INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 11(4):567-579, August 2016.

An Abnormal Network Traffic Detection Algorithm Based on Big Data Analysis

H.P. Yao, Y.Q. Liu, C. Fang

Haipeng Yao*
1. State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
No 10, Xitucheng Road, Haidian District, Beijing, PRC
2. Beijing Advanced Innovation Center for Future Internet Technology
Beijing University of Technology
100 Ping Le Yuan, Chaoyang District, Beijing, PRC
*Corresponding author: yaohaipeng@bupt.edu.cn

Yiqing Liu
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
No 10, Xitucheng Road, Haidian District, Beijing, PRC
colin617@qq.com

Chao Fang
1. Beijing Advanced Innovation Center for Future Internet Technology
Beijing University of Technology
100 Ping Le Yuan, Chaoyang District, Beijing, PRC
fangchao.bupt@gmail.com
2. College of Electronic Information and Control Engineering
Beijing University of Technology
100 Ping Le Yuan, Chaoyang District, Beijing, PRC
fangchao.bupt@gmail.com

Abstract: Anomaly network detection is a very important way to analyze and detect malicious behavior in network. How to effectively detect anomaly network flow under the pressure of big data is a very important area, which has attracted more and more researchers' attention. In this paper, we propose a new model based on big data analysis, which can avoid the influence brought by adjustment of network traffic distribution, increase detection accuracy and reduce the false negative rate. Simulation results reveal that, compared with k-means, decision tree and random forest algorithms, the proposed model has a much better performance, which can achieve a detection rate of 95.4% on normal data, 98.6% on DoS attack, 93.9% on Probe attack, 56.1% on U2R attack, and 77.2% on R2L attack.
Keywords: Anomaly Traffic Detection, Big Data, K-means, Decision Tree, Random Forest.

1 Introduction

With the improvement of network, storage, calculation and transmission,the Internet is interacting more closely with people than ever before. While the Internet is making our life more convenient, it brings about some potential risks. For example, malicious attacks involving user privacy and security become more and more frequent.

The changes of how people use the Internet is a new challenge for traditional abnormal network event detection techniques. It is more hard for researchers to get aware of some new kinds of attacks. To resolve these problems, some abnormal network traffic detection methods

have been proposed. Traditional abnormal traffic detection method can be classified into two categories [1–3]. One is misuse detection, and the other is abnormal detection. The two methods have their own pros and cons. Misuse detection has a high accuracy but needs support from known knowledge. Abnormal detection do not need known knowledge, but cannot categorize the type of attacks, the accuracy is also lower. For example, Hari Om [4] designes a hybrid detection system, which is a hybrid anomaly detection system considering k-means, k-nearest neighbor and Naïve Bayes methods.

However, the explosive increase of network traffic has directly or indirectly pushed the Internet into the big data era, which makes anomaly traffic detection more difficult to deal with because of high calculation volume and constant changes of network data distribution caused by big data [5–8]. Because the speed of network data generation is fast, it makes the volume of normal traffic and abnormal traffic differ a lot, and the distribution of the data change. Besides, with big data, the difference between normal traffic and abnormal traffic is increasing. It makes the traditional methods unable to effectively detect abnormal traffic.

Therefore, to increase the accuracy of abnormal traffic and avoid the loose caused by false negative detection, we propose a novel model based on big data analytics, which can avoid the influence brought by adjustment of network traffic distribution, increase dectection accuracy and reduce the false negative rate. The core of the proposed model is not simply combination of traditional detection methods, but a novel detection model based on big data. In the simulation, we use k-means, decision tree and random forest algorithms as comparative objects to vertify the effectiveness of our model. Simulation results reveal that the proposed model has a much better performance, which can achieve a detection rate of 95.4% on normal data, 98.6% on DoS attack, 93.9% on Probe attack, 56.1% on U2R attack, and 77.2% on R2L attack.

The rest of this paper is organized as follows. In Section 2, related work of this paper is presented. The system model is given in Section 3. Simulation results are presented and discussed in Section 4. Finally, we conclude this study in Section 4.3.

