International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.11 (2015), pp.253-264
http://dx.doi.org/10.14257/ijmue.2015.10.11.24

A Feature Selection Algorithm towards Efficient Intrusion Detection

Chunyong Yin[1], Luyu Ma[1], Lu Feng[1], Zhichao Yin[2] and Jin Wang[1]

[1]*School of Computer and Software, Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[2]*Nanjing No.1 Middle School, Nanjing 210001, China*

Abstract

Feature selection algorithm plays a crucial role in intrusion detection, data mining and pattern recognition. According to some evaluation criteria, it gets optimal feature subset by deleting unrelated and redundant features of the original data set. Aiming at solving the problems about the low accuracy, the high false positive rate and the long detection time of the existing feature selection algorithm. In this paper, we come up with a feature selection algorithm towards efficient intrusion detection, this algorithm combines the correlation algorithm and redundancy algorithm to chooses the optimal feature subset. Experimental results show that the algorithm shows almost and even better than the traditional feature selection algorithm on the different classifiers.

Keywords: *feature selection, the largest correlation, the minimum redundancy, intrusion detection*

1. Introduction

1.1. The background and Significance of the Research

With the rapid development of information technology and network, people can get more data from the network, and at the same time, the dimension of the data characteristics dimension is becoming more and more. In intrusion detection, in theory, the comprehensive information of the data will help to improve the classification accuracy of the intrusion detection; but in the actual detection, the existence of too many redundant and unrelated features will lead to the opposite effect. This redundant and useless information, not only can reduce the accurate precision of detection, will increase the time needed for testing, but also makes the overall effect of detection reduced greatly. The so-called "data rich, information redundancy" and "dimension disaster" is the embodiment of the informative but lack of effective information. So, the feature selection algorithm achieve the purpose of dimension reduction by analyzing and removing redundant and useless information, it can reduce the testing time and improve the accuracy of detection obviously. Based on this, feature selection in the intrusion detection has become a research hotspot.

1.2. The Status of Feature Selection Algorithm at Home and Abroad

In the early '60 s, the research about the feature selection algorithm is pulled open heavy curtain. Backer and Shipper [1] used the maximum and minimum algorithm for the selection of feature, this algorithm achieved the fast effect but sacrificed the detection performance, and it couldn't guarantee that the characteristics obtained is optimal feature subset; Siedlecki and Sklansky [2] cited genetic algorithm to feature selection, Though it

ISSN: 1975-0080 IJMUE

was verified by experiment, the algorithm obtained good experimental effect, but the genetic algorithm was easy to premature convergence problem.

In recent years, the applications of feature selection in intrusion detection are becoming more [3-8]. In 2012, S Suthaharan and T Panchagnula [9] used the method based on rough theory (RST) to select feature, experiments showed that the method can promise a better precision and less calculation time, but couldn't eliminate the presence of outliers; In 2013, Fengli Zhang and Dan Wang [10] carried on the effective feature selection according to the bayesian network classifiers, the experiment results showed that this method can get high accuracy and low false alarm rate in the detection of the Probe, Dos and R2L attack, but when it detects the U2R attacks, detection effect needs to be improved; In 2015, Nannan Xie and Yafen Cui [11] used the Fisher to select feature, they extracted the key features of the data set, This method in the experiment is showed that it can achieve a high detection rate in a data set with less characteristics, but it failed to show that it could get the same good effect when it detect the data with larger dimensions and more information.

At present, the feature selection as a variety of algorithms for data preprocessing steps received widespread attention, many scholars at home and abroad on the study continuously put forward the methods and ideas of their own.

1.3. The Problem Faced by Feature Selection Algorithm

1) Data quantity, high dimension, difficult to handle; Network intrusion detection problem, too large amount of data not only can increase the detection time but also reduce the accuracy of detection, at the same time, detection of computer also have a memory problem due to the amount of data to test.

2) Too much data redundancy, too little effective information, too much useless and redundant information will result in the problem of "data rich, information redundancy, these problems can bring tremendous problem of network intrusion detection, they will not only reduce the detection accuracy but also can make the rate of false positives be improved.

3) In view of the different algorithms, the difficulties of the optimal feature subset selection are big. The optimal feature subset of the data is relative, not every kind of the optimal feature subset of solutions can satisfy each kind of intrusion detection algorithm. So how to find the suitable for common intrusion algorithm of the optimal feature subset is particularly important.

