2017 IEEE 3rd International Conference on Big Data Security on Cloud

A Novel Unsupervised Anomaly Detection Approach for Intrusion Detection System

Weiwei Chen
State Grid Zhejiang Pingyang Power Supply Co. ltd
Zhejiang, China
Email: chen_weiwei@zj.sgcc.com.cn
Feng Mei
State Grid Zhejiang electric power company
information and telecommunication branch
Hangzhou, China
Email: mei_feng@zj.sgcc.com.cn

Fangang Kong
State Grid Zhejiang Electric Power Company
Zhejiang, China
Email: kong_fangang@zj.sgcc.com.cn
Guiqin Yuan and Bo Li
School of Computer Science and Engineering
Beihang University
Beijing, China
Email:{yuangq, libo}@act.buaa.edu.cn

***Abstract*—Network Anomaly Detection plays an important part in network security. Among the state-of-the-art approaches, unsupervised anomaly detection is effective when dealing with unlabelled data. However, these approaches also suffer from high false positive rate. We observed that different methods have their own defects and advantages. Inspired by this observation, we provide a new ensemble clustering(NEC) method to detect novel anomalies. In our system,we can get higher detection rate and lower false positive rate compared with existed apporaches as verified over NSL-KDD 2009 dataset.**

Keywords*—*Anomaly detection; unsupervised; ensemble clustering

I. Introduction

Intrusion detection system [1] is an important part of any well managed network systems. It include misuse detection system and anomaly detection system[2]. The misuse detection system try to detect anomalies with signatures. The system based on signature [3] is a common way in commercial system. Our experts try to get efficient signatures form anomalies. We can match the testing dataset with the patterns. If a record can find corresponding pattern from the signatures, we will treat the record as an anomaly. On the contrary, we will treat the record as normal. Consequently, if we want the system to keep pace with the anomalies, we have to update the signatures endlessly. Although the system based on signatures is an efficient way to detect anomalies in real life, but it can't detect novel anomalies. In addition, it's expensive to extract corresponding signatures for anomalies. Our system is vulnerable to the novel attacks until our experts manually to expend our signatures. In order to solve the problem, our researchers try to find a new way to solve the problem.

The anomaly detection system needn't that our experts manually update the signatures. The anomaly detection system includes three parts, which is supervised anomaly detection, semi-supervised anomaly detection and unsupervised anomaly detection. The supervised anomaly detection system [4] [5] [6] [7] [8] try to build an model based on the labelled normal records and abnormal records. If a record is more similar to the normal than the anomaly, we will think the record is normal. In the supervised anomaly detection system, we need quality labelled dataset. But in real life, it is difficult to produce quality label for the dataset. In addition, the network is changing all the time, supervised anomaly detection system often shows high false positive rate. The semi-supervised anomaly detection[9][10] try to build a model based on normal data. If an record can't match with the model, it will be treated as an anomaly. The semi-supervised anomaly detection also can build a model based an abnormal data. But anomalies are difficult to get, and we can't get all anomalies. For this reason, the semi-supervised anomaly detection based on anomalies is not common. The semi-supervised anomaly detection also need labelled dataset and often shows high false positive rate.

In order to solve the problem, unsupervised anomaly detection system[11] [12] is an appropriate way to solve the problem and find novel anomalies, which needn't labelled dataset. Moreover, it can be deployed in any systems without change. In [11], researcher try to provide an robust method. The model firstly transform the data into an common high-dimensional feature space. Then it tries to use common unsupervised anomaly detection ways to detect anomalies. But different unsupervised anomaly detection systems have their own defects and advantages. It will get better performance if it can match the data structure and take appropriate metric for distance. There are different ensemble clusterings in[16][17][18][19][20]. And in supervised anomaly detection systems, many researchers have do some effective research on ensemble classification [13][14][15]. Moreover, the adaboost-based algorithm often show better performance than basic classifiers.

The unsupervised anomaly detection system can detect new anomalies without labelled dataset. But the systems often show lower detection rate and higher false positive rate. In order to solve the problem and improve the robustness of our system, we provide an ensemble clustering based on subspace ensemble clustering. We have verified our algorithm on KDD[21]. In this paper, we show an novel ensemble unsupervised anomaly detection system. We have the following contributions:

- We provide an novel ensemble unsupervised anomaly detection system, which shows higher detection rate and lower false positive rate.

978-1-5090-6296-6/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataSecurity.2017.56

- We try to use more meaningful features based on subspace, which can show higher speed.