2 Related work

2.1 k-means

k-means is a classic clustering algorithm [9,10], which uses simple iteration algorithm to cluster the data set into certain amount of categories.Commonly, the number of clusters is annotated to be K. The four steps of k-means are:

1. Initialization: Randomly select K data points from the data set as the centers of the K-clusters;
2. Distribution: Assign each point in the data set to the nearest center;
3. Update: calculate new centers according to the cluster assignment, the new center is the average point of all the points in a cluster;
4. Repeat: Repeat these steps until no center is updated in this round, and the clustering is converged.

k-means needs the number of classification K to be specified. If K is not chosen properly, it will lead to an improper result of classification. So choose a proper cluster number is crucial to the result of k-means.

Another disadvantage of k-means is that, k-means can only use Euclidean distance. Even though Euclidean distance is convenient to calculate, but it cannot take the difference between two features into consideration, it means it treats all features as same. In the reality, it will sometimes lead to poor performance.

Anyway, k-means has its own advantages when dealing with big data.

1. k-means is simple. The time complexity is $n\left(n^{d*k+1}\log n\right)$, it can be fast when the number of clusters and the number of features are small;
2. k-means can be well adjusted to big data set and has high performance.

2.2 Decision tree

Decision Tree [9] is a common algorithm used in machine learning. A complete decision tree is composed by three kind of elements:
1. Decision Node, indicating which feature is used in split;
2. Chance nodes, indicating possible values of each features;
3. Leaf node, indicating which category is the record in.

There are two steps needed to use a decision tree:
1. Tree generation: Generate a tree according to training set. Need to determine which feature need to use in the split, and determine which category the result is in.
2. Classification: Classify new records from the root of the decision tree, and compare the record with each of the decision node, move to corresponding branch with the result. Repeat this process, and after a data reaches the leaf node, the category of leaf node is the new category of the node.

Quinlan proposed C4.5 algorithm in [11], which is a well known decision tree algorithm.The main method is to generate the decision tree from root to leaf, in order to reduce the level of uncertainty. Therefore, this algorithm can be described as follows.

Gain ratio is the index C4.5 used to select feature. Define a feature in the feature set to be A_k, the training set to be Tď ż˝ď ż˝definition of information gain is defined like this:

$$Gain(T, A_k) = Info(T) - Info_{A_k}(T) \tag{1}$$

where

$$Info(T) = -\sum_{i=1}^{n} \frac{freq(c_i, T)}{|T|} \log_2 \frac{freq(c_i, T)}{|T|} \tag{2}$$

$$Info_{A_k}(T) = -\sum_{a_k \in D(A_k)} \frac{|T_{a_k}^{A_k}|}{|T|} Info(T_{a_k}^{A_k}) \tag{3}$$

$freq(c_i, T)$ means the number of records belongs to c_i in T. $T_{a_k}^{A_k}$ express that subset which A_k is a_k, and domain of A_k is $D(A_k)$.

$SplitInfo(A_k)$ is defined to be:

$$SplitInfo(T, A_k) = -\sum_{a_k \in D(A_k)} \frac{|T_{a_k}^{A_k}|}{|T|} \log_2 \frac{|T_{a_k}^{A_k}|}{|T|} \tag{4}$$

Gain ratio is

$$Gainratio(T, A_k) = \frac{Gain(T, A_k)}{SplitInfo(A_k)} \tag{5}$$

The advantages of decision tree are:
1. The tree generated is easy to generate and easy to explain;
2. Performs well when dealing with large data set.

2.3 Random forest

Random Forest algorithm [9, 12] is a classification algorithm and contains multiple decision trees, where each tree has a vote, and result is the one with highest vote.

When generating decision tree, feature selection and pruning can be used to avoid over fitting. But when the number of features is large, the problems can hardly be avoided. Random forest consists of multiple decision trees, which can effectively avoid those problems.

Random forest has following advantages:

1. It can be used in various situation with a pretty high accuracy on classification;
2. It can effectively support multi-feature situation without feature selection;
3. It can report the importance distribution of features.

3 System model

Influenced by big data, network data distribution is gradually changing. This paper try to solve the problem that caused by the increasing difference between normal traffic and abnormal traffic. Therefore, we proposed a new abnormal traffic detection model based on big data analysis, and this model includes three sub-models.

3.1 Normal traffic selection model

Normal traffic selection model uses classification and clustering algorithm to distinguish normal and anomaly behaviors, rather than involved specific anomaly behaviors. This model includes two stages:

1. Training stage: training model uses data that labeled normal or abnormal, and the model applies in test stage.

2. Test stage: test stage is similar to detection in practice. Using unlabeled date, the model classifies traffic data into normal or abnormal, and labels them.

