Journal of Network and Information Security
Volume 05 Issue 01, 2017
ISSN.: 2321-6859

Optimized Outlier Based Web Bot Detection

R. Peter[1], D. Divya[2]
[1]M.Tech Scholar, Department Computer Science & Engineering, Adi Shankara College of Engineering & Technology Ernakulam, Kerala, India. Email: respapeter@gmail.com
[2]Assistant Professor, Department Computer Science & Engineering, Adi Shankara College of Engineering & Technology, Ernakulam, Kerala, India. Email: a.divya.d@gmail.com

Abstract: By the turn of century, the use of computers and accessing internet were rapidly increases. As the increasing the network access it increases the network attacks also. The nature of attacks may vary in each day. Today's trends of attacks are web bots. Web bots can be used for both useful and destructive purposes. Now a day's attackers use bot nets for malicious intents. Bots are basically a computer program that surf multiple websites without the intention of the user to perform variety of tasks. If any web bots were present in network it may distort the analysis process which leads to incorrect pattern and cause wrong decision making. The web bots requests were different from genuine request. So it can consider web bots are example of outliers and detect them using outlier detection methods. In this project use Swarm Intelligent (SI) based technique called Particle Swarm Optimization technique (PSO) for detect outliers or web bots. The efficiency of PSO algorithm depends on its parameters. For improving the efficiency of PSO algorithm it need some changes in its parameters. So for improving the efficiency of outlier detection optimization based HPSO (Hierarchical Particle Swarm Optimization) algorithm were used.

Keywords: Clustering, Optimization, Outlier, PSO.

I. Introduction

Data mining is search large stores of data automatically to find patterns and trends that go beyond simple analysis process. To segment data and evaluate the probability of future events data mining uses sophisticated large algorithms. One of the major research topics in Knowledge Discovery in Data (KDD) is Outlier Detection. Outliers can be defined as data objects which deviate from other data objects.

The outliers are formed because of any experiment errors or any measurement changes. If detecting outliers in data mining it improve the data processing task more fruitful by removing the abnormal or error data from the dataset (M. M. Breunig, 2000), (L. Duan, 2009). The application of outlier detection method can be used in different environments such as Fraud Detection (Credit card, telecommunications, criminal activity in e-Commerce) Customized Marketing (high/low income buying habits), Medical Treatments (unusual responses to various drugs), Analysis of performance statistics (professional athletes), Weather Prediction Financial Applications (loan approval, stock tracking). Detection of web bots use the application of outlier detection. Outlier detection mainly came in data mining tasks.

Web bots can be used for both useful and destructive purposes. Now a day's attackers use bot nets for malicious intents. Bots are basically a computer program that surf multiple websites without the intention of the user to perform variety of tasks. If any web bots were present in network it may distort the analysis process which leads to incorrect pattern which leads to wrong decision making. The web bots requests were different from normal different. The attackers will send spam mail to practicing click-fraud it may deviate the users to un wanted pages and they cause attacks without the users willing. For web bot detection process we can use optimization based technique

There are several algorithms are in data mining used for the outlier detection, If use data mining algorithms for detecting web bots or malware it may not be effective. For effectively detecting outliers can use optimization based algorithms like PSO (Particle Swarm Optimization). PSO is one of the optimization techniques in today's technology. PSO is stochastic population based optimization technique. Kennedy and Eberth were introducing PSO in 1995. The PSO algorithm belongs to Swarm Intelligence (SI) computation technique and which inspired from social behavior of birds flocking or fish schooling.

The Computation of PSO algorithm significantly affect by the parameters. It is possible to improve the efficiency of PSO algorithm for Outlier or web bot detection by making changes in its parameter setting through this way it exposes desirable computational behavior with some settings. So the important feature in PSO is that the different parameter settings which exposes different computational behaviors. That is based on the parameter settings the PSO algorithm may work efficiently with some parameters in other case it may work undesirably with some other settings. That's the reason to say that PSO is Parameter-Sensitive algorithm.

Outlier

An outlier in a dataset can be define as a measurement which deviates from the other values. The most established definition

for outlier given by (Hawkins, 1980) as "A judgment which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Another commentary for outlier is obsessed by (Barnett and Lewis, 1994) and defines an outlier as "an measurement (or subset of measurements) which looks to be conflict with the rest of data set". Fig. 1 Illustrate the idea of outlier .

