International Journal of Research in Advent Technology, Vol.2, No.2, February 2014

E-ISSN: 2321-9637

PCA Based Anomaly Detection

P. Rameswara Anand[1], , Tulasi Krishna Kumar.K [2]

Department of Computer Science and Engineering, Jigjiga University, Jigjiga, Ethiopi[1],

Department of Computer Science and Engineering,Yogananda Institute of Technology & Science,Tirupathi, Andhra Pradesh, India [2]

ABSTRACT- Anomaly detection is the process of identifying unusual behavior. It is widely used in data mining, for example, to identify fraud, customer behavioral change, and manufacturing flaws. We discuss how a probabilistic framework can elegantly support methods to automatically explain why observations are anomalous, assign a degree of anomalies, visualize the normal and abnormal observations and automatically name the clusters. To our knowledge, interactive visualization of anomalies has not previously been addressed, nor automatic naming of clusters for verification in the anomaly detection field. We specifically discuss anomaly detection using mixture models and the EM algorithm, however our ideas can be generalized to anomaly detection in other probabilistic settings. We implement our ideas in the SGI MineSet product as a mining plug-in re-using the MineSet visualizers. 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, MineSet**.**

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[1]. 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; ontologies are proposed [5] 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.

2 RELATED WORK

In the past, many outlier detection methods have been proposed [1], [2], [5], [10], [11], [12], [13], [14], [15]. Typically, these existing approaches can be divided into three categories: distribution (statistical), distance and density based methods. Statistical approaches [1], [11] 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 [10], [13], [14], 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 (e.g. multi-clustered structure). 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 [2], [12]. 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 [2]. 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 dataset is large.

Besides the above work, some anomaly detection approaches are recently proposed [5], [15], [16]. Among them, the angle-based outlier detection (ABOD) method [5] 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 large-scale problems (batch or online modes), 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 incremental based anomaly detection methods have been recently proposed [17], [18], 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 local outlier factors 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 $O(np)$ [17], [18], where n and p are the size and dimensionality of the data subset of interest, respectively. In [18], Ahmed proposed an online kernel density estimation for anomlay detection, but the proposed algorithm requires at least $O(np^2 + n^2)$ for computation complexity [18]. In online settings or large-scale data problems, the aforementioned methods

3 ANOMALY DETECTION VIA PRINCIPAL COMPONENT ANALYSIS

We first briefly review the PCA algorithm in Section 3.1.Based on the leave-one-out (LOO) strategy, Section 3.2 presents our study on the effect of outliers on the derived principal directions.

3.1 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.

$$\max_{\mathbf{U}\in\mathbb{R}^{p\times k},\|\mathbf{U}\|=\mathbf{I}} \sum_{i=1}^{n} \mathbf{U}^{\top}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\mathbf{U},$$

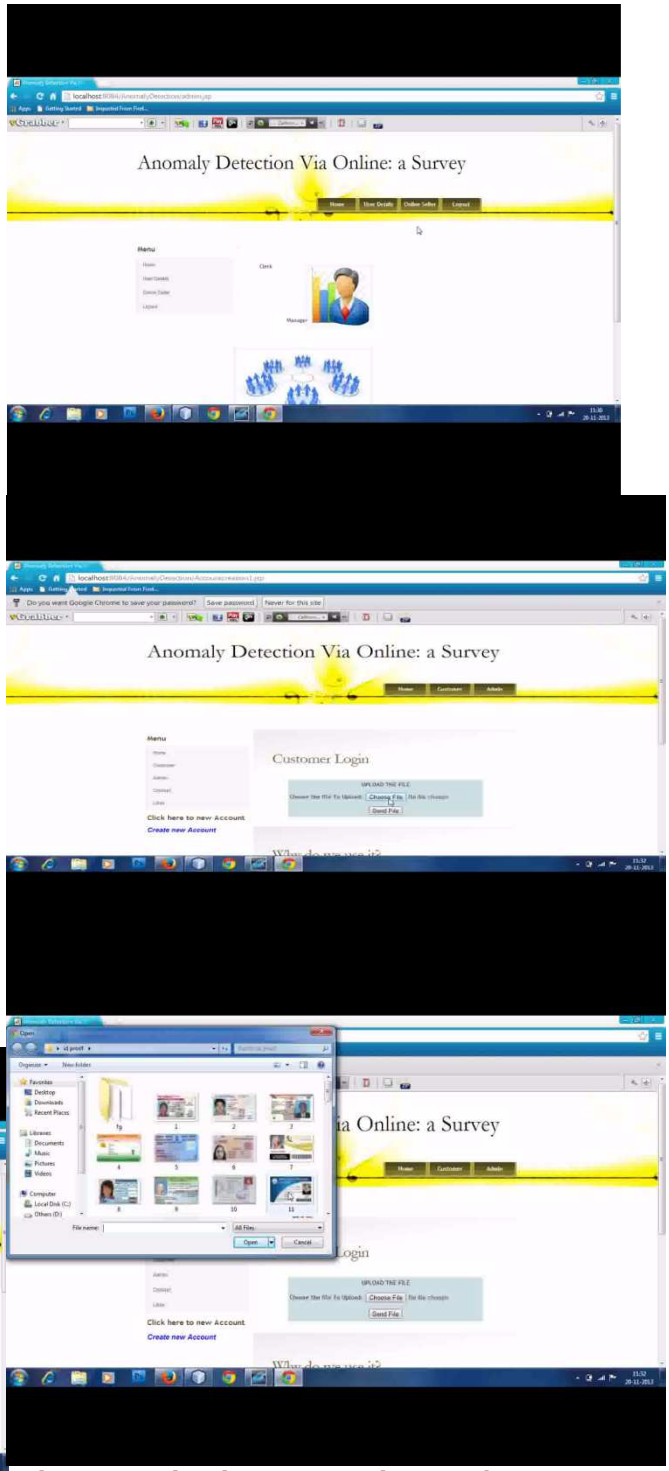
where U is a matrix consisting of k dominant eigenvectors. From this formulation, one can see that the standard PCA can be viewed as a task of determining a subspace where the projected data has the largest variation. Alternatively, one can approach the PCA problem as minimizing the data reconstruction error, 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. As shown in [19], 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.

