International Journal of Engineering Research & Technology (IJERT)
NCRTS`14 Conference Proceedings
ISSN: 2278-0181

Detection of the Outliers for Large Scale Data

Authors: Prapulla C [1], Prof. Malatesh S H [2]
4th sem Mtech MSEC, HOD of CSE MSEC

Abstract: In data mining and machine learning the detection of the outliers has become a very important topic. An effective and efficient framework is needed to identify deviated data in many real world applications such as intrusion detection and credit card fraud. Many methods used for detection of the outliers are typically implemented in batch mode and that implementation on large scale data is difficult. For implementation of the batch mode on large scale data will lead to sacrifice of computation and memory requirement. In this paper, I address the computational and memory management issues and propose Online oversampling principal component analysis algorithm that aims at detecting outliers from large scale data via online updating technique. The prior principal component analysis based approaches, the data would not be stored in covariance matrix, and thus our approach would be basically interested in large scale data problems. In this algorithm the oversampling of target instances and extracting the principal direction of the data the algorithm allows us to determine the target instances according to the variation of the resulting eigen vector. The proposed framework need not have to explicitly compute the eigen vector and hence this favours the problem that is being addressed in the paper. The eigen vector is not computed explicitly and hence the limitation of computation and memory management is favoured. On comparison of other methods the proposed experimental method will be both accurate and efficient.

Keywords: Data mining, Outlier detection, Eigen vector, computational requirement, Memory management , Large scale data.

I INTRODUCTION

Detection of the outliers that aim at identifying a small group of instances that deviate remarkably from the existing data. A well -known definition of the outliers is given as observation that deviates so much from other observation that rise suspicious that it was deviated for other mechanism. This gives the general idea on outliers and motivates us to many outliers detection methods. The outliers can be found in many practical application like homeland security, credit card fraud detection .fault detection, cyber security. Now the question that rises here is that there is only few examples in real world but how to identify problems in unseen data and researches have to be done in this format. Our main concentration is drawn towards datamining and machine learning. The researches have been done for small data but finding data in large scale data is difficult hence in this paper we concentrate on large scale data. Distribution or principal direction of the data might be rare to indicate the deviation of the data but it enormously affect the solution model. There exists a sensitivity for the outliers for example the calculation of the data mean or least square solution of the associsated data. The detection of the outliers needs solution in unsupervised yet unbalanced data learning problems. We observe that principal directon of resulting normal data does affect , adding or deleting the abnormal data instantces. We can calculate the principal direction of the data set without the presences of target data set or that of the normal data set this is done by using "LOO strategy". Thus detection of the Outliers can be done by determining the variation of resulting principal directions. The eigen vectors will be calculated and differences between the eigen vector will indicate the outliers in large scale data. By understanding the differences in eigen vector of data it will be easy to identify the outliers in the defined data threshold or predefined data. The decremental PCA based approach for outliers detection can be considered for the above framework . This is classic for the application with moderate data size ,the variation in the principal direction. The framework will not be significate for large scale data set. Large scale data are basic example for real-world outlier detection problem . In large scale data adding and deleting of the data on target instance and cannot simply apply decremental PCA for finding outliers as they produces only negligible differences in the eigen vector. To be more practical in the approach to the problems we have direct our research towards Oversampling strategy . This advance framework will duplicate the target instances and by using new algorithm Oversampling of Principal component analysis(osPCA) will overcome problems in large scale data. The outlier instances will be amplified as the target data will be duplicated in the presences principal component analysis. Using this method the detection of the outliers will become easy. A dense covariance matrix has to be created for each target instance and this associated with PCA will solve the problem. Large scale data application will prohibit the use of our proposed model. There is requirement of storage for covariance matrix to produce approximated PCA solution and that can not be easily extended for large scale data or on online data. The osPCA algorithm is basically used for Online updating technique. The calculation of eigen vector is done efficiently with this algorithm witout performing any eigen analysis or storing of data. Compared with any other method that are popularly used for detection of outliers this method is more efficient and shows less computational costs and has significantly less requirement of memory, which is more important in large scale data.

II RELATED WORK:

Many outliers detection method were proposed in the past. These existing approaches are divided into three categories : Statistical approach , Distance based, distance based. In

Statistical approach there exists a predetermined distribution of data adthis approach main aim is to find outliers that deviate from such distribution. However it is assumed that most of the distributed data are univariante and hence lack of robustness for multidimensional data. However these methods are implemented on original data space directly, the solution might suffer from noise that is present in the data. The practical problem can never be understood from prior knowledge of distribution of data.