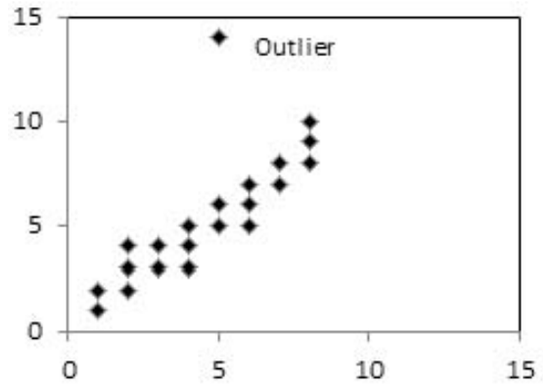

Fig. 1: Illustrate Outlier Data from a Given Data Set

An outlier may be due to inconsistency in the measurement or it may caused by any experimental error. Outlier points can therefore point out erroneous data points, fault procedures. Suppose a considerable datasets will contain modest number of outliers.

II. Particle Swarm Optimization

There are several algorithms are in data mining used for the outlier detection, If use data mining algorithms for detecting web bots or malware it may not be effective. For effectively detecting outliers can use optimization based algorithms like PSO. PSO is one of the optimization techniques in today's technology. PSO is stochastic population based optimization technique. Kennedy and Eberth were introducing PSO in 1995. PSO algorithm will solve discrete and continuous optimization problems. The PSO algorithm belongs to Swarm Intelligence (SI) computation technique and which inspired from social behavior of birds flocking or fish schooling.

To solve optimization problem the concept of PSO algorithm obtained from animals behavior. The population is called swarm and the members in population are called its particles. The optimization problem Starting with a randomly initialization of the particles in population and moving in randomly chosen directions, each particle goes through the searching space and keep the track of best previous positions of itself and its neighbors. Particles of a swarm communicate good positions to each other as well as dynamically adjust their own position and velocity derived from the best position of all particles. The next iteration starts with all particles have been moved.

PSO make is an optimization based algorithm there were no any other optimization algorithms were not used for web bot or also for outlier detection. In data mining for outlier detection many algorithms were used but they are not any optimization based. So for efficient and also to get optimal result use optimization based algorithm such as HPSO algorithm. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position but, is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions. In the problem space each particle keeps track of its coordinates which are related with best fitness it has obtained earlier. The fitness value will be store in each iteration. This value is known as *pbes t*. When another best value called lbest value obtained by the particle swarm optimizer finding *best* value with its neighbors of the particle. When a particle takes all the population as its topological neighbors, the best value is a global best and is called *gbest.* That is in each iteration they search better position using three basic learning components. they are:

- Cognitive Component (*pBest*)
- Social Component (*gBes*t)
- Self-Organizing Component

Parameter Selection in PSO

The Computation of PSO algorithm significantly affect by the parameters. It is possible to improve the efficiency of PSO algorithm for Outlier or web bot detection by making changes in its parameter setting through this way it exposes desirable computational behavior with some settings. So the important feature in PSO is that the different parameter settings which exposes different computational behaviors. That is based on the parameter settings the PSO algorithm may work efficiently with some parameters in other case it may work undesirably with some other settings. That's the reason to say that PSO is Parameter-Sensitive algorithm.

The imperative issue in PSO is that distinctive parameter settings can prompt altogether unmistakable computational practices. It may work effectively with some parameters, while it may work undesirably with some different parameters. Despite the fact that the sensitivities of different PSO variations as for their parameters are diverse, PSO is completely considered as a parameter-delicate calculation. In this way, the push to discover ideal PSO parameters that lead to the best computational conduct is of critical significance and much research exertion has been put on it. For the most part, for setting PSO parameters, there exists taking after systems in writing:

- Utilizing PSO variations with altered parameters. This is the most generally utilized system as a part of PSO writing. For this situation, parameter qualities are either set as broadly acknowledged qualities from writing or set on experimentation premise.

- Utilizing without parameter PSO variations which don't require any parameter setting procedure.
- Utilizing composite PSO. In this methodology, a heuristic calculation is utilized to discover ideal PSO parameters, however it has from time to time been utilized as a part of PSO writing. The huge issue of this system is that it adds fundamentally to the intricacy of the issue.
- Utilizing PSO variations with dynamic or versatile parameters.

III. Related Works

An important research problem in data mining is an Outlier detection which aims to recognize objects that are dissimilar with other existing objects in data set. Outlier detection, also known as anomaly detection in some literatures, has become the enabling underlying technology for a wide range of practical applications in industry, business, security and engineering, etc. According to the verities in the data used for different domains there were different types of outlier detection method can choose. The selection of the method for detecting outliers are depends on various factors. The selection of outlier detection method mainly cares about the input data type and it also depends on so many other factors such as data type distribution, data availability and resource constraints etc.