Normal traffic selection model uses k-means clustering algorithm, KNN, decision tree and random forest classification algorithms. Traditionally, before using k-means algorithms, it is very important to set the number of categories, because we don't know how many categories. But in order to distinguish normal and abnormal behavior, the normal traffic selection model uses k-means as following way.

In training stage, using labeled information classify data into normal and abnormal. These two categories use k-means separately instead of clustering all data at once, getting the center of the data set respectively. Then using the center of the data set, KNN clustering algorithm classifies test data. Decision tree and random forest classification algorithms train with labeled normal and abnormal data.

3.2 Abnormal traffic selection model

The purpose of abnormal traffic selection model is avoid influence caused by too many normal traffic than abnormal traffic. This model classifies anomaly traffic into specific categories, and includes two stage as well:

1. Training stage: this stage only use abnormal data to train classification model, and every data label specific attack group. Using classification algorithms learns classified rules.

2. Test stage: test stage is similar to detection in practice, using unlabeled data (including normal behavior data). The classification model classifies anomaly traffic into specific categories according to the classified rules, and gives specific label to every data.

Table 1: Distribution of KDDCUP99 data set

Data set	Normal	DoS	Probe	R2L	U2R
10 percent of training data set	97278	391458	4107	1126	52
test data set	60593	229853	4166	16189	228

Abnormal traffic selection model uses decision tree and random forest classification algorithms. Abnormal traffic selection model and normal traffic selection model are independent, without order of priority in training stage or test stage.

Mixed compensation model combines the result from normal traffic selection model and abnormal traffic selection model to produce a final result. Although abnormal traffic selection model is more effective because without influence of normal traffic data, the model has high false negative rate due to this characteristic. Therefore use normal set produced by normal traffic selection model to compensate abnormal set $A = \{A_1, A_2, \cdots, A_k\}$ produced by abnormal traffic selection model. $A_i, i \in [1, k]$ denote specific attack category. If c denote detection result, rule of compensation as follow:

$$\begin{cases} \text{if } c \in A_i,\ c \in N, \text{ then } c \in N \\ \text{if } c \in A_i,\ c \notin N, \text{ then } c \in A_i. \end{cases} \tag{6}$$

4 Simulation results and discussions

Before using three sub-models of anomaly detection based on big data analysis, data set needs be preprocessed with label for training model. It should be noted that rightly selecting feature is a good way to reduce dimension and increase efficiency of running. In the simulation, three different algorithms are used to verify validity of the proposed model.

4.1 Data set

In the simulation, we use KDDCUP99 [13] data set to test my model. KDDCUP99 data set is widespread use for testing abnormal detection model, which is obtained and processed from KDDCUP99 [14]. KDDCUP99 data set has 41 features and been sorted into three group: basic feature, content feature and time feature [15].

The distribution of data set is shown as Table 1, where training data has 5 million records, 10 percent of training data has 494021 records, and test data has 311029 records. Every record is labeled to be normal or abnormal, and abnormal data can be classified into four groups: Dos, U2R, R2L and Probe. From Table 1, we find that normal data in training data set is more than abnormal data in test data set. Therefore, this data set can be used to test the performance of the proposed model under different circumstances.

4.2 Simulation results

As shown in Table 2, we have done eight experiments with the model based on big data analysis, and three control experiments which used k-means, decision tree or random forest respectively. In the control groups, training classify model uses all training data set with five categories, then classifying test data into five categories. Another control group is winner of KDDCUP99.

In the simulation, prediction accuracy is used as a simulation metric of detection effect, which is shown in Table 3. Besides, we adopt way of sorting and grading for every type. For example,

Table 2: Number of experiments

No.	Normal traffic selection model	Abnormal traffic selection model	No. of control group	Algorithm
1	k-means1*	Random Forest	9	k-means
2	k-means1*	Decision Tree	10	Decision Tree
3	k-means2*	Random Forest	11	Random Forest
4	k-means2*	Decision Tree	12	Winner of KDDCUP99
5	Decision Tree	Decision Tree		
6	Decision Tree	Random Forest		
7	Random Forest	Decision Tree		
8	Random Forest	Random Forest		

*note: In the normal traffic selection model, the number of cluster of normal and abnormal respectively is 4 and 30 in $k-means1$, and the number of cluster of normal and abnormal respectively is 100 and 300 in $k-means2$.

