International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 07, July-2015

Web Usage Analysis and Web Bot Detection based on Outlier Detection

Respa Peter
Computer Science and Engineering
Adi Shankara Institute of Engineering and Technology
Kalady, India

Divya D
Computer Science and Engineering
Adi Shankara Institute of Engineering and Technology
Kalady,India

***Abstract*— For securing the ones network it is important to detect botnet. For improve the efficiency and accuracy in data mining use optimization techniques. One of the current application areas is outlier detection that has not been fully explored yet but has enormous potential. Web bots are type of outliers, They can be found in the web usage analysis process. .Particle Swarm Optimization (PSO) based on Hierarchical method is used in this paper to detect web bots among genuine user requests. In proposed scenario deal with tuning parameters in PSO algorithm for selecting the process. It is necessary to set different strategies for each PSO parameter. These parameter selections should ne optimum. HPSO algorithm provides high accuracy, fast convergence results. Less computational time to execute this process is the main advantage.**

Keywords- Particle Swarm Optimization, web bots, web usage mining.

I. INTRODUCTION

The outliers are data object which do not have the general behavior or characteristic or model of the data. An outlier is an observation that is numerically distant from the rest of the data. We introduce a novel method to find the outliers and strong outlier groups, based on the Maximum Flow Minimum Cut theorem from graph theory, and evaluate the outliers by outlier degrees. They are entirely different from existing set of data objects these type of data objects are considered as outliers. Web bots are included in this category. These web bots detection process is still unresolved efficiently, if they become undetected it will cause many errors in data mining processes. Detecting botnets is a critical need for securing one's network and the Internet at large. Bot traffic skews overall metrics, meaning site managers are unable to provide accurate reports for their dealership partners. Though the bot traffic will result in more visits to the dealer's listings, it won't lead to a spike in sales. This can devalue the auto shopping site in the dealer's eyes, and makes it difficult to sustain a successful relationship. Web bots can inflate the total number of clicks your PC ads receive by huge margins. This causes cost Per click (CPC) to go up and makes it impossible for you to tell how much of your traffic is genuine customers and how much of it is automated and irrelevant. As a result, without the use of web bot detection systems, our reports might show huge spikes in traffic but an equal decline in conversions.

Optimization based techniques have emerged as important methods to tackle the problems of efficiency and accuracy in data mining. One of the current application areas is outlier detection that has not been fully explored yet but has enormous potential. Detecting botnets is a critical need for securing one's network and the Internet at large. The problem of botnet detection is still unresolved. Optimization based techniques have emerged as important methods to tackle the problems of efficiency and accuracy in data mining.

Since the particle swarm optimization method is a global optimization algorithm based on iterative computing, it can find the global optimal lane model by simulating the food finding way of fish school or insects under the mutual cooperation of all particles.

Such an algorithm is an iterative operation optimization algorithm that further refines the.

II. STATE OF ART

There are number of outlier detection algorithms are exist today that can be used for detecting various outliers presented in web usage data .But there is no particular techniques currently designed for web bot detection. These techniques are classified in to different number of categories. They are

- Statistical based Outlier Detection
- Clustering based Outlier Detection
- Distance based Outlier Detection
- Density based Outlier Detection
- Machine learning & Computational Intelligence

a) Statistical Based Outlier Detection

Statistical outlier detection methods rely on the statistical approaches .That assume a distribution or probability model to fit the given dataset. In this approach outliers are points that have a low probability to be generated by the overall distribution Statistical outlier detection technique is also known as parametric approach. This technique is formulated by using the distribution of data point available for processing. Detection model is formulated to fit the data with reference to distribution of data.

The statistical outlier detection can be divided in to two

- *The parametric methods*
- *The non-parametric methods.*

Parametric statistical outlier detection methods explicitly assume the probabilistic or distribution model(s) for the given data set.

The outlier detection techniques in this category do not make any assumptions about the statistical distribution of the data. Statistical outlier detection methods typically take two stages for detecting outliers.

Training Stage

Testing Stage

In Training Stage it will train the input data set. The trained data set is tested in testing phase. There will different kind of clustering algorithms in statistical method.

Advantages

- They are mathematically justified.
- If a probabilistic model is given, the methods are very efficient and it is possible to reveal the meaning of the outliers found.
- It possible to detect outliers without storing the original datasets that are usually of large sizes.