A. Density Based Outlier Detection

This strategy contrasts the density around a point and its nearby neighbors densities. The relative density of a point contrasted with its neighbor is registered as an outlier score. Density based outlier detection system utilizes density distribution of data points inside of data set. The thought of density based neighborhood outlier utilizing correlation with density of nearby neighborhood was presented by Breuing et al. This method correlate the density over a point with its regional neighbors densities. The corresponding density of a point compared to its neighbors is measure as an outlier score. Density based outlier detection approach uses density. Compared to distance based approach density based method is higher convoluted mechanisms to model the outlier ness of the data items. By analyzing the central aspect or feature of object in database it can determine the outlier. The nearest neighbors are comparitatively proximate in this approach, then the data point is considered to be normal, diversely it is considered as an outlier.

B. Clustering Based Outlier Detection

Cluster analysis is prevalent unsupervised methods to bunch comparative information occurrences into clusters. Clustering parcels the information into gatherings, in which comparative articles are contained. The expected conduct of outliers is that they either don't have a place with any cluster, or are compelled to have a place a cluster where they are altogether different from different individuals then again fit in with little clusters. If the number of normal attribute is more than abnormal behavior attribute then use cluster based outlier detection method .If an object does not resides to any cluster or there is high gap between the object and its adjacent cluster or it fit to inadequate cluster then that object is considered as an outlier. By applying this process it will achieve more confident result. This method 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.

C. Distance Based Outlier Detection

Distance based outlier detection method judge a point based on the distance(s) to its neighbors. Express distance-based methodologies are based on the well known nearest-neighbor standard. Ng and Knorr propose an all around characterized distance metric to identify outliers. They characterize outlier as the object which is more prominent in distance to its neighbors. The essential algorithm, the settled circle (NL) algorithm, figures the distance between each pair of objects and afterward set as outliers those that are far from most objects. The NL algorithm has quadratic unpredictability, concerning the quantity of objects, making it inadmissible for mining extensive databases, for example, those found in government review information, clinical trials information, and system information.

D. Statistical Based Outlier Detection

Statistical outlier detection utilizes certain sort of statistical distribution and processes the parameters by expecting all information focuses have been created by statistical distribution. In this methodology outliers are focuses that have a low probability to be created by the general distribution Statistical outlier detection strategy is otherwise called parametric methodology. This strategy is detailed by utilizing the distribution of information point accessible for preparing. Detection model is defined to fit the information with reference to distribution of information.

In Table 1.1 show the some popular techniques existed in today.

TABLE 1: LITERATURE SURVEY

Method	Advantage	Disadvantage
Local outlier factor (LOF)	Can applied to various other problems	Less Sensitive. Computationally expensive. Requires a large number of k-nearest neighbors search. Can't detect a cluster of outliers. Less Accurate
Micro cluster based LOF	Eefficiently found top-n outliers in large database. Introduce a cut-plan solution for over-lapping data. Good performance to find the most outstanding local outliers.	Runtime complexity High cost. They compute it for the top n-outliers only. Inefficiency Data must be loaded in to the memory to implement algorithm
FP-Outlier: Frequent Pattern Based Outlier Detection	It is very robust. High performance on low dimension data. Eefficient implementation	Not appropriate for discovering outliers in a high dimensional space. Algorithm has a high computational cost. Computationally intensive. Causes the abnormality.
bin PSO	Less computation time. Can handle both discrete binary and continuous variable.	Need high memory Parameters will effects its working

There are number of outlier detection methods are used in data mining but for detecting web bots efficiently use optimized based detection method that is PSO based outlier detection. So propose Hierarchical clustering based PSO algorithm.

IV. 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

A. Clustering Module

For the clustering process we use hierarchical 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 centroid of the clusters to better positions use cognitive and self-organizing components of the swarm.

$$Vel_{i(t+1)} = \omega \times Vel_{i(t)} + q_1 r_1 (pBest_i(t) - X_i(t) + q_2 r_2 (Y_i(t) - X_i(t)))$$

The new position of particle is X_i $(t + 1)$ and the current position is X_i (t). 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. Here $p\text{Best}_i$ $(t) - X_i(t)$ 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.

C. Outlier Detection Module

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.

The Outlier detection decision done based on the average and maximum intra cluster distance based calculations. The Calculation mainly done based on following equations.

C. Average Intra-Cluster Distance

For calculating the average intra cluster distance we use the equation

$$\text{Thresh Dist}\left(X_i\right) = \frac{D_t}{k} \times \sqrt{\sum_{j=1}^{k} (Y_j - X_i)^2}$$

Based on the threshold distance value will implement the dataset. It chose different values of threshold distance as a function of average intra-cluster distance and record the number of outliers found in the log. At larger values offewer web bots have been detected and vice versa. The right most part represents the large number of outliers found at smaller values of *Dt*.

