A hybrid approach for anomaly detection using K-means and PSO

Ke-Wei Wang[1, a], Su-Juan Qin[1,b]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China

[a]13141221443@163.com, [b]qsujuan@bupt.edu.cn

Keywords: Anomaly detection, Particle Swarm Optimization, K-Means, Clustering analysis.

Abstract.

The network intrusion detection systems which based on anomaly detection techniques plays an important role in protection network and systems from harmful attacks. With increasing in attacks and the new security challenges, The lower accuracy of anomaly detection method based on cluster analysis network traffic is a big question, In this paper, we proposed a hybrid anomaly detection method by combining the Particle Swarm Optimization(PSO) and K-Means clustering algorithms improving the accuracy. We first preprocess features of data traffic, extract the characteristics of the various categories of attack, and then use parallel PSO calculation, to find the best or a little approximations to optimal clustering initial point. Finally, we perform the K-Means clustering algorithm. Experiment results show the effectiveness of the proposed optimization scheme.

Introduction

In the past few decades, with the development of information technology, there is growing recognition of the importance of network and data security, Network intrusion and attacks are likely to cause serious disasters, therefore, the development of intrusion detection systems to reduce the impact of the attack traffic is the new trend [1].

Anomaly detection methods include supervised learning, semi-supervised learning, and unsupervised learning methods. Supervised-learning method requires to establish behavior patterns, but it is difficult to obtain labeled data sets in the actual network environment, and cannot guarantee the accuracy of the data set markers. The new trend is to use unsupervised learning methods for the not tag data sets , to directly establish a detection model for automatically generated labeled data sets [4].Unsupervised learning or clustering analysis is usually used to solve the abnormal traffic and new attacks [5].K-Means algorithm is one of the most efficient partitioning clustering algorithm. The algorithm is simple and easy to implement, suitable for large data sets, and very highly active. However, it has two major drawbacks: 1).the number of randomly chosen points and the centroid of clusters may lead to different clustering results, 2).K-Means algorithm may contain many local convergence. In some of the previous anomaly detection systems have adopted the K-Means clustering, but the result is not ideal. There are a lot of clusters and local optima, this algorithm run several times will get different results.

In order to overcome these shortcomings of K-Means, We use the Particle Swarm Optimization (PSO) to combine the K-Means algorithm.PSO is a heuristic algorithm, it can be a minimum number of iterations to find the optimal or near-optimal solutions. PSO can find good initial cluster centers by its’ global search capability and then avoid K-Means algorithm falling into local optimal solution.

This paper is organized as follows: the second part is related to work; the fourth and fifth section briefly describes the PSO and K-Means algorithm; part VI describes the idea of this paper; part VII is the result; part VIII is the conclusion; part IX of the relevant literature.

Related work

Tarek F.Ghanem [6] use the genetic algorithm, negative selection algorithm to combine K-Means,improve the detection rate of abnormal network,but after combining algorithm the complexity is too high.Chen [7] by using particle swarm optimization (PSO) and K-Means algorithm combined to categorize web pages,Zhen Kuai et al [8] use of Iris data, but also to reduce the false detection rate of onlyusing the K-Means by combining PSOwith K-Means,Amin Karami [9] for CCN new network, combined PSO with K-Means to improve the detection rate and reduce the false detection rate.

Particle Swarm Optimization (PSO)

PSO is proposed by the Kennedy and Eberhart in 1995. It is based on Swarm Intelligence optimization algorithm. The process of PSO is as follows.

1) *Initialization*: Initialize particle population (population size is N), random position, velocity, and other parameters.
2) *Assessment*: Evaluate the fitness of each particle based on the fitness function
3) *Finding the best individual:* For each particle, compare the current best fitness value with its own history, and if the current adaptation value is higher than the value of the best fitness value, then the best fitness value are updated to the value of the current adaptation values.
4) *Finding the global optimum*: For each particle, compare the current fitness whit the best fitness value of the global, if the current adaptation value is higher than the best fitness values of the global, then the best fitness values of the global are updated to the value of the current adaptation values.
5) *Updating*: update the velocity and position of each particle according to the formula.
6) *Cyclic iteration*: If the stop condition of the algorithm is not reached, returns to the step of the Assessment. Usually algorithm runs until the maximum number of iterations is reached or the incremental of the optimum fitness value less than a given threshold.