1.4 The Work in this Paper

In this article, we will focus on intrusion detection algorithm to select the optimal feature subset. We will select the optimal feature subset by combining the maximum correlation algorithm and minimum redundancy algorithm, and according to the contrast test, we will evaluate the optimal feature subset under the standard of modeling time and the accuracy of detection. Besides, we carry comparison experiment in the different classifiers about the subset, and we evaluate the subset through the experimental results.

2. Feature Selection Algorithm

2.1. Definitions

Feature Selection has been viewed and defined by different researchers, and been studied and analyzed from the perspective of different fields. John [12] presents that feature selection should not reduce the classification accuracy of classifiers in 1994. Koller [13] presents that feature selection should select data as few as possible without changing the distribution of date in 1996. Dash [14] presents that feature selection should

work by reducing the dimension of data without influencing the detection precision in 1997. Thus, the definitions of feature selection have different emphases. At the same time, there is no unified standard when different classifiers test different data sets under different algorithms. So, in terms of mathematics, feature selection is essentially a process of search optimization, which would choose the best optimum combination by solving different combinations. Here are some typical definitions.

(1) There are M features in a data set and we select N features from it (N <= M), which is the optimization when compared with the original data set under the same evaluation standard.

(2) Reduce the total number of features as far as possible without lower detection classification standards under the same classifier [15].

(3) Select features when ensure there are no large deviation probability distribution in the feature selected and original data set.

Feature selection is a kind of feature space dimension compression method. It computes the original features through the method of transformation, which makes the secondary characteristics after transformation can remove some component [16].

Here is an example of feature selection. For a feature vector $x = (x_1, x_2, ..., x_n)^T$, which has n original features. Feature selection is to transform vector x and generate a vector $y = (y_1, y_2, ..., y_d)^d$, which has d features and $d \le n$. We can present the vector y as:

$$y = W^T x$$

In which $W = W_{(n \times d)}$ is called feature selection matrix or transformation matrix. Feature selection based on separability criterion is to compute a transformation matrix W under a set of separability criterion basis.

2.2 The Commonly Used Method of Feature Selection

In intrusion detection, if you want to ensure that the final classification accuracy, you must first start with the test data set. We preprocess the data set for testing, and make it achieve the ideal effect at the time of detection classification. At present, the main methods used in intrusion detection are as follows:

1) IG (Information Gain) Information Gain method

Information gain method is used to judge a character appears or not the average amount of information, in order to measure the feature can bring information for system classification. The information carried by feature is much more, that means the feature is more important. The information gain method of formula is:

$$IG(t) = -\sum_{i=1}^{M} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{M} P(c_i \mid t) \log P(c_i \mid t) + P(\bar{t}) \sum_{i=1}^{M} P(c_i \mid \bar{t}) \log P(c_i \mid \bar{t})$$

When $\bar{t} \to t$ don't occur, where P is the probability of t appearing in event c_i, and t not appearing in event c_i, and according it to measure the amount of information contained by t.

2) The CHI (statistics) chi-square statistics

Chi-square statistic determines the correlation between features by calculate the independence of the each other between each feature. The independence is lower, then the greater the relevance; on the contrary, the higher the independence, the smaller the correlation. Chi-square statistic calculation formula is:

$$\chi^2(t, c_i) = \frac{n \times (ad - bc)^2}{(a + c) \times (b + d) \times (a + b) \times (c + d)}$$

$$\chi^2_{avg}(t) = \sum_{i=1}^{M} P(c_i)\chi^2(t, c_i)$$

Where, a means that t occurs and it belongs to c_i, b mans that t occurs but it don't belong to c_i, c means that t don't occurs but it belongs to c_i, d mans that t don't occurs and it don't belong to c_i.

These four situations is used to calculate t and c_i relative independence.

3. Feature Selection Algorithm in this Paper

Based on the above the existing intrusion detection methods, we will be combined with maximum and minimum redundancy correlation method for feature selection in this paper. Mutual information is a kind of used to represent the largest minimum redundancy algorithm of nonlinear relationship. Two random variables x and y, their density distribution function is $p(x)$ and $p(y)$, the joint probability distribution is $p(x, y)$, the mutual information of x and y are as follows:

$$I(x, y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)\, p(y_j)}$$

Mutual information can well describe the relationship of the selected features and output category. The smaller the correlation, the greater the redundancy feature subset of the more conducive to the classification of the classifier, we determine final selection of the feature by calculating the redundancy between features and characteristics of correlation between the label in this paper.