Disadvantages

- They are unsuitable even for moderate multi-dimensional data sets.
- Limited applications
- The quality of results cannot be guaranteed.
- It is not easy to characterize.
- They are typically not applied in a multi-dimensional scenario.

b) Clustering Based Outlier Detection

If the number of normal attribute is more than abnormal behavior attribute then we use cluster based outlier detection .This technique provides more positive result. This approach is used in those situation when large and dense cluster have normal data and data which does not belong to any cluster or small cluster (low dense cluster) are consider as outlier. In cluster based approach normal data records belong to large and dense clusters, while outliers do not belong to any of the clusters or form very small clusters

Advantages

- Reduce the size of database that will reduces computation time.
- To each cluster user can give certain radius to find outliers.
- Clustering based techniques can operate in an unsupervised mode.
- Such techniques can often be adapted to other complex data types by simply plugging in a clustering algorithm that can handle the particular data type.
- The testing phase for clustering based techniques is fast.

Disadvantages

- Highly dependent on the effectiveness of clustering algorithm.
- The computational complexity for clustering the data is often a bottleneck.
- They are effective only when the anomalies do not form significant clusters among themselves.

I. LOF : Identifying Density-Based Local Outliers

The LOF method is introduce for finding outliers in a multidimensional dataset. LOF will introduce for each object in the dataset. It will indicate the degree of outlier-ness, related to density-based clustering. For each object assign an outlier factor, which is the degree the object is being outlying. The LOF efficiently select the exact isolation degree. The Local Outlier Factor (LOF) is based on a local density. The locality is based on K nearest neighbor this distance is used for estimate the density. By computing the density it can find the regions which have similar density. Points which have a lower density than their neighbors. These are considered to be outliers. There are two parameters that define the notion of density:

• A parameter minpts specifying a minimum number of objects;

• A parameter specifying a volume.

For operating the clustering algorithms these two parameters determine a density threshold. The Objects or regions are connected if their neighborhood densities exceed the given density threshold. It is important to compare the densities of different sets of objects, to detect density based outliers.

Disadvantages

• Local Outlier Probability (LoOP) is a method derived from LOF but using inexpensive local statistics to become less sensitive.

• The computation of LOF value for every data object requires a large number of k-nearest neighbors search.

• Computationally expensive.

Advantages

• Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set.

• The geometric intuition of LOF is only applicable to low-dimensional vector spaces, the algorithm can be applied in any context a dissimilarity function can be defined.

• The LOF family of methods can be easily generalized.

• Can applied to various other problems, such as detecting outliers in geographic data, video streams or authorship networks

II. MINING TOP-N LOCAL OUTLIERS IN LARGE DATABASE[3]

The LOF concept is very useful. The value of LOF in each data require number of k-nearest neighbor search which is computationally expensive. It is necessary to the users to find n most outstanding local outliers. According to LOF the top-n data objects are most likely to be local outliers. If The pruning is not properly done then the top-n outliers results the same amount of computation in finding LOF for all objects. To compress the data it introduce the concept of "Micro clusters". For efficiently handle the overlapping micro clusters it introduce a cut-plane solution. The Micro cluster

$$mc\ (n, c, r)$$

Are introduce to represent the summarized form of group of data .The mean center will be

$$c = \frac{\sum_{i=1}^{n} p_i}{n} \quad (1)$$

III. FindOut: Finding Outliers in Very Large Datasets[13]

The main idea in FindOut is to remove the clusters from the original data and then identify the outliers. FindOut can identify various percentage of outliers from the large data set. Through a unified approach we can combine both the clustering and outliers detection. Experimental results on very large datasets are presented which show the efficiency and effectiveness of the proposed approach. FindOut is works based on the wavelet transform. The byproduct of Wave Cluster is FindOut. It will remove the outliers through this way. It can combine both clustering and outlier detection in a unified approach.WaveCluster which partitions the feature space into cells to generate a quantized space then apply wavelet transform on that quantized space.

After finding the clusters in the transformed space it assigns labels to the cells according to the cluster that they belong to Finally it assigns labels to the objects in the cells based on the cluster label of each cel.The main advantage of wave Cluster is we can detect or identify arbitrary shape clusters such as convex concave or nested clusters. The wavelet transform has multi resolution property, using this wavelet can detect clusters at different degrees. Which does not require the exact number of clusters as input. Wavelet transform can removes the noise from the feature space by taking the merits of Iters in wavelet transform which perform without requiring extra processing time. It uses low pass filter with capability of removing noises.

Advantages

• It is a very fast and efficient method for very large databases.

• Using parallel processing so it speed up processing.
• Take the advantage of low pass filter to remove outliers.

• Can predict future data set.

Disadvantages

• Need higher amount of memory.

• The result will depend on the value of the threshold value.

• Highly cost-effective.

IV. FP-OUTLIER: FREQUENT PATTERNBASED OUTLIER DETECTION

Method for detecting outliers by discovering frequent patterns. In this method we consider outliers are as the data transactions, which contain less frequent patterns in their item sets .For detecting the outlier transactions it define a measure called FPOF (Frequent Pattern Outlier Factor), for this purpose propose the FindFPOF algorithm to discover outliers. If data object contains more frequent patterns, it means that this data object is unlikely to be an outlier because it possesses the "common features" of the dataset. Those infrequent patterns that are contained in few data objects can be used as descriptions of outliers. In this approach the outlier is detect based on distance of points in the full dimensional space. The user will define threshold minisupport,it will find all itemset which equal to or greater than the give threshold minisupport. FPOF-Frequent Pattern Outlier Factor Let be a database containing set of transactions with items I .The threshold minisupport. For each transaction t , FPOF of t.