Table 3: Prediction accuracy

No.	Experiment	Normal	DoS	Probe	U2R	R2L
1	k-means1+Random Forest	0.632	0.814	0.939	0.561	0.679
2	k-means1+Decision Tree	0.656	0.791	0.878	0.500	0.772
3	k-means2+Random Forest	0.945	0.983	0.910	0.513	0.510
4	k-means2+Decision Tree	0.946	0.979	0.852	0.500	0.504
5	Decision Tree+Decision Tree	0.951	0.984	0.829	0.500	0.512
6	Decision Tree + Random Forest	0.951	0.986	0.831	0.550	0.517
7	Random Forest + Decision Tree	0.954	0.980	0.861	0.500	0.521
8	Random Forest + Random Forest	0.952	0.985	0.872	0.520	0.510
9	k-means	0.938	0.968	0.785	0.500	0.528
10	Decision Tree	0.951	0.983	0.793	0.500	0.500
11	Random Forest	0.952	0.985	0.875	0.522	0.507
12	Winner of KDDCUP99	0.995	0.971	0.833	0.132	0.084

all experiments are sorted by prediction accuracy of normal. The first grades 1 point, the second grades 2 points, and so on. Finally, adding grade of five groups is final grade.

As shown in Table 4, the experiment group and winner of KDDCUP99 are sorted by final grade. While the later has high detection rate in normal data, as for four attack types, the result of model based on big data analysis is better than winner of KDDCUP99.

Algorithm of winner of KDDCUP99 is C5 decision tree [16–19]. Training data of winner of KDDCUP99 is a little different with my experiment. Thus for evaluating detection effect of the proposed mode, we did three control experiments with same training data and test data, used with k-means, decision tree or random forest respectively. The number of these experiments is noted as 11, 10 and 9.

Sorting result shows that detection effect of algorithm of the proposed model is better than no use, as shown as Table 5. We will discuss experiments results, compared No.8 with No.11, No.7 with No.5 and No.3 with No.4.

Discussing result of no.8 and no.11

Score of top three are same. Judging No.8 and No.11 with final grade, detection result of two experiments are almost same. And both of them use random forest algorithm. But the difference is:

Table 4: Compared with winner of KDDCUP99

No.	Experiment	Normal	DoS	Probe	U2R	R2L	Final Socre	Rank
8	Random Forest+Random Forest	3	2	2	2	4	13	1
6	Decision Tree+Random Forest	4	1	6	1	2	14	2
7	Random Forest +Decision Tree	2	5	3	4	1	15	3
2	k-means2+Random Forest	7	4	1	3	5	20	4
5	Decision Tree+Decision Tree	4	3	7	6	3	23	5
4	k-means2+Decision Tree	6	6	4	5	6	27	6
12	Winner of KDDCUP99	1	7	5	7	7	27	6

Table 5: Compared with control group

No.	Experiment	Normal	DoS	Probe	U2R	R2L	Final Socre	Rank
6	Decision Tree+Random Forest	4	1	6	1	3	15	1
8	Random Forest + Random Forest	2	2	3	3	5	15	1
11	Random Forest	2	2	2	2	7	15	1
7	Random Forest + Decision Tree	1	7	4	5	2	19	4
3	k-means2+ Random Forest	8	5	1	4	5	23	5
5	Decision Tree + Decision Tree	4	4	7	5	4	24	6
10	Decision Tree	4	5	8	5	9	31	7
9	k-means	9	9	9	5	1	33	8
4	k-means2+ Decision Tree	7	8	5	5	8	33	8

1. Importance of variable used in classifying is different;
2. No.8 has lower false negative rate.

• **Importance of variable**

As shown as Fig. 1, variables chosen by random forest in No.8 and No.11 are different. Random forest algorithm can output importance of variables, noted Gini index [9]. Fig. 1 shows that top 20 have important variables in comparison with top 1, whose value is higher and more important.