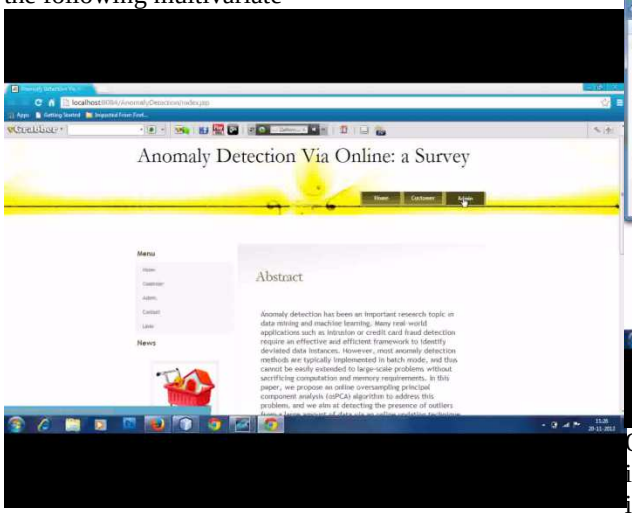
3.2 The Use of PCA for Anomaly Detection

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 outlierness 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 outlierness 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).

4 EXPERIMENTAL RESULTS

To verify the feasibility of our proposed algorithm, we conduct experiments on both synthetic and real data sets. We first generate a 2-D synthetic data, which consists of 190 normal instances and deviated instances . The normal data points are sampled from the following multivariate





Online anomaly detection results on the KDD intrusion detection data set. Note that TP and FP indicate true and false positive rates, respectively. The runtime estimate reports the testing time in determining the anomaly of a newly received target instance.

5. 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.

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: Identifyingdensity-based local outliers," in Proceeding of the 2000 ACMSIGMOD International Conference on Management of Data, 2000.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: Asurvey," ACM Computing Surveys, vol. 41, no. 3, pp. 15:1–58, 2009.

[4] L. Huang, X. Nguyen, M. Garofalakis, M. Jordan, A. D. Joseph,and N. Taft, "In-network pca and anomaly detection," in Proceedingof 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," in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008.

[6] A. Lazarevic, L. Ert oz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in Proceedings of the Third SIAM International Conference on Data Mining, 2003.

[7] X. Song, M. Wu, and C. J. andSanjay Ranka, "Conditional anomaly detection," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 5, pp. 631–645, 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," in Proceeding of the International Symposium on Neural Networks, 2004.

[10] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," IEEE Transactions on Knowledge and Data Engineering, 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," in Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2006.

[13] N. L. D. Khoa and S. Chawla, "Robust outlier detection using commute time and eigenspace embedding," in Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2010.

[14] E. M. Knox and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," in Proceedings of the International Conference on Very Large Data Bases, 1998.

[15] H.-P. Kriegel, P. Kr oger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data,"in Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009.

[16] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in Proceeding of ACM SIGMOD international conference on Management of data, 2001.

[17] D. Pokrajac, A. Lazarevic, and L. Latecki, "Incremental local outlier detection for data streams," in Proceeding of IEEE Symposium on Computational Intelligence and Data Mining, 2007.

[18] T. Ahmed, "Online anomaly detection using KDE," in Proceedings of IEEE conference on Global telecommunications, 2009.

[19] Y.-R. Yeh, Z.-Y. Lee, and Y.-J. Lee, "Anomaly detection via oversampling principal component analysis," in Proceeding of the First KES International Symposium on Intelligent Decision Technologies,
2009, pp. 449–458.

[20] G. H. Golub and C. F. V. Loan, Matrix Computations. Johns Hopkins University Press, 1983.

[21] R. Sibson, "Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling," Journal of the Royal Statistical Society B, vol. 41, pp. 217–229, 1979.

Mr. P. Rameswara Anand obtained first class MCA degree from S.V.University in the year 1993 – and ever since he is in the teaching line taking classes to BTech and MCA students in various institutions. In the year 2010, he secured first class M.Tech degree from Nagarjuna University. In the year 2012, he was qualified in "Andhra Pradesh State Eligibility Test(APSET) for Assistant Professors" Conducted by Osmania University-Hyderabad. Currently, he is working as a Senior Lecturer in Jigjiga University – Ethiopia.

Tulasi Krishna Kumar.K: is an assistant professor of Computer Sciences & Information Technology, works as Placement Officer at YITS, Tirupati, India. As he is in the field of teaching and research, He received his Master of Technology degree in the field of Computer Science & Engineering. Contact him at tulasikrishnakumar@gmail.com