The distance between each point of interest and its neighbours is calculated for distance based methods. The target instance will be considered as outliers when the result is above some predetermined threshold. The distribution of data can become complex when there is no prior knowledge on data distribution. There are possibilities of getting improper result when outliers are detected improperly.

To overcome these problem density based problem was identified. One of the type of approach is to identify the outliers using density based local outliers factor that is used tom identify each data instances. LOF determines the degree of outliers based on the local density of each data which provides suspicious ranking score. The ability to estimate data structure via density estimation is the most important property of LOF. The user is allowed to identify outliers using sheltered data structure. However when the size of the data is large each instances is computationally expensive and it is not worth estimation of local data density.

There are many newly proposed approach proposed called as angle based outliers detection. Among them unique method the angle based outlier detection. In Angle based outliers detection we calculate the variation of angle between remaining data set and each target instance. It is observed that the outlier will produce smaller angle variation from the target data set. The computation complexity is the major concern of ABOD. This is not surprising as huge amount of data is paired. To generate approximate and original ABOD solution a fast ABOD algorithm is proposed. The variance of angles between target data instances and Kth nearest neighbour is the only difference between standard and fast algorithm ABOD.

III OUTLIER DETECTION VIA PCA

3.1 Principal Component Analysis

PCA is used to determine the principal direction of distribution of data ,which is well known for unsupervised dimension reduction method. The data covariance matrix and calculation of eigen vector is needed to obtain principal direction. Thus considering the principal direction these eigen vector are most informative among the vector in the original data space.

The last few eigen vectors will be discarded due to their negligible contribution. The major purpose of discarding the last few eigen vector is for dimension reduction. The global mean and data covariance matrix needs calculation. on these calculation for PCA we understood that they both are sensitive to the presence of Outliers. Dominate eigen vectors produced by PCA will remarkably affect the outliers present in the data. thus produce a significant variation that will result in principal direction.

3.2 Use of PCA for Outlier Detection

In this section we will study the variation that are seen in the principal direction when we add ,delete data instances and also how we utilize this property
to determine various outliers on the target data.

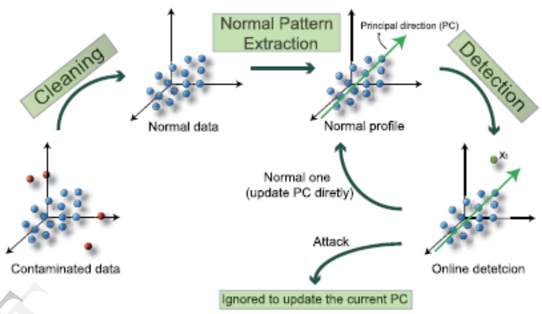

Fig 3.1

IV Outlier detection using Oversampling of Principal Component analysis

The size of data is typically large for practical detection of outliers and thus its not very easy to identify data that are variant from principal direction. These variation is caused by the presence of single outlier in the target data . In the frame work for detection of the outliers we have to perform p-direction space for nPCA analysis of data set with n data instances for large scale or online data that is not computationally feasible. The algorithm that we proposed will address all the above issues to get with online updating strategy.

Even when the size of data is large according to the principal component analysis we will understand and discuss how and why we are able to detect the presence of abnormal data. To solve the Eigen vector decomposition problem we apply a very well known power method to determine the principal direction. Next important issues we are addressing in this paper is about computational cost and this issue is solved by thinking on the issue in determining the principal direction ,we will also discuss the limitation and explain why we use the power method is not practical in online setting. In the next section we will present the least squares of osPCA followed which will efficiently solve problem in online updating algorithm.

ISSN: 2278-0181

4.1 Oversampling Principal Component Analysis

The resulting principal direction of the data on large scale will not be affected significantly by adding or removing a single outlier instances. Hence we advance to simple strategy and present an Oversampling of principal component analysis algorithm for large scale data for determining the outliers. The proposed algorithm will duplicate the target instances multiple times, and idea is to amplify the effect of outlier than that of normal data. performing the detection of outliers based on the dominate eigen vector is not sufficient. The osPCA framework mainly aims at determining the outliers of each target instances without scarifying memory and computation.

The oversampling might overemphasize its effect on most dominate eigen vector if the target instance is an outlier. Instead of calculating multiple eigen vector carefully it is better to focus on extracting approximate principal direction in an online fashion.

Clearly our major concern is about the computational and utilization of less memory. We cant perform cross validation or similar strategies to determine the parameter in advance as there is no training or validation data for practical detection of outliers.

V RESULTS

The results here are for credit card fraud identification. The algorithm used here is osPCA .