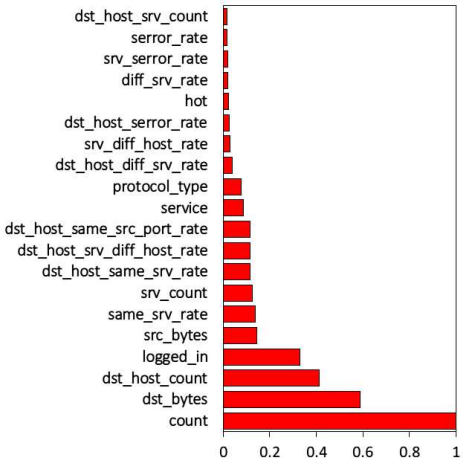
In No.8, rank of variables is different between normal traffic selection model and abnormal traffic selection model. This means that variable used for predicting normal or abnormal and specific attack is different. Therefore, choosing variable in No.11 is influenced by both sides, and output a compromised result when choosing variables, that's why prediction of model in No.11 has deviation.

• **Comparison of false negative rate**

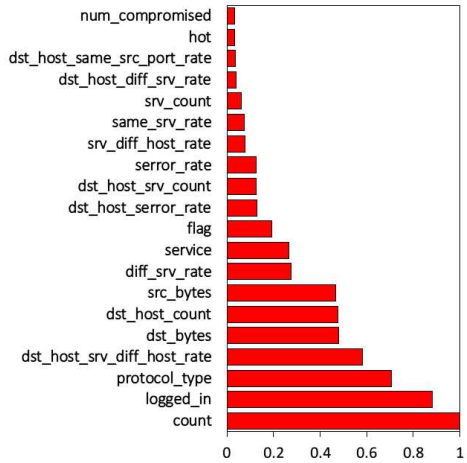
In order to evaluate effect on predicting abnormal behavior, false negative rate is used as an important index, which can measure how many attack events are omitted. Table 6 shows confusion matrix of results of experiments No.8 and No.11 when using random forest. Row express information of prediction, and column express actual information. False negative rate of No.8 in normal type is very low, but high in U2R and R2L type. In No.8, false negative rate of normal selection model in normal is low. Without influence of normal training data, false negative rate of abnormal selection model in four specific attack types are lower than No.11.

Discussing result of no.5 and no.7

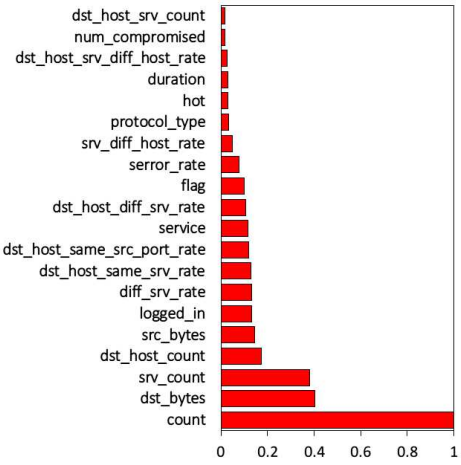
No.5 and No.7 respectively compare with No.6 and No.8 by using same algorithm in normal traffic selection model, and their ranks are lower when using decision tree in abnormal traffic



(a) No.11



(b) No.8 Normal traffic selection model



(c) No.8 Abnormal traffic selection model

Figure 1: Importance of Variables in Random Forest.

Table 6: Confusion matrix

No.11					
Prediction	**Normal**	**DoS**	**Probe**	**U2R**	**R2L**
Normal	60287	5967	847	159	15839
DoS	69	223814	191	8	0
Probe	233	72	3128	50	104
U2R	1	0	0	10	5
R2L	3	0	0	1	241
False Negative	0.00505	0.026273	0.24916	0.95614	0.985113
No.8 **Normal traffic selection model**					
Prediction		**Normal**		**Abnormal**	
Normal		60289		22853	
Abnormal		304		227583	
False Negative		0.005017		0.091253	
No.8 **Abnormal traffic selection model**					
Prediction	**DoS**	**Probe**	**U2R**	**R2L**	
DoS	229231	769	20	4693	
Probe	297	3393	135	5646	
U2R	0	0	39	32	
R2L	325	4	34	5818	
False Negative	0.002706	0.18555	0.828947	0.64062	

Table 7: Confusion Matrix of abnormal traffic selection model with decision tree

Predition	**DoS**	**Probe**	**U2R**	**R2L**
DoS	227792	589	34	6245
Probe	1434	3192	20	283
U2R	0	0	0	0
R2L	627	385	174	9661

selection model.

