IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Survey on Anomaly Detection in Web Usage Mining

Navareena .A [1], Kathiresan .V [2], Dr. S. Gunasekaran[3]

Student-M.E, Dept of Software Engineering, Coimbatore Institute of Engineering and Technology, Coimbatore, India [1]

Assistant Professor, Department of Computer Science, Coimbatore Institute of Engineering and Technology, Coimbatore, India [2]

Head of the Department, Department of Computer Science, Coimbatore Institute of Engineering and Technology, Coimbatore, India [3]

Abstract: In machine learning, Anomaly detection is widely processed for many real-world applications mainly intrusion or fraud detection which also require an actual and efficient framework to classify differed data instances. In this paper, the study of anomaly detection in web usage mining is done. In which, the various methods and functions that are used in various papers in various methods of analysis. This survey attempts to provide broad overview of the research on anomaly detection. Also have done the work in the form of collected existing techniques into different categories based on the method adopted by each procedure. For every paper the used techniques and what that author process are all explained in various behaviors.

Keywords: Anomaly detection, outlier detection.

I. INTRODUCTION

Web mining is the means of utilizing the methods of data mining in order to induce and extract useful information from the web. It especially aims in discovering massive data sources available in the web, which have distinctive features from the data in conventional database management system.

II. WEB MINING CLASSIFICATION

Basically, web mining could be classified into three, they are,

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Web content mining tries to discover valuable information from the web document. It is also termed as text mining.
Content data is the collection of facts a web page is designed to contain. It may consist of text, audio, video, images. Web structure mining does the work of modeling websites to enhance linking structures. This can be further classified as hyperlinks and document structure.

III. WEB USAGE MINING IN DETAIL

Web usage mining involve in revealing the underlying work in access patterns in web transactions. It is done to find the pattern from the web usage data for serving the needs of web based applications. For this purpose user logs like IP address, page reference, and access time are collected by the server. It is further classified depending on the data that is being used.

- Application level data
- Application server data
- Web server data

Many corporations and even government agencies are getting benefitted by web usage mining. One among the upcoming trend that has emerged out for the welfare of people's service is e-commerce. By this better customer relationship can be built over, more and more modernized ways are adopted for this. Resources are being utilized properly.

IV. APPLICATIONS

- ❖ Mining web logs to improve the website organization
- ❖ Clustering use queries from web logs for related query
- ❖ Using ontology based user preferences to improve web search

V. ANOMALY DETECTION

Something which gets deviated from the normal order or rule is called as anomaly. It is determined out and further used in certain applications. There are three broad categories of anomaly detection. For prolonged period of time, this field is of a very active research in computer science field. An object's outliners' will be more different from the local neighborhood.
They are,

- Supervised learning
- Unsupervised learning
- Semi-supervised learning

The various fields in which anomaly detection have found applications are detecting eco-system disturbances, intrusion detection, fraud detection, fault identification.
It has some peculiar benefits like

- Trouble shooting time will be reduced to the half
- No requirement in subject matter knowledge

- Involve predicting real time results
- More and more human interpretation will be reduce to equal half

VI. TECHNIQUES DEPLOYED

- Density based technique
- Fuzzy logic based outlier detection
- Cluster analysis based outlier detection
- Deviation from association rules and frequent item sets.

VII. LITERATURE SURVEY

Markus et all.,[3] described in the paper about many KDD applications, such as detecting criminal activities in E-commerce, finding the rare instances or the outliers, can be more interesting than finding the common patterns. In this paper the existing work has done in the form of outlier detection concerns being an outlier as a binary property. So they have done the proposed as in many scenarios, which is more important to assign to each object a grade of existence to an outlier. The formation of this degree is called the local outlier factor (LOF) of an object. Also the author proves that the LOF can be used to find outliers which perform to be important, but this is not identified with the defined existing approaches. Finally, he has done a performance for the evaluation of the defined method in which it confirms that the proposed approach of finding local outliers can be practically deployed.

Ling Huang et al, [5] explain about the process of PCA in network in which the large amounts of distributed systems are connected to the network. This PCA has processed in the form of tracking projection of the data in subspace. They also proved that the process of work done through for the highly aggregated networks, with a limited number of large nodes and at coarse time scales. In their approach, however, has scalability limitations and overcome these limitations by means of PCA-based anomaly detector in which adaptive local data send to a coordinator just enough data to enable accurate global detection. For this they used stochastic matrix perturbation analysis which characterizes the tradeoff between the accuracy of anomaly detection and the amount of data communicated over the network.

Xiuyao Song et al, [8] have done the work based on Conditional Anomaly Detection. If there are the processes in the software for about too many unremarkable data points are returned to the user labeled as candidate anomalies, the software will be in disuse. To ensure the refunded anomalies are useful which are make use of domain knowledge that are provided by the user. Values are never in directly indicate the anomaly, but these attributes cant able to be ignored in which they have a direct effect on the expected distribution of the outcome attributes whose values can specify an anomalous remark. The authors have shown that a general-purpose method called conditional anomaly detection for taking such modifications among attributes into consideration.

Wen Jin et all [13] described in his paper about the in the database how the process of mining takes place in outliers which are deviate from the remaining dataset. Further, classical outlier analysis algorithms, current studies have focused on mining local outliers, such that the outliers that has density spreading from their nearby community. The approximation of density circulation at the location of an object has so far been based on the density distribution of its k-nearest neighbors. Finally the author concluded that their method is not only efficient and also effective in ranking outliers.

Yuh-Jye Leeetall, represented his concept in outlier as in the form of PCA it means principle component analysis. This processed as the information in which the huge amount of data's available. Such example in credit card fraud detection and intrusion detection which are used to deviated from the instance of data. They have used an online over sampling principal component analysis (osPCA) also they have aim at detecting the presence of outliers from a large amount of data via an online updating technique. In PCA method cant able to store the entire data as in the form of matrix or covariance matrix, and thus this 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 in this efficiency and management will be high.

VIII. CONCLUSION

In this survey, discussed possible different methods in which the problem of anomaly detection has been expressed in literature, and have tried to provide an overview of the various techniques in a process. For each and every category of anomaly detection methods, have identified a similar assumption regarding the idea of anomalous and normal data. With this work in future, the work concentration is about the assumptions made by different methods are considered and a try to make a real time application by using online oversampling principal component analysis method.

REFERENCES

[1] Yuh-Jye Lee,Yi-RenYeh, and Yu-Chiang Frank Wang,"AnomalyDetection via Online OversamplingPrincipal Component Analysis", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 7, July 2013.
[2] D.M. Hawkins, Identification of Outliers. Chapman and Hall, 1980.
[3] 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.
[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 15:1-15:58, 2009.
[5] 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.
[6] 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.
[7] A. Lazarevic, L. Ertoz, 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.

[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley&Sons, 1994.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] C.C. Aggarwal and P.S. Yu, "Outlier Detection for High Dimensional Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2001.

DOI 10.17148/IJARCCE.2015.41152