The redundancy of features is put forward to measure the common information contained by features, Contains the more mutual information, then the greater the degree of redundancy between them; and vice versa, so in the case of contains a lot of the same information, only to select the better one.

Generally the basic steps of feature selection is shown in Figure 1.

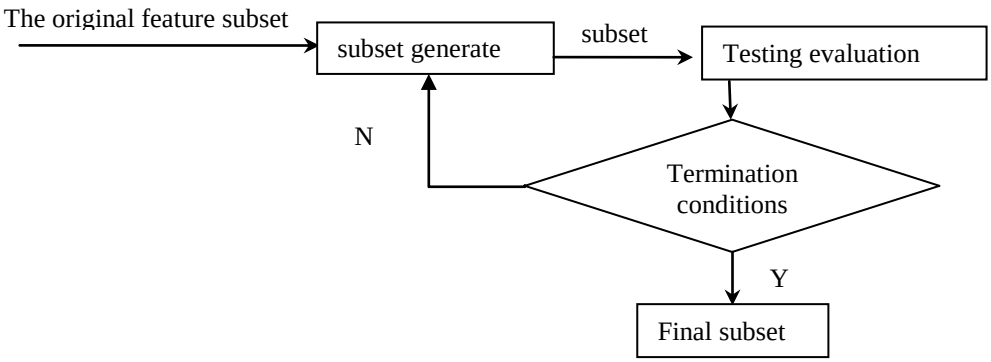


Figure 1. The Basic Steps of Feature Selection

According to the basic framework of the algorithm, in feature selection, there are four basic steps: candidate feature subset generation, evaluation standard, verify the termination conditions and final feature subset to generate. The present study is focused on the candidate feature subset generation and evaluation standard and test conditions set, so in this paper, we study how to carry out the feature subset selection and the final judgment standard setting.

In this paper, we analyze the redundancy between the attributes of a data set, the correlation between various characteristics and tags. We will combined with different

measurement algorithm of attributes to do comprehensive analysis of the data sets, finally choose the optimal feature subset of common, the feature selection algorithm flow chart as shown:

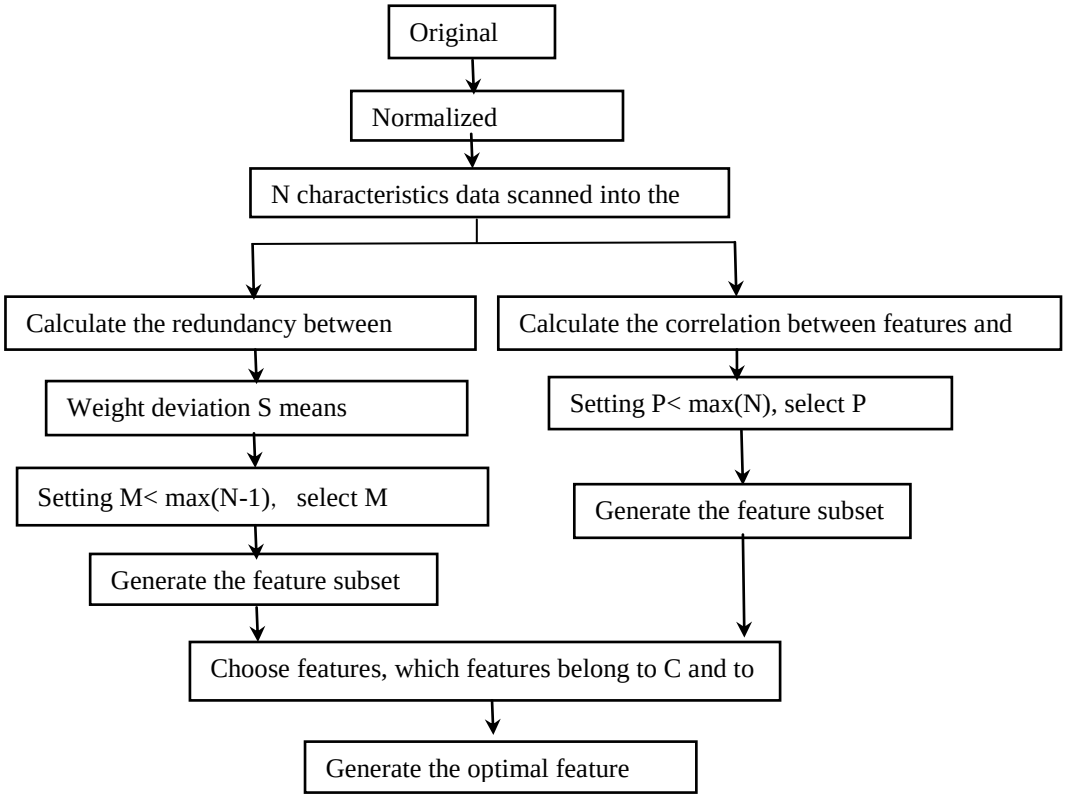


Figure 2. The Paper's Feature Selection Algorithm Flow Chart

The feature selection algorithm is described as follows:

a) The original data should be set normalization and numerical processing firstly, so that the next step is calculated, and the normalization method USES min - Max method, the range of the attribute of all are summarized in the [0, 1].

b) Scan and set the data which has been preprocessed into the Treasury.

c) Calculate all of the features besides tag attributes, use deviation weight method to calculate each characteristic and other characteristics of its own redundancy, The calculation formula is:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - y_i)^2}$$

Where x and y respectively for the two different characteristics, is the amount of data for the data, the value of is bigger, that means the redundancy of two characteristics is greater.

d) For each feature, we choose the biggest value of S to determine the feature about it and put in storage.

e) Setting the value of M, make M<max(N-1), choose M features which appear most frequently and form feature subset C.

f) Calculate the correlation of each feature and tag, the greater the relevance, the more that the feature can identify the data of real property, we calculate correlation of the tag after numerical transformation with each feature. And finally we choose p features and put them into storage, form feature subset C *, where P< max(N).

g) Choose the same feature which belong to set C and belong to set C* and final form feature subset.

4. Experimental Study

4.1. Data Sets

We use KDD CUP99 data set from MIT in our experimental study, which is one of the world recognized off-line intrusion detection data set. This data set contains three sub data set, which are the whole data set, ten percent data set and the test set with correct labels named correct.gz. Specially, we sample 1% and 2% data set from the ten percent KDD CPU99 data set respectively in our experiments, which contains 49402 and 98804 samples in corresponding. There are 41 features in each sample and a label which denotes whether the sample is a normal or attacked, and if it is an attacked data, the label denotes the attack method. The attack can mainly divided into the following four categories.

(1) DoS represents denial of service attack. The attackers make the memory of the computer too busy and cannot handle legitimate requests or refuse to legitimate user's access to the machine.

(2) U2R represents illegal access to the local super user. The attackers access the root permissions using a loophole through a user without permissions or lower permissions, then login and make illegal operations using root.

(3) R2L represents remote user attack. The attackers remote login the computer, then use the account and password to access to the computer and make illegal operations.

(4) Probe represents the port scanning and vulnerability scanning. The attackers detect and search the computers and ports, and collect all kinds of information and system vulnerabilities, then use these information to attack the target.

Each sample of the data set is a vector, which is collected network connection data from the simulated invasions. The last feature of a sample is the label which denotes whether the sample is normal, and the other 41 feature are listed in Label [4-1], in which Feature 1 - 9 are the base attributes, Feature 10 - 22 are the content attributes, Feature 23 - 31 are time flow attributes, Feature 32 - 41 are host flow attributes. Both normal discrete features and continuous features, which are also listed in Label 1.