Our paper is organised as follows. In section 2, we will show the related work. And in the following section, we will introduce the dataset that we used and preprocessing. In section 4, we will introduce our model. And we mainly introduce the results in section 5. In section 6 and section 7, we will introduce the conclusion and references.

II. Related work

From the base of our paper, we mainly present there related areas, which are ensemble clustering and misuse detection.

And as we all known, there no ideal algorithm. Each algorithm has their own advantages and defects. The ensemble algorithms provide a way to combine different algorithms. In [22], the authors summarize different ways to combine the clustering. They product different clustering result from the same data set or supplement part of the same data set, and then try to combine them into a final result. It provides an efficient alternative way to solve the clustering analysis problem.

The supervised anomaly detection, also known as misuse detection. Although misuse detection has its own defects, which are unconquerable. They can't detect the new anomalies and need the labelled data which produced by people or the signature-based system. The original of the data is not reliable. It means that we couldn't sure whether the label is right or not. On the other hand, anomaly detection faced a common problem. Nobody is willing to provide their data sets for research, because of the privacy, economy or safety. Many authors have make great contribute to the misuse detection in recent years. In [23], Christopher Kruegel proposed an algorithm that based on different attributes to build different attacks. One of the models predict the result by a Markov model. In [24], three independent traffic sets was detected by the SVM. Moreover, the model could finish the training with reduced data set. In [25], Yun Wang indicated that multinomail logistic model can be used to solve the problem of multi-classifier.

III. Data Set Collection and Preprocessing

A. Data set

The data set that we used is the KDD CUP data set, which is 21.6 MB of data. It includes 148517 samples, with 71463 anomaly records and 77054 normal records. The data set solves some problems of KDD'99 data set. Firstly, it does not include redundant records in the data set, which can ensure the model will not be biased. More precisely, it include two data types, which are real and symbolic. And it include 41 features, which can be divided into three types:basic features, extracted form TCP/IP features; and content features, derived from packet contents; and traffic features, mainly referred to the same host or some service. In order to get appropriate rate of data set, we randomly choose 70000 normal records and 700 anomaly records, which is an reasonable rate in real life. Moreover, our data set includes 11 kinds of anomalies, such as neptune, warezclient, portsweep and so on.

B. Preprocessing

In the last part, we introduce the data set that we used. Firstly, we need to transform the symbolic into real. The symbolics that in the same row are put into a set, and the symbolics in the original data set are replaced by the index of the symbolic in the set. In addition, different rows used various matric. For example, src_bytes and land, the former stands number that one host send how many bytes to the destination, while the latter stands whether the source IP and ports is the same ad destination. The former values range from 0 to a few thousand, which the latter values only can be 0 and 1. To solve this problem, we use the normalizer to normalize the data set. Fristly, we can use equation 1 and 2 to get the mean value and standard deviation.

$$mean\_value[j] = \frac{1}{N}\sum_{i=1}^{N} instances_i[j] \tag{1}$$

$$std[j] = (\frac{1}{N-1}\sum_{i=1}^{N}(instance_i[j] - mean\_value[j])^2)^{1/2} \tag{2}$$

After that, the instances can be transformed into uniform expression using equation 3.

$$new\_instance[j] = \frac{new\_instance[j] - mean\_value[j]}{std[j]} \tag{3}$$

IV. Algorithm

In this paper, we proposed an unsupervised ensemble clustering to detect anomalies. It can be divided into three parts: feature selection, ensemble clustering and evaluation. Firstly, we transform features into real and get many subspaces of original feature set. Secondly, different models use different subspaces to get clustering labels, then all these clustering labels will be passed to the voting model. The voting model will ensemble all these results and output the final clustering result. Finally, the predicted results will be compared with true labels.

Figure 1 shows details of our unsupervised ensemble clustering model. From [21], NSL-KDD data set is downloaded. Then our model try to choose some records from original NSL-KDD data set for the reason that normal records is far greater than the number of anomalous records. Then the selected subset of original data set is passed to preprocessing model. The preprocessing model transforms all features into real numbers and normalise data set, in order to scale all samples into unit norms. Nextly, the transformed data set are split into different subsets. DBSCAN model can get more effective results in lower dimensional space. In addition, not all features benefits anomaly detection. Then our DBSCAN, One-SVM, Agglomerative Clustering and expectation-maximization(EM) model will split the data set into normal records and anomalous records. All these results are passed to voting model. Voting model try to ensemble all of these results and get an consistent result. Finally, the evaluation component will compare predicted result and true result.

