IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VIII (Mar-Apr. 2014), PP 31-34
www.iosrjournals.org

An Enhanced Detection of Outlier using Independent Component Analysis among Multiple Data Instances via Oversampling

R.Krithigarani[1], R.Karthik[2]
[1] *(Department of CSE, SVS College of Engineering/Anna University, India)*
[2] *(Department of CSE, SVS College of Engineering/Anna University, India)*

Abstract: *Anomaly is a pattern of data that does not conforms to expected behavior. It is also referred as outlier, exceptions, peculiarities, surprise etc. Anomaly detection aims to identify a small group of instances which deviates from the existing data. It needs to solve an unsupervised yet unstable data learning problem.Detecting an anomaly is an essential research topic in data mining to solve the real world applications like intrusion detection, homeland security to identify the deviated data instances. Mostly anomaly detection methods are implemented in batch mode it requires more computation and memory. Existing system online oversampling Principal Component Analysis (osPCA) algorithm to address this problem and for detecting the presence of outliers from a large amount of data via an online updating technique. In PCA normal data with multi clustering structure and data is in an extremely high dimensional space is not supported. It is typically not easy to use linear models such as PCA to estimate the data distribution if there exists multiple data instances. To overcome these problems and support multiple data instances we proposed a system called Independent Component Analysis (ICA) in which it is a technique of array processing and data analysis aiming at recovering unobserved data samples from oversampled dataset and it is also used to reduce the computation and memory requirement for outlier detection.*

Keywords: *Anomaly detection, independent component analysis, local outlier factor, principal component analysis.*

I. INTRODUCTION

The definition of outlier is ”an input data which deviates so much from existing input data as to cause in to activity that it was given by a some other method”. Anomaly detection is to define a region representing normal behavior and declare any observation in the data which does not belong to the normal region declared as an anomaly.Anomaly detection is used in the applications such as credit card fraud detection, intrusion detection system.

Deviated(anomaly) data can be obtained by finding principal direction on data. For calculating principal direction of data set uses the Leave One Out (LOO) strategy which does not considers the target instance and the original data set. The variation in the principal direction determines that the data is an outlier. Anomaly can also be identified by defining an threshold value. If the data set is large, the principal direction is not significant. In decremental PCA (dPCA) method in which it performance is good for limited size of data and not for high dimensional data. To find the duplicate instance we used oversampling PCA method on the over sampled data set. The outlier identified easily due to duplicates in the PCA, since outlier instances amplified. Oversampling requires more load on computing data by creating data matrix for every target instance. The over sampling is not suitable for the applications with streaming data.

II. RELATED WORK

Most of the anomaly detection algorithms are proposed in [2], [6], [8].The methods can be density based and distance based, statistical approach. In density based method [8] it uses Local Outlier Factor (LOF) in which it assigns a degree to each of the data instance and LOF is approximately determined to be 1.Higher the LOF declares that the data point is an outlier and lower the LOF declares that the data point is not an outlier. LOF judges the structure of data through the density of the data point. In distance based method [6] the distance between the data point and the neighbor data point is evaluated, if the distance between them is beyond the estimated threshold value then it is considered as an outlier. In statistical approach in which data is provide with certain standard and it identifies the outlier when data deviates the standard and also it is unfair for the data with noise. It is also not feasible for assumption of online and offline data. The latest outlier detection methodology is Angle Based Outlier Detection(ABOD) method in which it measure the angle between target instance and the existing data points and it determines data point as an outlier only when the angle between them is large. It has the limitation that it requires high computation complexity for measuring the angle for instance pairs is tedious.The above methods are not suitable for online and continuous data and also it requires higher memory

requirement and computation in high dimensional space. In this paper, we propose the ICA will substantially used among multiple instances to find an outlier.

III. OUTLIER DETECTION PCA AND ICA

3.1 Outlier detection via PCA

PCA is a well known unsupervised dimensionality 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 considered as the principal directions. The black circle represents normal data instances, the red circledenotes an outlier and the arrow is the dominant principal direction. 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.PCA requires the calculation of global mean and data covariance matrix, we found that both of them are sensitive to the presence of outliers. There are outliers present in the data dominant eigenvectors produced by PCA will be remarkably affected by themand this will produce a significant variation of the resulting principal directions.