Table 7 is confusion matrix of abnormal traffic selection model with decision tree algorithm. It shows that U2R can not be detected and false negative rate of R2L is higher. In order to find the reason, classify tree is checked in Fig. 2, where the classification model prefers DoS and Probe attack, then R2L attack, and no result point of U2R attack. Distribution of training data can explain this phenomenon, which can be shown in Fig. 3.

When generating decision tree, the obtained information will cause results in favor of feature which have more samples. Therefore, if the number of training data set in every group is different enough, it cannot get efficient classification model for small samples. Moreover, because the number of between training data is comparatively equal, classification result is better, such as No.6, when normal traffic selection model uses decision tree.

Discussing result of no.3 and no.4

No.3 and No.4 use k-means in normal traffic selection model to choose clustering center. Table 8 shows final prediction accuracies in No.3 and No.4. Because final results are lower than that of normal traffic selection model or abnormal traffic selection model, we find that this problem is caused by using k-means in normal selection model. Table 9 shows confusion matrix

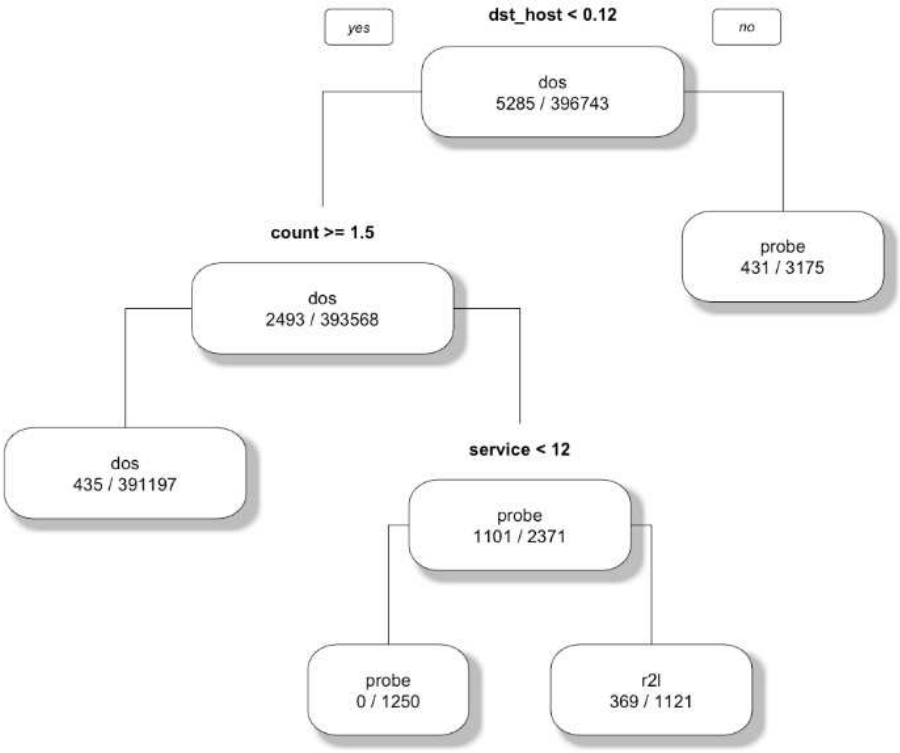


Figure 2: Classify tree of abnormal traffic selection model.