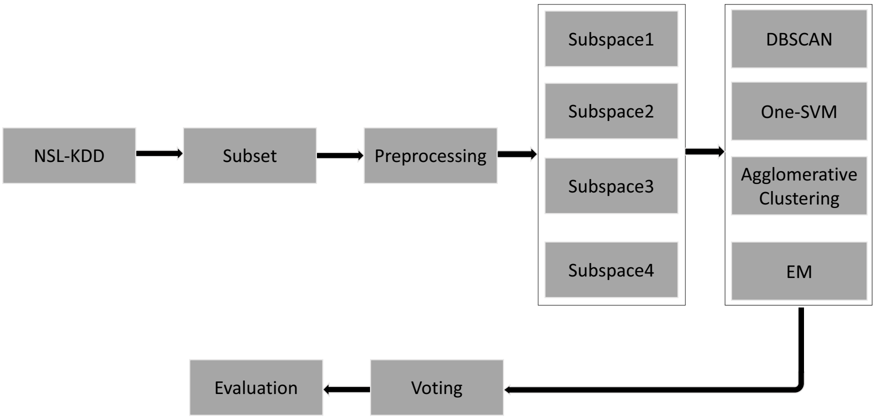

Fig. 1: Architecture of Unsupervised Ensemble Clustering

A. DBSCAN

DBSCAN [26] splits data set into different clusters according to density. If a point includes $MinPts$ points in the d-dimensional ball centered at point p with radius r. $MinPts$ is a real number and stands the number of points. p stands a point in d-dimensional ball. r stands the radius of d-dimensional ball. The DBSCAN algorithm can be down like follows. Firstly, we randomly choose a point p_0 than calculate the number of points($NumPts$) in the d-dimensional ball centered at point p_0 with radius r. If $NumPts$ is larger than $MinPts$, p_0 can be treat as an core point, all these points can be treat as an cluster C_0. Then traversing all point in cluster C_0 to expand the cluster C_0 until C_0 is stable. Then we randomly choose one point from the rest of data set and repeat the previous process until all points are distributed into different clusters. The DBSCAN model needn't us to give number of clusters and can distinguish all kinds of clusters. On the other hand, DBSCAN model don't care the order of records which is appropriate for real life.

B. One-SVM

Traditional SVM classifier need labelled data. In [27], the SVM classifier is adjusted into an unsupervised algorithm(One-SVM). It try to separate the whole data set from original data set. This algorithm try to find a region where most of points lie and labelled as +1, then assigned the rest of points as -1. The algorithm tries to find a hyperplane to separate data point from the origin with maximal margin. The optimization can be finished by equation 4 and 5.

$$\min_{\omega\in Y,\zeta_i\in\Re,\rho\in\Re} \frac{1}{2} \parallel \omega \parallel^2 + \frac{1}{\nu\ell}\sum_{i}^{l}\zeta_i - \rho \tag{4}$$

$$subject\ to: (\omega\cdot\Phi(x_i)) \geq \rho - \zeta_i, \zeta_i \geq 0 \tag{5}$$

Once we get the hyperplane, we can get the labels of all points like equation 6.

$$f(x) = sgn((\omega\cdot\Phi(x)) - \rho) \tag{6}$$

In equation 4, 5 and 6, p stands the origin, and w stands hyperplane in the feature space. $\Phi(x)$ stands our point in the feature space. One-SVM is an unsupervised anomaly detection model which do not need labelled data, and can split one data set into two sets.

C. Agglomerative Clustering

Agglomerative clustering[28] is a child of hierarchical clustering and do not need know the number of clusters. Firstly, we treat n points as a clusters. Then let two nearest clusters to be one cluster and repeat this process until all points are in the same cluster. This process can be showed in Figure 2. Points with the same color are treated as a cluster. In Figure 2, all points are divided into two clusters.

D. EM

Expectation Maximization(EM) [29] includes expectation step and maximization step. Firstly, we have observable variable Y, hidden variable Z, their joint distribution $P(Y,Z|\theta)$, and their conditional distribution $P(Z|Y,\theta)$. The aim of expectation maximization is getting the θ. Algorithm 1 shows the details of expectation maximization. In M-step, we can get θ when Q get the maximum. θ stands parameters, $\theta^{(i+1)}$ stands the parameters while running $i+1$ iteration. ε_1 and ε_2 stand two small positive real numbers.