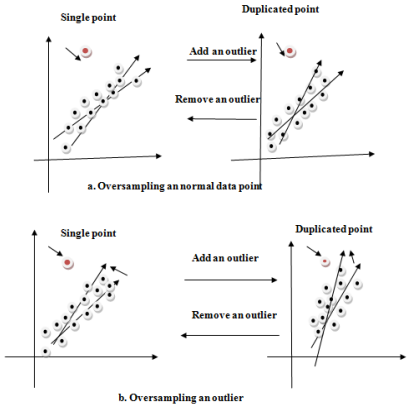

Fig. 1 The effect ofaddition and deletion of an outlier as a single/duplicated points.

3.2 Oversampling PCA (osPCA)

The 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, online osPCA method aims to efficiently determine the anomaly of each target instance without sacrificing computation and memory efficiency. More specifically, if the target instance is an outlier, this over- sampling scheme allows us to overemphasize its effect on the most dominant eigenvector and thus we can focus on extracting and approximating the dominant principal direction in an online fashion, instead of calculating multiple eigenvectors carefully.

The osPCA will duplicate the target instances n no of times from data set and computes the score of outlier.If the score is above the threshold it is determined to be an outlier.The power method for osPCA is also provides the solution in solving the eigen value decomposition problem. The solving of PCA to calculate the principal directions n times using n instances and it is very costlier and it prohibits the online data in anomaly detection. Using LOO strategy it is not necessary to recomputed covariance matrix for all instances and thus find the difference between the normal and outlier data is easily identified.

The dominant eigen value is more than others and it is shown that power method requires only multiplication of matrix not matrix decomposition. So this method is better for reducing the computation cost in calculation of dominant principal direction. The cost of memory is O(np) for matrix multiplication. However the use of power method does not guarantee fast convergence and it also need to store covariance matrix which not used in high dimensional data, so we prefer the least square approximation method.

3.3 Oversampling PCA with Anomaly Detection

Effective and efficient updating technique for osPCA, which allows us to determine the principal direction of the data. This updating process makes anomaly detection in online or streaming data settings feasible. More importantly, since we only need to calculate the solution of the original PCA offline, we do not need to keep the entire covariance or outer matrix in the entire updating process. Once the final principal direction is determined, we use the cosine similarity to determine the difference between the current solution and the original one (without oversampling), and thus the score of outlierness for the target instance can be

determined accordingly. An algorithm anomaly detection via online oversampling PCA in which data matrix with values are transposed and it is weighted with a value. Score of outlinerness is found if s is higher than the threshold x is an outlier for computing the principal directions and finds the cosine similarity to obtain the difference between the existing value and new instance.

3.4 Independent Component Analysis (ICA)

ICA finds the independent components (aka factors, latent variables or sources of outlier data detection) by maximizing the statistical independence of the estimated components. We may choose one of many ways to define independence, and this choice governs the form of the ICA algorithms. The Minimization-of-Mutual information (MMI) family of ICA algorithms uses measures like Kullback-Leibler Divergence and maximum-entropy. Independence assumption is correct, blind ICA separation of a mixed data gives very good results. It is also used for mixed data and that are used to generate by a mixing for analysis purposes. It is closely related to (or even a special case of) the search for a factorial code of the data, to detect the outlier in the data i.e., a new vector-valued representation of each data vector such that it gets uniquely encoded by the resulting code vector (loss-free coding), but the code components are statistically independent or not. If the data of the outlier is statically independent it detects the outlier.

The procedure is ICA can use a statistical "latent variables" model. Assume that we observe *n* linear mixtures of the outlier data $x_1, ...,x_n$ of *n* independent components $x_j= a_{j1}s_1+a_{j2}s_2+...+a_{jn}s_n$, for all *j*. Then now dropped the time index *t*; in the ICA model we assume that each mixture x_j as well as each independent component $s_k$is a random variable of the data instead of a proper time signal. The observed values $x_j(t)$, e.g., the number of outlier in the data samples are then a sample of this random variable. It is convenient to use vector-matrix notation instead of the sums like in the previous equation. Let us denote by **x** the random vector whose elements are the mixtures $x_1, ...,x_n$, and likewise by **s** the random vector with elements $s_1, ..., s_n$. Let us denote by A the matrix with elements a_{ij} . Generally, bold lower case letters indicate vectors and bold upper-case letters denote matrices. All vectors are understood as column vectors; thus $\mathbf{x}_T$, or the transpose of **x**, is a row vector. Using this vector-matrix notation, the above mixing model is written as