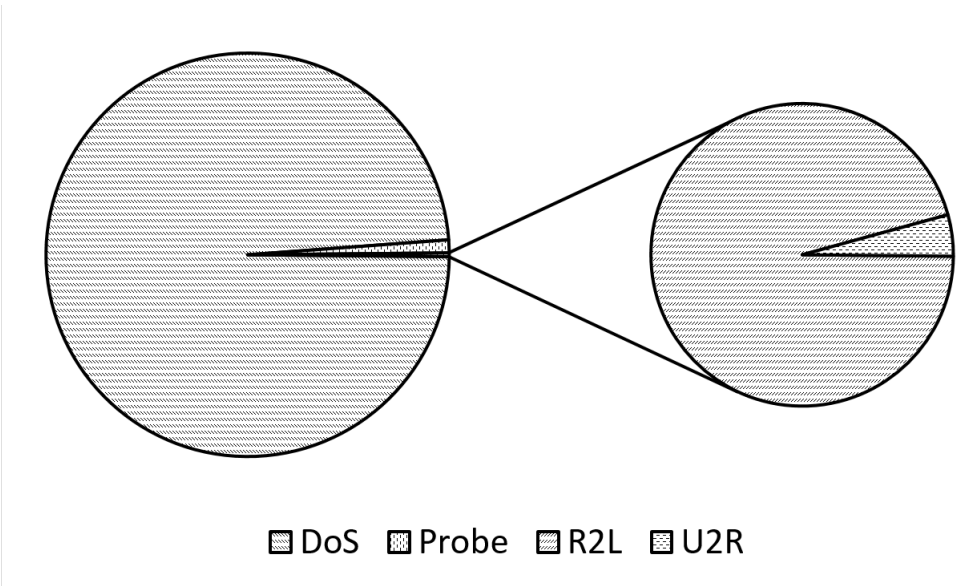


Figure 3: Distribution of training data.

Table 8: Accuracy of no.3 and no.4

No.	Model	Algorithm	Accuracy
No.3	Normal Traffic Selection	k-means	0.926
	Abnormal Traffic Selection	Random Forest	0.948
	Mixed Compensation Model		0.923
No.4	Normal Traffic Selection	k-means	0.925
	Abnormal Traffic Selection	Decision Tree	0.961
	Mixed Compensation Model		0.918

Table 9: Confusion matrix of normal traffic selection model of no.3 and no.4

No.	Prediction	Normal	Abnormal
No.3	Normal	59189	21663
	Abnormal	1404	228773
No.4	Normal	59428	22221
	Abnormal	1165	228215

of normal traffic selection model of No.3 and No.4. Many abnormal records are predicted as normal, which cause high false negative rate. Therefore, many abnormal records predicted by abnormal traffic selection model will be regarded as normal after mixed compensation model.

Nowadays, many novel attacks are unknown to researchers, and many attacks will be disguised as normal. It's very dangerous to have high false negative rate, and it does not fit the proposed model.

Because the effect of k-means has great correlation with the number of centers chosen to cluster, and we can fine tune the strength of clustering, and lower the false negative rate to establish a strict normal selection model.

In No.3 and No.4, the number of centers for normal traffic and attacks is 100 and 300, respectively. Although it can achieve a good overall accuracy, its false negative rate is higher than other model. However, according to Table 10, by choosing 4 and 30 in No.1 and No.2, it has lower false negative rate, and only classify four kinds of attacks. Besides, a strict normal detection model is established.

By adjusting the parameters and reducing false negative rate in No.1 and No.2, we can find that the rank has increased rapidly compared with No.3 and No.4. Especially, when K-means combines with random forest, it has a very high accuracy on Probe, U2R and R2L attack. Therefore, we can draw the conclusion that by adjusting the parameters of K-means, the strength of abnormal traffic detection can be controlled by adjusting the strength of normal traffic identification.

4.3 Summary

Based on the results analyzed above, as shown in Table 11, the following conclusions can be drawn:

1. Random forest classification algorithm can adapt to the change of distribution of network data, and this algorithm by using the proposed model can reduce false negative rate.

2. If the number of training data in different group is largely different with each other, the classify model built by decision tree will prefer to attack types, which have more training data. So we should avoid using decision tree in abnormal traffic selection model. However, in the normal traffic selection model, the difference between different groups is comparatively small. In this

Table 10: Results of experiments

No.	Experiment	DoS	Probe	U2R	R2L	Final	Rank
1	k-means1+Random forest	10	1	1	2	14	1
6	Decision tree+ Random forest	1	8	2	5	16	2
8	Random forest + Random forest	2	5	4	7	18	3
11	Random forest	2	4	3	9	18	3
3	k-means2+ Random forest	5	2	5	7	19	5
2	k-means1+ Decision tree	11	3	6	1	21	6
7	Random forest + Decision tree	7	6	6	4	23	7
5	Decision tree + Decision tree	4	9	6	6	25	8
9	k-means	9	11	6	3	29	9
4	k-means2+ Decision tree	8	7	6	10	31	10
10	Decision tree	5	10	6	11	32	11

Table 11: Summary of model

	Model 1	Model 2	Model 3
Normal traffic selection model	k-means1	Decision Tree	Random Forest
Abnormal traffic selection model	Random Forest	Random Forest	Random Forest

situation, using decision tree can fast get classify model, and the results have higher accuracy.
3. There are more and more unknown abnormal events in the future. In order to avoid loss of false negative prediction, we can change the number of clustering in the normal traffic selection model with k-means algorithm to reduce false negative rate and increase the accuracy of detecting abnormal events.