V. Experiment

To evaluate our algorithm, we concentrate on two indications of performance:detection rate and false positive rate. If

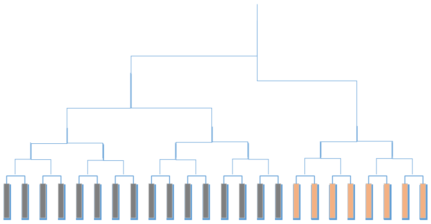

Fig. 2: Agglomerative Clustering

Algorithm 1 Expectation Maximization

```
INPUT: $P, Z, P(Y,Z|\theta), P(Z|Y,\theta)$
OUTPUT: $\theta$
begin
set the start value, $\theta^{(0)}$
while $\parallel \theta^{(i+1)} - \theta^{(i)} \parallel < \varepsilon_1$ or
        $\parallel Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i),\theta^{(i)}}) \parallel < \varepsilon_2$ do
    E-step:
        $Q(\theta, \theta^{(i)}) = E_z[logP(Y,Z|\theta)|Y, \theta^{(i)}]$
                          $= \Sigma_z logP(Y,Z|\theta)P(Z|Y,\theta^{(i)})$
    M-step:
        $\theta^{(i+1)} = arg\max_\theta Q(\theta, \theta^{(i)})$
end while
return $\theta$
```

one sample is an anomaly and our predicted label also stands anomaly, we call it true positive. If one sample is an anomaly, but our predicted label stands normal, we call it false negative. If one sample is a normal and our predicted label also stands normal, we call it true negative. If one sample is normal, but our predicted label stands anomaly, we call it false positive. TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives. From equation 7 and 8, we can get the detection rate and false positive rate.

$$detection\ rate = \frac{TP}{TP + FN} \tag{7}$$

$$false\ positive\ rate = \frac{FP}{FP + TN} \tag{8}$$

A. NEC

The aim of experiments in this part is to verifying that our model can get better results than single unsupervised anomaly detection models. DBSCAN, One-SVM, Agglomerative Clustering, EM and NEC all are run in the same data set. In the DBSCAN, we set $r = 0.0003$, and $MinPTs = 20$. In the agglomerative clustering, we set $n - clusters = 2$, which means we only split original data set into normal and anomaly. In order to get the best results, we run every model ten times and record the best results. And DBSCAN, One-SVM, Agglomerative Clustering and EM all are run ten times, and all these results are transformed into our voting component. It's worth that we need to unify all of these result, in order to ensure that 0 stands normal and 1 stands anomaly. The voting component will output the label which most results holds. Such as, there are 40 results, of which 30 is 1, while the rest of 10 0, voting component will treat it as 1 not 0. Table I shows the results of different models while running on the same data set. And we show three sets of results for every model. From table I, we can see that our NEC model shows higher positive rate and lower false positive rate. DBSCAN, One-SVM have high false positive rate and high detection rate, while Agglomerative Clustering and Em have lower detection rate and lower false positive rate. In real intrusion detection system, we wish that our system has high detection rate and low false positive rate. Our NEC can combine all models and has higher detection rate, lower false positive rate. Moreover, NEC model have more robust results compared with other models.

TABLE I: Results

Algorithm	Detection rate	False positive rate
DBSCAN	86.8%	10%
DBSCAN	82.7%	7%
DBSCAN	72%	5.66%
One-SVM	95%	8.31%
One-SVM	85.6%	6.36%
One-SVM	75%	4.43%
Agglomerative Clustering	91.6%	3.8%
Agglomerative Clustering	85.7%	3.6%
Agglomerative Clustering	79.6%	2.1%
EM	87.8%	2.3%
EM	85.66%	1.53%
EM	84.27%	0.34%
NEC	92.3%	3.66%
NEC	91.77%	2.35%
NEC	89.03%	0.78%

VI. Conclusion

In this paper, we provide a novel unsupervised anomaly detection model called NEC, which try to combine classical unsupervised anomaly detection models. This model are verified on the NSL-KDD. Compared with single models, our model is more suitable for real system, for the reason that it show stronger robustness, higher detection rate and lower false positive rate.

Acknowledgment

We thank all the anonymous reviewers for their valuable comments and helps in improving this paper.

References

[1] OLeary D. Intrusion-detection systems[J]. Journal of Information Systems, 1992, 6(1): 63-74.

[2] Patcha A, Park J M. An overview of anomaly detection techniques: Existing solutions and latest technological trends[J]. Computer networks, 2007, 51(12): 3448-3470.

[3] Roesch M. Snort: Lightweight Intrusion Detection for Networks[C]//LISA. 1999, 99(1): 229-238.