Fig 5.1 Home page

Fig 5.2 Login Page

Fig 5.3 Add items

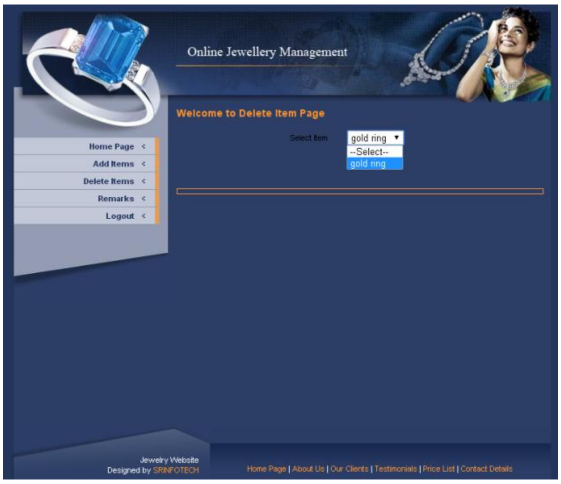

Fig 5.4 Delete Item

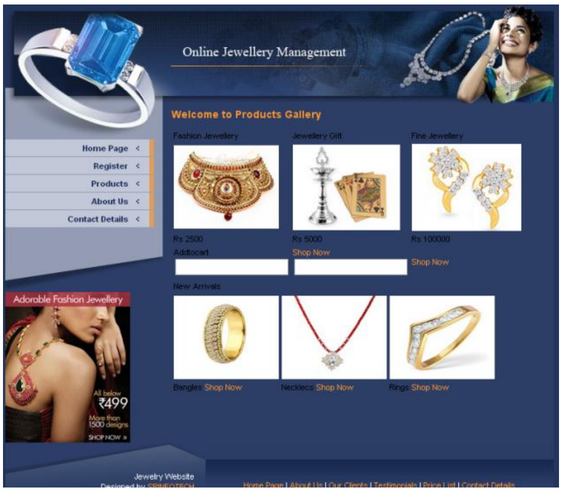

Fig 5.6 Product List

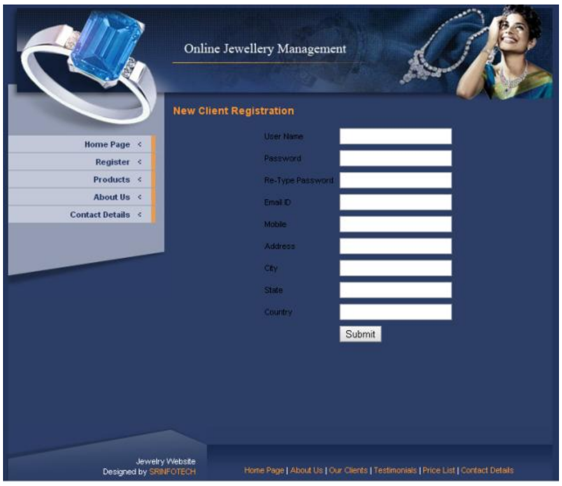

Fig 5.7 New client registration

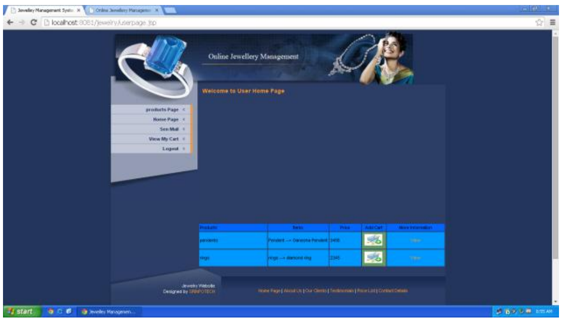

Fig 5.8 Added list

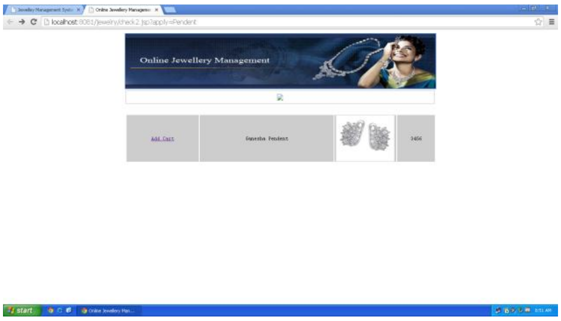

Fig 5.8 Carted Items

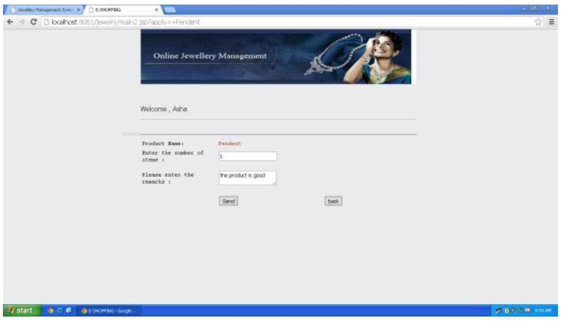

Fig 5.9 Shopped Items

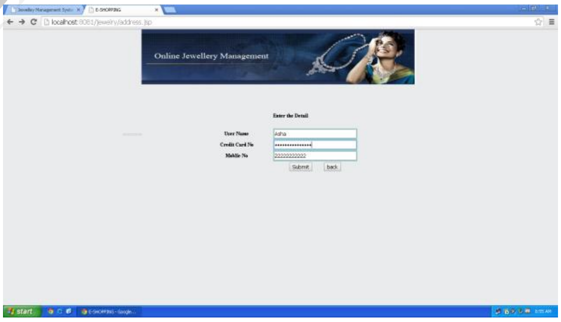

Fig 5.10 credit card details

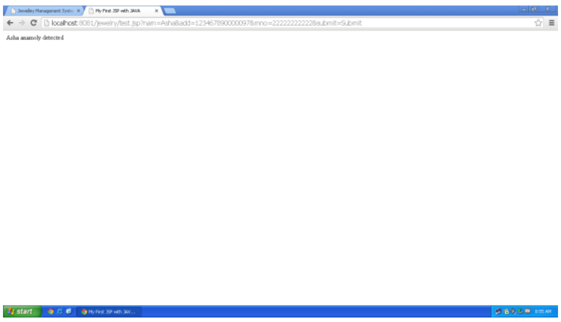

Fig 5.11 Anomaly deducted

VI CONCLUSIONS

In this paper I have proposed and algorithm oversampling PCA. This is a algorithm with an enhancement to PCA algorithm. This algorithm is used to reduce computational power and memory consumption. This algorithm is mainly developed for online database. As database are instantly being updated in an online shopping. Here we have computed eigen vector, covariance matrix and score. Some variation in the score can be detected as anomaly.

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] A. 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.
[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.
[13] N.L.D. Khoa and S. Chawla, "Robust Outlier Detection Using Commute Time and Eigenspace Embedding," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2010.
[14] E.M. Knox and R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. Int'l Conf. Very Large Data Bases, 1998.
[15] H.-P. Kriegel, P. Kro¨ger, E. Schubert, and A. Zimek, "Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2009.