Conclusion

With the change of distribution of network data, traditional anomaly traffic detection techniques can not fit this situation. In order to solve the problem, we propose an anomaly traffic detection model based on big data analysis. Simulation results show that the proposed model achieves a detection rate of 95.4% on normal data, 98.6% on DoS attack, 93.9% on Probe attack, 56.1% on U2R attack, and 77.2% on R2L attack.Therefore, the model can increase the accuracy of attack behavior, and reduce false negative rate.

Acknowledgment

This work was supported by NSFC (61471056) and China Jiangsu Future Internet Research Fund (BY2013095-3-1, BY2013095-3-03).

Bibliography

[1] Patcha, A.; Park, J.M. (2007); An overview of anomaly detection techniques: Existing solutions and latest technological trends, *Computer Networks*, ISSN 1389-1286, 51(12): 3448-3470.

[2] Lazarevic, A.; Kumar, V.; Srivastava, J. (2005); Intrusion detection: A survey, *Managing Cyber Threats*, ISSN 0924-6703, 5: 19-78.

[3] Axelsson, S. (1998); Research in intrusion-detection systems: a survey, *Department of Computer Engineering, Chalmers University of Technology, Goteborg. Sweden*, Technical Report 98-17.

[4] Om, H.; Kundu, A. (2012); A hybrid system for reducing the false alarm rate of anomaly intrusion detection system, *IEEE 1st International Conference on Recent Advances in Information Technology (RAIT)*, ISBN 978-1-4577-0694-3, 131-136.

[5] Kaisler, S. et al (2013); Big data: Issues and challenges moving forward, *IEEE 46th Hawaii International Conference on System Sciences (HICSS)*, ISSN 1530-1605, 995-1004.

[6] Michael, K.; Miller, K.W. (2013); Big Data: New Opportunities and New Challenges, *Computer*, ISSN 0018-9162, 46(6):22-24.

[7] Russom, P. et al (2011); Big Data Analytics, *TDWI Best Practices Report*, Fourth Quarter.

[8] Fan, W.; Bifet, A. (2013); Mining big data: current status, and forecast to the future, *ACM SIGKDD Explorations Newsletter*, ISSN 1931-0145, 14(2): 1-5.

[9] James, G. et al (2013); An introduction to statistical learning, *Springer*, ISSN 1431-875X.

[10] Guan, Y.; Ghorbani, A.A.; Belacel, N. (2003); Y-means: A clustering method for intrusion detection, *IEEE Canadian Conference on Electrical and Computer Engineering*, ISSN 0840-7789, 2:1083-1086.

[11] Quinlan, J.R. (1993); C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers Inc.*, ISBN 1558602402.

[12] Elbasiony, R.M. et al (2013); A hybrid network intrusion detection framework based on random forests and weighted k-means, *Ain Shams Engineering Journal*, ISSN 2090-4479, 4(4): 753-762.

[13] KDD Cup 1999, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. May 2015

[14] Lippmann, R.P. et al (2000); Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation, *IEEE Proceedings of DARPA Information Survivability Conference and Exposition (DISCEX)*, ISBN 0-7695-0490-6, 2:12-26.

[15] Tavallaee, M. et al (2009); A detailed analysis of the KDD CUP 99 data set, *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA)*, ISBN 978-1-4244-3763-4, 1-6.

[16] Pfahringer, B. (2000); Winning the KDD99 classification cup: bagged boosting, *ACM SIGKDD Explorations Newsletter*, ISSN 1931-0145, 1(2): 65-66.

[17] Yu, G. D. et al (2014); Multi-objective rescheduling model for product collaborative design considering disturbance, *International journal of simulation modelling*, ISSN 1726-4529, 13(4): 472-484.

[18] Gusel, L. R. et al (2015); Genetic based approach to predicting the elongation of drawn alloy, *International journal of simulation modelling*, ISSN 1726-4529, 14(1): 39-47.

[19] Prasad, K. et al (2016); A knowledge-based system for end mill selection, *Advances in Production Engineering & Management*, ISSN 1856-6250, 11(1): 15-28.