$$\mathbf{x = As} \qquad (1)$$

It is called independent component analysis, or ICA model. The ICA model is a generative model, which means that it describes how the observed data are generated by a process of mixing the components s_i. The independent components are latent variables, meaning that they cannot be directly observed. Also the mixing matrix is assumed to be unknown. All we observe is the random vector **x**, and we must estimate both **A** and **s** using it. This must be done under as general assumptions as possible. The starting point for ICA is the very simple assumption that the components s_i are statistically *independent*. It will be seen below that we must also assume that the independent component must have *non-gaussian* distributions. However, in the basic model we do *not* assume these distributions known (if they are known, the problem is considerably simplified.) For simplicity, we are also assuming that the unknown mixing matrix is square, but this assumption can be sometimes relaxed. Then, after estimating the matrix A that is the outlier data, we can compute its inverse, say W, and obtain the independent component simply by,

$$\mathbf{s = Wx} \qquad (2)$$

Finally based on the matrix we find the outlier for anomaly detection via oversampling.

Table 1 Comparison of several anomaly detection algorithms

Terms	ICA	osPCA	Online osPCA	Local Outlier Factor
Memory Requirements	O(n)	O(np)	O(p)	O(np)
Computation Complexity	O(nk)	O(mnp)	O(p)	$O(n^2p+k)$

From the table1, it compares the memory requirement and computational complexity of anomaly detection methods that are LOF, osPCA,OnlineosPCA and proposed ICA. The table represents the memory requirements and computation complexity of all methods for finding the outlier in new data instance. ICA requires memory as O(n) and its computational complexity as O(nk), osPCA needs memory as O(np) and computational complexity as O(mnp), Online osPCA is in need of memory and computational complexity as O(p) and finally LOF requires memory as O(np) and computational complexity as $O(n^2p+k)$. The memory requirement is reduced from O(np) to O(n) and Computation complexity is reduced from O(mnp) to O(nk).Thus the memory requirement and the computational complexity reduced even though the data are of multiple instances and it shows that proposed method ICA is efficient in finding outlier when comparing with other anomaly detection algorithms. Comparisons of Proposed ICA, osPCA, online osPCA, LOF is in table1 for anomaly detection in terms of memory requirements and computational complexity. Where n and p are the number of values and data

respectively. It requires the number of iterations m and the number of nearest neighbor k is used in local outlier factor.

IV. RESULTS

4.1 Anomaly Detection using real-world data

The data set used for detecting anomaly is kdd intrusion detection and its available in http://database//kdddatabase/kdddata.html. The system uses graph in which to assess the outlier detection and larger the value more the chances of outlier. So from the performance of data we found that the proposed ICA is computationally better and efficient when comparing with other algorithms. This confirms that the use of Independent Component Analysis reduces the time for finding the outlier and categories used are DOS is for Denial of Service, Probe is for monitoring, and S23 is for unauthorised access of an anonymous user.The following table shows that the results on kdd dataset.

Table 2 Results of detection of anomaly in KDD dataset

Type of an anomaly	Anomaly size	ICA	osPCA	Online osPCA
		Time Utilised(in Seconds)	Time Utilised(in Seconds)	Time Utilised(in Seconds)
DOS	50	1.596	33.81	2.596
Probe	50	1.594	27.60	2.595
S23	30	1.590	33.05	2.593
Attacks as a whole	70	1.699	34.23	2.823

V. CONCLUSION

In this paper we proposed an anomaly detection method based on Independent Component Analysis. We depicted that amplification of outliers using LOO strategy and found the outlier with the help of principal directions. This method uses only the simple matrix calculation and also it does not require large memory and high computations. Finally our ICA is suitable for data in high dimensional space and also reducing the time to identify the outlier among multiple instances. Thus our approach is also preferable for large stream of data.

References

[1] Yuh-Jye Lee, Yi-RenYeh, and Yu- Chiang Frank Wang, Member, "Anomaly Detection via Online Oversampling Principal Component Analysis",*IEEE Trans.on Knowledge and Data Eng, vol 25, no.7,pp 1460-1470, July 2013.*

[2] 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.*

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

[4] X.Song, M.Wu, and C.J., and S.Ranka, "Conditional Anomaly Detection", *IEEE Trans.on Knowledge and Data Eng., vol. 19, no. 5, pp.631-645,May 2007.*

[5] L.Huang, X.Nguyen, M.Garofalakis, M.Jordan, A.D. Joseph, and N.Taft, "In Network PCA and AnomalyDetection", *Proc. Advances in Neural Information Processing Systems 19, 2007.*

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

[7] A.Lazarevic, L.Ertoz, V.Kumar, A.Ozgur, and J.Srivastava,"A Comparitive Study of Anomaly Detection Schemes In Network Intrusion Detection", *Proc. Third SIAM Int'l Conf. Data Mining 2003.*

[8] 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.*