K-Means

It is based on Euclidean distance between two data points (x and y) calculation of N objects and in n-dimensional space. The process is as follows.

1) Randomly select *k* data points for each cluster most initial centroid.
2) Repeat.
3) Calculate the distance between each data point with each cluster centroid. Dividing each data point to its nearest cluster, and re-calculated for each new cluster centroid
4) Until the cluster centroid is unchanged.

Methodology

In this section, a new anomaly detect approach is proposed based on using K-Means and Particle Swarm Optimization(PSO). The main idea is based on using PSO compensate for K-Means's shortcomings and Using parallel PSO computing to find the best initial point.

Data Normalization.

In the network, data traffic generally consist of some features. We should take different measures of data pre-processing and distance metric for different types of characteristics.

For enumerated disorder features, we use the index of the array to convert, and use integer numbers to represent the characteristics of the different values (including non-numeric types). for the type of {0,1}, it doesn’t need data pre-processing.

For the characteristics of non {0,1}, using the formula of (X-mean)/std, when calculated separately for each property. Each property of the data minus the mean and in addition to its variance. After the data processing, the mean for each attribute is gathered in the vicinity of 0, and the variance is 1. So this will take convenient to the following algorithm.

Group And Reduce Dimensionality.

Table 1: the features are extracted from the 41 characteristics.

Index	Name	Index	Name	Index	Name
2	protocol_type	18	num_shells	30	diff_srv_rate
3	service	19	num_access_files	31	srv_diff_host_rate
4	flag	22	is_guest_login	32	dst_host_count
5	src_bytes	23	count	33	dst_host_srv_count
6	dst_bytes	24	srv_count	35	dst_host_diff_srv_rate
12	logged_in	25	serror_rate	36	dst_host_same_src_port_rate
14	root_shell	27	rerror_rate	38	dst_host_serror_rate
16	num_root	28	srv_rerror_rate	40	dst_host_rerror_rate
17	num_file_creations	29	same_srv_rate	41	dst_host_srv_rerror_rate

Grouping the data which extracted features according to Several categories of attack to, to reach the dimension reduction, so make PSO process to be more pertinent to find the best spot. According to the characteristics of various types, the original data is converted into many different dimensions of data (Their data size won’t be changed, but their dimensions will be reduced), and then a plurality of data parallelly computing PSO to draw into targeted optimum point. The output point of PSO will input to the K-Means. Firstly, we reduce dimension according to all kinds of characteristics of attack in the network. For example, the attacks of U2R and R2L are different from DOS attacks that have frequent sequential patterns in the data record, so the times of remote login failures, the times of using root, and the times of using the shell command that different from the feature of Dos attack. So we extract features to reduce dimension according to the categories of attack. The following, we will introduce the extracted characteristics depending on 41 characteristics:

So, the features will be extracted as follows:

Dos. Feature extraction [2,3,4,5,6,23,24,27,28,29,31,32,33,34,35]

R2L. Feature extraction [2,3,4,6,12,16,17,18,19,22,23,24,32]

U2R. Feature extraction [2,3,4,13,14,17,18,23,24,27,28,29,32,33,,40,41]

PROBING. Feature extraction [2,3,4,23,24,25,27,28,29,30,32,33,34,35,36,38,40,41]

This step is to reduce the dimension, the original data is n * m m-dimensional matrix, n is the amount of data size, m is the dimensional (m less than 41), for example, Dos is n * 15, because this class’s dimensional will be reduced to 15, other class is the same method, so that we can get four groups of n * k data, k represents all kinds of different dimensions. We should consideration of the normal flow, and normal traffic is arbitrary, so we need to add a set of data without dimensionality reduction, so that the data can be divided into five groups. Each group has the same data, except for the different dimension of the data in each group.