Advantages

• Efficient and easy to implement

Disadvantages

• Not appropriate for discovering outliers in a high dimensional space.

• The outlier factor of each data object is determined only by the projection with the lowest density of data, without considering the effect of other projections.

• Algorithm has a high computational cost.

• Computationally intensive.

• Causes the abnormality.

V. TWO-PHASE CLUSTERING PROCESS FOR OUTLIER DETECTION

Firstly Partition the entire data set in to several clusters, these clusters may be outliers or non-outliers. If points and they may not be similar in such case it may split in to different clusters. Such partitioning of data points may reduce the time complexity for finding the outliers. If different clusters contained similar points it will be merged. It contains two phases. The first phase use a modified k-means algorithm i.e. using a heuristic. If a new input pattern is far enough away from all clusters centers, then assign it as new cluster center. If data points are in same cluster the diameter of each cluster may be reduced, may be mostly likely all outliers or all non-outliers. In second phase for finding outliers introduce an Outlier Finding Process (OFP), which obtained from the clustered result in Phase 1.

In process it construct a minimum spanning tree (MST) for these clusters constructs. To remove the longest edge of the MST tree from the forest and replace the original tree with two newly generated sub trees. When the cluster is small the tree with less number of nodes are select and regarded as outlier. The use of many classification method make it to implementation become difficult. Also the classification task also more difficult due to the use of various clustering algorithms.

Advantages

• Robust method for outlier detection.

Disadvantages

- Noise is present, so bad effects on results.
- The use of number of clustering method make the classification become difficult.
- Difficult to implement.
- Time consuming and highly expensive implementation.

VI. PARTICLE SWARM OPTIMIZATION FOR OUTLIER DETECTION

It is a cluster based approach, In this approach outlier is detected based on the distance to the centroids. The approach is commonly used for outliers detection techniques also validate the detected outliers. This is based on creating an unbalance distribution of different classes of data and makes one class as very Rare class. The rare class data is considered as group of outliers. From application to application the value of outlier distance threshold of outlierness may varied. The distance value evolves for each cluster also differently. HPSO-clustering uses a partitional approach to generate a hierarchy of clusters. Each level of the hierarchy is treated as a generation of the swarm. The initial generation consists on the entire swarm. Each particle of the swarm represents a centroid of a cluster. The swarm is then evolved towards a single cluster by merging two clusters of the swarm in each successive generation.

Advantages

- High Scalability
- Less complexity.

III. METHODOLOGY

For this work we use Hierarchical Particle Swarm Optimization (HPSO) algorithm for the efficient clustering purpose. Use this HPSO algorithm for perform agglomerative manner clustering. Firstly partition the dataset in to tiny clusters and perform merging of smaller clusters. The merging form a hierarchy clusters. There are two modules included in this work.

- Clustering Module
- Outlier Detection Module

1) *HPSO-Clustering*

HPSO clustering initially partition the data set in to tiny clusters and then merge them in an agglomerative manner. The merging performed based on the learning of particle swarm optimization. For moving to the centroids of the clusters to better positions use cognitive and self-organizing components of the swarm. Particle move from one position to another using the below equation.

The new position of particle is $X_i(t+1)$ and the current position is X_i . From the cognitive and self-organizing components of the swarm we get new velocity $Vel_i(t+1)$ for calculating the new velocity of particle.

$$Vel_i(t+1) = \omega \times Vel_i(t) + q_1 r_1 \left(pBest_i(t) - X_i(t)\right) + q_2 r_2 \left(Y_i(t) - X_i(t)\right)\big) \quad (2)$$

Here $pBest_i(t) - X_i(t)$ is the cognitive learning component and $Y_i(t) - X_i(t)$ is self-organizing component of the swarm. The cognitive component means the learning of particle from its own experience. The particles best position termed *pBest* .This variable need to update each updating to get the better position of the particle. In each iteration the swarm moves to better position and merge smaller clusters to nearest higher cluster.

Here $pBest_i(t) - X_i(t)$ is the cognitive learning component and $Y_i(t) - X_i(t)$ is self-organizing component of the swarm. The cognitive component means the learning of particle from its own experience. The particles best position termed *pBest* .This variable need to update each updating to get the better position of the particle. In each iteration the swarm moves to better position and merge smaller clusters to nearest higher cluster.