[16] C.C. Aggarwal and P.S. Yu, "Outlier Detection for High Dimensional Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2001.
[17] D. Pokrajac, A. Lazarevic, and L. Latecki, "Incremental Local Outlier Detection for Data Streams," Proc. IEEE Symp. Computational Intelligence and Data Mining, 2007.
[18] T. Ahmed, "Online Anomaly Detection using KDE," Proc. IEEE Conf. Global Telecomm., 2009.
[19] Y.-R. Yeh, Z.-Y. Lee, and Y.-J. Lee, "Anomaly Detection via Oversampling Principal Component Analysis," Proc. First KES Int'l Symp. Intelligent Decision Technologies, pp. 449-458, 2009,
[20] G.H. Golub and C.F.V. Loan, Matrix Computations. Johns Hopkins Univ. Press, 1983.
[21] R. Sibson, "Studies in the Robustness of Multidimensional Scaling: Perturbational Analysis of Classical Scaling," J. Royal Statistical Soc. B, vol. 41, pp. 217-229, 1979.
[22] B. Yang, "Projection Approximation Subspace Tracking," IEEE Trans. Signal Processing, vol. 43, no. 1, pp. 95-107, Jan. 1995.
[23] S. Papadimitriou, J. Sun, and C. Faloutsos, "Streaming Pattern Discovery in Multiple Time-Series," Proc. 31st Int'l Conf. Very Large Data Bases, 2005.
[24] S. Haykin, Adaptive Filter Theory. Prentice Hall, 1991.
[25] A. Asuncion and D. Newman, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/mlearn/mlrepository.html, 2007.
[26] A.P. Bradley, "The Use of the Area under the Roc Curve in the Evaluation of Machine Learning Algorithms," Pattern Recognition, vol. 30, pp. 1145-1159, 1997.

Author: Prapulla C [1] is a student of VTU presently presuming MTech in computer Science Engineering in M S Engineering College

Author: Prof.Malatesh S H [2] is HoD for department Computer Science and is a professor in Computer Science.