Parallel PSO.

Then each set of data in parallel computing PSO. As long as the best value or the best approximation of each data point, therefore, five groups of data can be parallel computing PSO, This consideration is to improve the efficiency of the algorithm, and to minimize the complexity. And each set of data do not affect each other in the process of computing, the result is reliable.

K-Means Computing.

This step is to use the five best points to perform K-Means Clustering. So the proposed algorithmic process:

Input data set.

Output clustering results.

Step 1 process data normalization.

Step 2 packet processing data dimensionality reduction.

Step 3 PSO parallel processing packet data obtained five best point.

Step 4 K-Means clustering use five points to give the best clustering results.

Experimental results and Analysis

In this paper, we use the data of KDD CUP 99 improved data set, and extracte three datasets which are 3,000 data (D1), 3 million data (D2) and 30 million data (D3) to do experiment. And we use the detection rate (DR) and false positive rate (FPR) to assess and comparative results of each algorithm.

$$DR = \frac{\mathrm{True}Positive}{\mathrm{True}Positive + FalseNegative} \tag{1}$$

$$\mathrm{FPR} = \frac{FalsePositive}{FalsePositive + TrueNegative} \tag{2}$$

True negative and true positive detection represent system accurately identified the normal and abnormal traffic flow. False positive shows a normal traffic identified as abnormal traffic, false negative shows abnormal traffic that will be recognized as normal traffic.

Experiment I.

Table 2: the DR and FPR of the five algorithms.

	D1		D2		D3	
	DR	FPR	DR	FPR	DR	FPR
k	61.4	14.7	65.3	15.9	72.4	12.5
kb	75.8	10.3	80.7	9.2	86.3	9.8
dpk	86.2	12.2	88.2	10.5	90.2	8.4
npk	87.2	13.6	88.6	11.5	89.3	10.2
pk	93.3	6.5	92.4	7.3	95.5	4.4

We take this paper’s algorithm (pk), general K-Means algorithm (k), the algorithm of taking the best of running a hundred times of the general circulation K-Means (kb), the algorithm of combining PSO and K-Means algorithm (dpk) with blocking data and the algorithm of combining the PSO and K-Means algorithm (npk) with not processing data to do experiment, Respectively compared the detection rate and false positive rate. Through this experiment can be seen in the detection rate of the proposed algorithm to improve a lot, the false detection rate is as low as possible, so the method in this paper is feasible and effective.

Experiment II.

We take this paper's algorithm (pk), n is the number of PSO algorithm iterating times(n = 10, 20,…, 100).Through experiments can verify the situation of PSO will be premature, and a large amount of data, a higher detection rate, which better reflect the advantages of PSO. Here are the results and comparison chart of each group of data sets:

Table 3: the values of this algorithm with different iteration times.

	10	20	30	40	50	60	70	80	90	100
D1	87.2	93.3	88.9	84.7	83.1	84.5	85.3	86.4	84.3	80.7
D2	85.7	87.4	86.9	88	92.4	87.6	84.9	83.1	82.5	84.1
D3	86.3	83.2	85.7	87.9	86.8	89.4	90.6	89.8	95.5	88.1

So we can summarize the trend from the values:

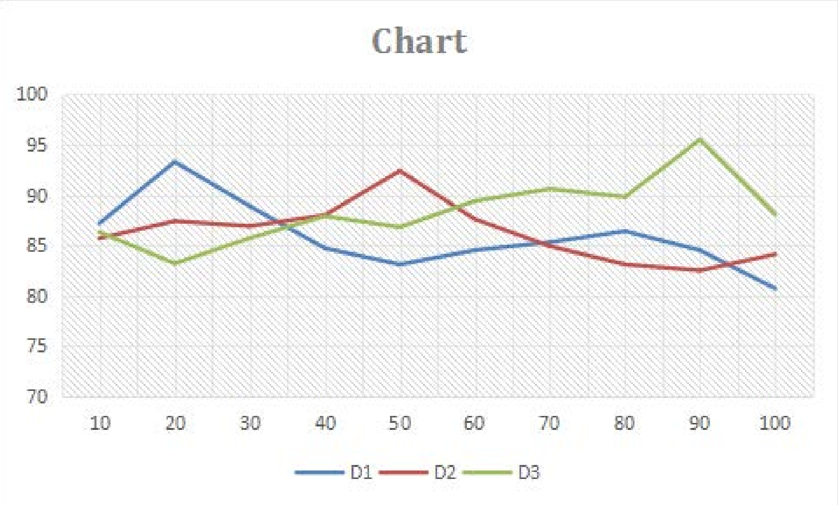

Fig. 1 the trend of this algorithm with different iteration times.

Experiment III.

We take this paper proposed algorithm (pk), the algorithm of combining PSO and K-Means algorithm (dpk) with blocking data and the algorithm of combining the PSO and K-Means algorithm (npk) with not processing data to compare with complexity:

Table 4: the complexity of five algorithms.

	complexity	Time[s]
pk	o(pso)+o（k-means）	798
dpk	o(pso)+o（k-Means）	923
npk	o(pso)+o（k）	1448

From the experimental results, we can learn that the ideas of this paper has a significant effect in the degree of complexity.

Summary

In this paper, we propose efficient algorithms which by reducing dimensionality of data and combine concurrent PSO and K-Means for the network traffic. Compared with several other algorithms in several different data sets, this algorithm have higher detection rate and low false positive rate. this effect which be achieved is inseparable with the combination of PSO algorithm. And on effectiveness for a given period of time have also been optimized by reducing dimension and paralleling PSO will make the complexity of the algorithm to be minimize.

Of course, this article only focuses on the optimization of the clustering results, in the clustering process is quite rough, it should be more detailed in future work, because for some points which are away from the normal cluster did not be considered that whether these are abnormal in this article, so the detection rate there is still room for improvement.

References

[1] Liao HJ, Lin CHR, Lin YC, Tung KY. Intrusion detection system: a comprehensive review. J Netw Comput Appl 2013;36(1):16–24.

[2] Patcha A, Park JM. An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw 2007;51(12):3448–70.

[3] Garcı´a-Teodoro P, Dı´az-Verdejo J, Macia´ -Ferna´ ndez G,Va´ zquez E. Anomaly-based network intrusion detection: techniques, systems and challenges. Comput Secur 2009;28(1–2):18–28.

[4] B. Krawczyk,M.Woźniak, Diversity measures for one-class classifier ensem-bles, Neuro computing 126 (2014)36–44.

[5] G. Corral, E.Armengol, A.Fornells, E.Golobardes, Explanations of unsuper-vised learning clustering applied to data security analysis, Neurocomputing72 (13–15) (2009) 2754–2762.

[6] Tamer f.Bhanem, A hybrid approach for efficient anomaly detection using metaheuristic methods, Journal of Advanced Research(2015) 6, pages 609-619.

[7] J. Chen, Hybrid clustering algorithm based on pso with the multi dimensional asynchronism and stochastic disturbance method, J.Theor.Appl.Inf.Technol.46 (1) (2012) 434–440.

[8] P.Zhenkui, H.Xia, H.Jinfeng, The clustering algorithm based on particle swarm optimization algorithm, in: Proceedings of the International Conference on Intelligent Computation Technology and Automation(ICICTA'08), IEEEComputer Society, Washington,DC, USA, 2008, pp.148–151.

[9] Amin karami, A fuzzy anomaly detection system based on hybrid Pso-KMeans algorithm in content-centric networks, Neurocomputing 149(2015), pages 1253-1269.