2) *HPSO- Outlier Detection*

When merging the smaller clusters it need to calculate a distance threshold to identify whether a particle cluster is genuine or not. We calculate this distance threshold relate to the configuration of data and average intra cluster distance and maximum intra-cluster distance. Average based intra-cluster distance. In this cluster we calculate the distance by using the below formula.

A. *Average intra-cluster distance.*

For calculating the average intra cluster distance we use the equation below.

$$ThreshDist(X_i) = \frac{D_t}{k} \times \sqrt{\sum_{j=1}^{k} (Y_j - X_i)^2}$$

Based on the threshold distance value we will implement the dataset.

B. *Maximum intra cluster distance*

$$ThreshDist(X_i) = D_t \times argMax_{i=0}^{n} \left\{ \sqrt{\sum_{j=1}^{k} (Y_j - X_i)^2} \right\}$$

REFERENCES

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, SIGMOD '00. NewYork, NY, USA: ACM, 2000, pp. 93–104.

[2] L. Duan, L. Xu, Y. Liu, and J. Lee, "Cluster-based outlier detection," Annals of Operations Research, vol. 168, no. 1, pp. 151–168, Apr. 2009.

[3] W. Jin, A. K. H. Tung, and J. Han, "Mining top-n local outliers in large databases," in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '01. New York, NY, USA: ACM, 2001, pp. 293–298.

[4] Z. He, X. Xu, J. Z. Huang, and S. Deng, "Fp-outlier: Frequent pattern based outlier detection,"

[5] Comput. Sci. Inf. Syst., vol. 2, no. 1, pp. 103–118, 2005.

[6] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 145–160, 2006.

[7] S.Ramaswamy,R.Rastogi,andK.Shim,"Efficient algorithms for mining outliers from large data sets," in SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM, 2000, pp. 427–438.

[8] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03. New York, NY, USA: ACM, 2003, pp. 29–38.

[9] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," in Proceedings of the 15th International Conference on Data Engineering, ICDE '99. Washington, DC,USA: IEEE Computer Society, 1999, pp. 512–521.

[10] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," IEEE Transactions on Knowledge and Data Engineering, vol. 14, pp. 1003–1016, 2002

[11] L. Kaufman and P. J. Rousseau, Clustering Large Applications (Program CLARA). John Wiley & Sons, Inc., 2008, pp. 126–163.

[12] M. F. Jaing, S. S. Tseng, and C. M. Su, "Two-phase clustering process for outliers detection," Pattern Recogn. Lett. vol. 22, pp. 691–700, May 2001.

[13] L. Duan, L. Xu, F. Guo, J. Lee, and B. Yan, "A local-density based spatial clustering algorithm with noise," Inf. Syst., vol. 32, pp. 978–986, November 2007.

[14] M.Ankerst,M.M.Breunig,H.-P.Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," SIGMOD Rec., vol. 28, pp. 49– 60, June 1999.

[15] S.Alam, G.Dobbie, P.Riddle, and M.A. Naeem, "A swarm intelligence based clustering approach for outlier detection," in IEEE Congress on Evolutionary Computation (CEC), 2010. IEEE, 2010, pp. 1–7.

[16] D.Yu,G.Sheikholeslami,andA.Zhang,"Findout;: Finding outliers in very large datasets," Knowledge and Information Systems, vol. 4, pp. 387–412, 2002.

[17] C. Aggarwal and S. Yu, "An effective and efficient algorithm for high dimensional outlier detection," The VLDB Journal, vol. 14, pp. 211–221, April 2005.

[18] G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu, "A comparative study of rnn for outlier detection in data mining," in Proceedings of the 2002 IEEE International Conference on Data Mining,ICDM '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 709–712.

[19] A. W. Mohemmed, M. Zhang,and W. N. Browne, "Particle swarm optimization for outlier detection," in GECCO, 2010, pp. 83–84.

[20] Hawkins, H.He, G.J.Williams, andR.A. Baxter, "Outlier detection using replicator neural networks," in Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2000. London, UK: Springer-Verlag, 2002, pp. 170–180.

[21] S.Alam, G.Dobbie, P.Riddle, andM.A.Naeem, "Particle swarm optimization based hierarchical agglomerative clustering," IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 64–68, 2010

[22] S.Alam,G.Dobbie,Y.Koh,andP.Riddle, "Clustering heterogeneous web usage data using hierarchical particle swarm optimization," in IEEE Symposium on Swarm Intelligence (SIS), 2013, 2013, pp. 147–154.

[23] S. Alam, G. Dobbie, and P. Riddle, "Exploiting swarm behavior of simple agents for clustering web users session data," in Data Mining and Multi-agent Integration. Springer, 2009, pp. 61–75.

[24] S. Alam, "Intelligent web usage clustering based recommender system," in Proceedings of the fifth ACM conference on Recommender systems. ACM, 2011, pp. 367–370.