[4] Khanna R, Liu H. System approach to intrusion detection using hidden markov model[C]//Proceedings of the 2006 international conference on Wireless communications and mobile computing. ACM, 2006: 349-354.

[5] Zolotukhin M, H?m?l?inen T, Kokkonen T, et al. Analysis of HTTP requests for anomaly detection of web attacks[C]//Dependable, Autonomic and Secure Computing (DASC), 2014 IEEE 12th International Conference on. IEEE, 2014: 406-411.

[6] Bhavsar Y B, Waghmare K C. Intrusion detection system using data mining technique: Support vector machine[J]. International Journal of Emerging Technology and Advanced Engineering, 2013, 3(3): 581-586.

[7] Parhizkar E, Abadi M. OC-WAD: A one-class classifier ensemble approach for anomaly detection in web traffic[C]//2015 23rd Iranian Conference on Electrical Engineering. IEEE, 2015: 631-636.

[8] Hasan M A M, Nasser M, Pal B, et al. Support vector machine and random forest modeling for intrusion detection system (IDS)[J]. Journal of Intelligent Learning Systems and Applications, 2014, 6(1): 45.

[9] Noto K, Brodley C, Slonim D. FRaC: a feature-modeling approach for semi-supervised and unsupervised anomaly detection[J]. Data mining and knowledge discovery, 2012, 25(1): 109-133.

[10] Wulsin D, Blanco J, Mani R, et al. Semi-supervised anomaly detection for EEG waveforms using deep belief nets[C]//Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on. IEEE, 2010: 436-441.

[11] Eskin E, Arnold A, Prerau M, et al. A geometric framework for unsupervised anomaly detection[M]//Applications of data mining in computer security. Springer US, 2002: 77-101.

[12] Zhang J, Zulkernine M. Anomaly based network intrusion detection with unsupervised outlier detection[C]//2006 IEEE International Conference on Communications. IEEE, 2006, 5: 2388-2393.

[13] Hu W, Gao J, Wang Y, et al. Online adaboost-based parameterized methods for dynamic distributed network intrusion detection[J]. IEEE Transactions on Cybernetics, 2014, 44(1): 66-82.

[14] Gowrison G, Ramar K, Muneeswaran K, et al. Minimal complexity attack classification intrusion detection system[J]. Applied Soft Computing, 2013, 13(2): 921-927.

[15] Hu W, Hu W, Maybank S. Adaboost-based algorithm for network intrusion detection[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 38(2): 577-583.

[16] Tumer K, Agogino A K. Ensemble clustering with voting active clusters[J]. Pattern Recognition Letters, 2008, 29(14): 1947-1953.

[17] Zheng L, Li T, Ding C. Hierarchical ensemble clustering[C]//2010 IEEE International Conference on Data Mining. IEEE, 2010: 1199-1204.

[18] Jing L, Tian K, Huang J Z. Stratified feature sampling method for ensemble clustering of high dimensional data[J]. Pattern Recognition, 2015, 48(11): 3688-3702.

[19] Ding H, Su L, Xu J. Towards distributed ensemble clustering for networked sensing systems: a novel geometric approach[C]//Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing. ACM, 2016: 1-10.

[20] Amini M, Rezaeenour J, Hadavandi E. Effective intrusion detection with a neural network ensemble using fuzzy clustering and stacking combination method[J]. Journal of Computing and Security, 2015, 1(4).

[21] https://web.archive.org/web/20150205070216/http://nsl.cs.unb.ca/NSL-KDD/

[22] Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(03): 337-372.

[23] Kruegel C, Vigna G, Robertson W. A multi-model approach to the detection of web-based attacks[J]. Computer Networks, 2005, 48(5): 717-738.

[24] Este A, Gringoli F, Salgarelli L. Support vector machines for TCP traffic classification[J]. Computer Networks, 2009, 53(14): 2476-2490.

[25] Wang Y. A multinomial logistic regression modeling approach for anomaly intrusion detection[J]. Computers & Security, 2005, 24(8): 662-674.

[26] Gan J, Tao Y. DBSCAN revisited: mis-claim, un-fixability, and approximation[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 519-530.

[27] Sch?lkopf B, Platt J C, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution[J]. Neural computation, 2001, 13(7): 1443-1471.

[28] Davidson I, Ravi S S. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results[C]//European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 2005: 59-70.

[29] Moon T K. The expectation-maximization algorithm[J]. IEEE Signal processing magazine, 1996, 13(6): 47-60.