M.Tech.)
Sindura College of Engineering and Technology
Ramagundam,Telangana

K.GEETA
Asst.Professor
Department of Computer Science & Engineering
M.tech
Sindura College of Engineering and Technology
Ramagundam,Telangana

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 over-sampling 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 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 over-sampling 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: *Clustering, Anomaly Detection, Multivariate Outlier Detection, Mixture Model, EM, Visualization, Explanation*

1. Introduction

The huge increase in the amount and complexity of reachable information in the World Wide Web caused an excessive demand for tools and techniques that can handle data semantically. Most people believe they can easily find the information they're looking for on the Web. They simply browse from the prelisted entry points in hierarchical directories (like yahoo.com) or start with a list of keywords in a search engine. However, many Web information services deliver inconsistent, inaccurate, incomplete, and often irrelevant results. For many reasons, existing Web search techniques have significant deficiencies with respect to robustness, flexibility, and precision. The disadvantage of the traditional search can be overcome with the proposal of semantic web. Semantic web also called the intelligent web or next generation web. Semantic web is approach towards understand the meaning of the contents. Semantic information is stored in the form of ontologies.To deal with this issue; ontologism are proposed for knowledge representation, which are nowadays the

backbone of semantic web applications. Both the information extraction and retrieval processes can benefit from such metadata, which gives semantics to plain text. The current WWW has a huge amount of data that is often unstructured and usually only human understandable.
The Semantic Web aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user. There are so many techniques to represent the semantic web information and the data mining techniques to retrieve the information from the semantic web.
We note that the above framework can be considered as a detrimental PCA (DPCA) based approach for anomaly detection. While it works well for applications with moderate dataset size, the variation of principal directions might not be significant when the size of the dataset 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.

2. 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 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.

3. Anomaly Detection via Principal Component Analysis

We first briefly review the PCA algorithm in Section III.A Based on the leave-one-out (LOO) strategy, Section III.B presents our study on the effect of outliers on the derived principal directions.

A. 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] where each row 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. Where U is a matrix consisting of 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, 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. if there are 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. We will further discuss this issue in the following subsections, and explain how we advance this property for anomaly detection.

1. The 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 out liernes of the target data point. We use Figure 1 to illustrate the above observation. We note that the clustered blue circles in Figure 1 represent normal data instances, the red square denotes an outlier, and the green arrow is the dominant principal direction. From Figure 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. On the other hand, this angle will be small when a normal data point is added. Therefore, we will use this property to determine the out liernes of the target data point using the LOO strategy. 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. If the target instance is x t , we next compute the leading principal direction without x t present. To identify the outliers in a dataset, we simply repeat this procedure times with the LOO strategy (one for each target instance).

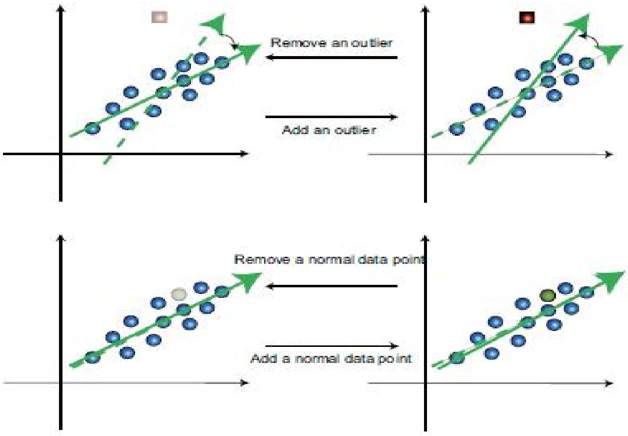

Fig.1. the effects of adding/removing an outlier or a normal data instance on the principal directions.

4. Existing System

The existing approaches can be divided into three categories:

1. distribution (statistical),
2. distance and
3. 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. 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. 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

A.Disadvantages of Existing System

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.

5. Proposed System

PCA is a well known unsupervised dimension reduction method, which determines the principal directions of the data distribution. 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.

A. Advantage of Proposed System

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.

6. System Architecture

We use binary and continuous features (38 features) and focus on the 10% training subset under the tcp protocol. The size of normal data is 76813. In this experiment, data points from four different attacks are considered as outliers. Table 6 shows detection performance (in terms of AUC) and the numbers of test samples of each attack category. Only LOF is used for comparison, since it is shown to outperform the ABOD method online osPCA again achieved comparable performance with LOF, while the LOF required significant longer computation time. Nevertheless, the effectiveness of our online osPCA is verified by the experiments conducted in this section, and it is clear that our approach is the most computationally efficient one among the methods we considered for comparison.

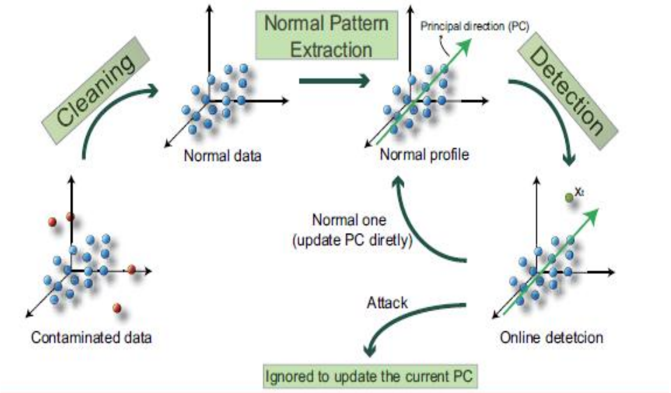

Fig. 2. The framework of our online anomaly detection.

7. Conclusion

In this paper, we proposed an online anomaly detection method based on over-sample PCA. We showed that the osPCA with LOO strategy will amplify the effect of outliers, and thus we can successfully use the variation of the dominant principal direction to identify the presence of rare but abnormal data. 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. Therefore, compared with other anomaly detection methods, our approach is able to achieve satisfactory results while significantly reducing computational costs and memory requirements. Thus, our online osPCA is preferable for online large-scale or streaming data problems. Future research will be directed to the following anomaly detection scenarios: normal data with multi clustering structure, and data in a 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 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, PCA might not be preferable in estimating the principal directions for such kind of data. Therefore, we will pursue the study of these issues in our future work.

References

1. M. Hawkins, Identification of Outliers. Chapman and Hall, 1980.
2. L. Huang, X. Nguyen, M. Garofalakis, M. Jordan, A. D. Joseph, and N. Taft, “In-network pca and anomaly detection,” in Proceeding of Advances in Neural Information Processing Systems 19, 2007. H.-P. Kriegel, M. Schubert, and A. Zimek, “Angle-based outlier detection in high-dimensional data,” in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008.
3. M.Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in Proceeding of the 2000 ACM SIGMOD International Conference on Management of Data, 2000.
4. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 15:1–58, 2009.