D. Maximum Intra Cluster Distance

$$\text{Thresh Dist}\left(X_i\right) = D_t \times \arg\,\text{Max}_{i=0}^{n} \left\{ \sqrt{\sum_{j=1}^{k} (Y_j - X_i)^2} \right\}$$

Maximum intra-cluster distance as a threshold distance calculated using equation and extracted the suspected web bots. At smaller values of Dt fewer web bots have been detected and vice versa.

E. Intersection of Maximum and Average Intra-Cluster Distance

The results of overlap when the intersection of avg. and max. intra-cluster distance was used to detect web bots. The overlap increases when the number of suspected web bots decreases because of the smaller number of true negatives.

V. Conclusion

To implement this system use one of the popular Swarm Intelligence (SI) based techniques called Particle Swarm Optimization (PSO) to detect Outliers. Here web bots are considered as web bots. So it need to detect web bot among genuine user requests. Use Particle Swarm Optimization (PSO) based clustering algorithm, Hierarchical Particle Swarm Optimization based clustering (HPSO-clustering) to cluster the web-usage data and detect the abnormal behavior caused by the web bots. However it is still not develop the issue of automating the parameter selection process in web bot detection. In proposed scenario going to deal with the tuning parameters in PSO algorithm for selecting the process. Different strategies for setting each PSO parameter are explained and analyzed deeply. After obtained optimized result we can stop the parameters based on the decision maker. Finally it provides high accuracy, fast convergence results. It takes less computational time to execute this process. Introduced a Particle Swarm Optimization (PSO) based clustering technique called Hierarchical Particle Swarm Optimization based clustering (HPSO-clustering) to detect web bots in the web usage data.

VI. Future Work

There are still some problems and inefficiencies can detect in the existing work. This technique still include some issues in performance, generalization and extendibility. It need to improve extendibility performance and generalization of this system, which would useful for further investigation. Need to find how to obtaining these efficiencies in to the system. In future work mainly focus to the efficiency improving process. The performance of PSO algorithms is a parameter dependent, that is its always performs based on its parameter. Scaling the outlier detection to high dimensional data is another future research direction.

The future research work contains the generalization of these techniques, automating the parameter selection process, and application of HPSO-clustering based outlier detection method in different domains other than with web usage data. Introduce new parameters for this achievement. Inertia weight is used to control the velocity in this scenario. It is the process of providing balance between exploration and exploitation process in general. The Inertia Weight determines the contribution rate of a particle's previous velocity to its velocity at the current time step. A large Inertia Weight facilitates a global search while a small Inertia Weight facilitates a local search. Inertia Weight is based on the function of local best and global best of the particles in each generation. It neither takes a constant value nor a linearly decreasing time-varying value. To overcome the weakness of premature convergence to local minimum, Adaptive Inertia Weight strategy is proposed to improve its searching capability.

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,* New York, NY, USA: ACM, pp. 93-104, 2000.

[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,* New York, NY, USA: ACM, pp. 293-298, 2001.

[4] Z. He, X. Xu, J. Z. Huang, and S. Deng, "Fp-outlier: Frequent pattern based outlier detection," *Computer Science and Information Systems,* vol. 2, no. 1, pp. 103-118, 2005.

[5] 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.

[6] S. Ramaswamy, R. Rastogi, and K. 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,* pp. 427-438, 2000.

[7] 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,* New York, USA: ACM, pp. 29-38, 2003.

[8] 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, pp. 512-521, 1999.

[9] 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.

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

[11] M. F. Jaing, S. S. Tseng, and C. M. Su, "Two-phase clustering process for outliers detection," *Pattern Recognition Letter,* vol. 22, pp. 691-700, May 2001.

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

[13] 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.

[14] 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),* pp. 1-7, 2010.

[15] D. Yu, G. Sheikholeslami, and A. Zhang, "Findout; Finding outliers in very large datasets," *Knowledge and Information Systems,* vol. 4, pp. 387-412, 2002.

[16] 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.

[17] 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,* Washington, DC, USA: IEEE Computer Society, pp. 709-712, 2002.

[18] A. W. Mohemmed, M. Zhang, and W.N. Browne, "Particle swarm optimization for outlier detection," In *GECCO,* 2010, pp. 83-84.

[19] S. Hawkins, H. He, G. J. Williams, and R. 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, pp. 170-180, 2002.

[20] S. Alam, G. Dobbie, P. Riddle, and M. 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.

[21] S. Alam, G. Dobbie, Y. Koh, and P. Riddle, "Clustering heterogeneous web usage data using hierarchical particle swarm optimization," In *IEEE Symposium on Swarm Intelligence (SIS),* pp. 147-154, 2013.

[22] 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, pp. 61-75, 2009.

[23] S. Alam, "Intelligent web usage clustering based recommender system," In *Proceedings of the fifth ACM Conference on Recommender Systems,* pp. 367-370, 2011.