Table 1. The Attributes of KDD CUP 99 Data

No.	Name	Type	No.	Name	Type
1	duration	continuous	22	is_guest_login	discrete
2	protocol_type	discrete	23	count,srv_count	continuous
3	service	discrete	24	serror_rate	continuous
4	flag	discrete	25	srv_serror_rate	continuous
5	src_bytes	continuous	26	rerror_rate	continuous
6	dst_bytes	continuous	27	srv_rerror_rate	continuous
7	land	discrete	28	same_srv_rate	continuous
8	wrong_fragment	continuous	29	diff_srv_rate	continuous
9	urgent	continuous	30	srv_diff_host_rate	continuous
10	hot	continuous	31	dst_host_count	continuous
11	num_failed_logins	continuous	32	dst_host_srv_count	continuous
12	logged_in	discrete	33	dst_host_same_srv_rate	continuous
13	num_compromised	continuous	34	dst_host_diff_srv_rate	continuous
14	root_shell	discrete	35	dst_host_same_src_port_rate	continuous
15	su_attempted	discrete	36	dst_host_srv_diff_host_rate	continuous
16	num_root	continuous	37	dst_host_serror_rate	continuous
17	num_file_creations	continuous	38	dst_host_srv_serror_rate	continuous
18	num_shells	continuous	39	dst_host_rerror_rate	continuous
19	num_access_files	continuous	40	dst_host_srv_rerror_rate	continuous
20	num_outbound_cmds	continuous	41	dst_host_srv_rerror_rate	continuous
21	is_host_login	discrete	42	Label	discrete

4.2. Pretreatment

The KDD CUP99 data set contains both the normal attributes and the numerical attributes, in which "protocol_type", "service", "flag" and "label" are normal attributes, and the others are numerical attributes.

Because some classifier and algorithms cannot handle normal attributes, we replace them with numerical value. For example, we replace labels with according to the order of them, in particular, we replace the normal label with 0. The rules of replacement are shown as Table 2.

Table 2. Category Numerical Value Correspondence Table

Category	Attack	Corresponding numerical
Normal		0
Dos	Land	1
	Neptune	1
	Smurf	1
U2R	Teardrop	2
	Buffer_overflow	2
	Perl	2
	Rootkid	2
Probe	Satan	3
	Nmap	3
	Mscan	3
R2L	ftp_write	4
	Imap	4
	Spy	4

After the numeralization, we normalize the data set to map all the attributes into [0-1]. The conversion formula we used is:

$$x^* = \frac{x - min}{max - min}$$

4.3. Experiment Results

4.3.1. Evaluation Criterion

In the experimental study, we apply the feature selection algorithm into the intrusion detection systems. In order to evaluate the results of experiments, we use four kinds of evaluation criterion to, which are the average modeling time, the average testing time, the true positive rate (TPR) and the false positive rate (FPR).

4.3.2. Results of Feature Selection

Through combining the largest correlation of text and the minimum redundancy algorithms, we choose the optimal feature subset of KDD CUP99. The chosen feature subset is shown as Table 3, in which the label represents the feature number.

Table 3. The Results of Feature Subset Selection

No.	Feature	No.	Feature
2	protocol_type	32	dst_host_count
3	service	33	dst_host_srv_count
4	flag	34	dst_host_same_srv_rate
5	src_bytes	35	dst_host_diff_srv_rate
23	count	36	dst_host_same_src_port_rate
24	srv_count	37	dst_host_srv_diff_host_rate
25	serror_rate	38	dst_host_serror_rate
26	srv_serror_rate	39	dst_host_srv_serror_rate
29	same_srv_rate	40	dst_host_rerror_rate
30	diff_srv_rate	41	Label
31	srv_diff_host_rate		

Through the Table 3 shows that, in this paper, the characteristics of the final selection has 2,3,4,5,23,24,25,26,29,30,31,32,33,34,35,36,37,38,39,40 and 41 (attributes), a total of 21 characteristics. Next, we will test and analyze the selected feature subset, and we evaluate the feature subset from the accuracy of detection of modeling time and *etc.*, and at the same time, we will verify the desirability of subset in different classifier algorithm.

4.3.3. Results and Analysis

We detected the sub feature set generated by Clonal Algorithm which is wildly used in intrusion detection. Then we compared the results from the selected sub feature set with eh results from the original feature set. In the experiment, we randomly sample 50 thousand samples from the KDD CUP99 data set, and we generate 10 data sets through increase five thousand samples at a time. And the TPRs are shown as Figure 3- Figure 5.

From the Figures above, it can be seen that the results from the selected feature set are basically the same as the original feature set. However, Figure 4 shows that the feature set after text feature selection would lead to the significantly lower TPR. And at the same time, from Figure 5 shows that the selected feature set significantly improved the efficiency, because the process of feature selected has been reduced the redundancy feature and has the lower dimension, which would greatly save the time of modeling. Hence, the sub feature set selected is effective in the same intrusion detection algorithm.

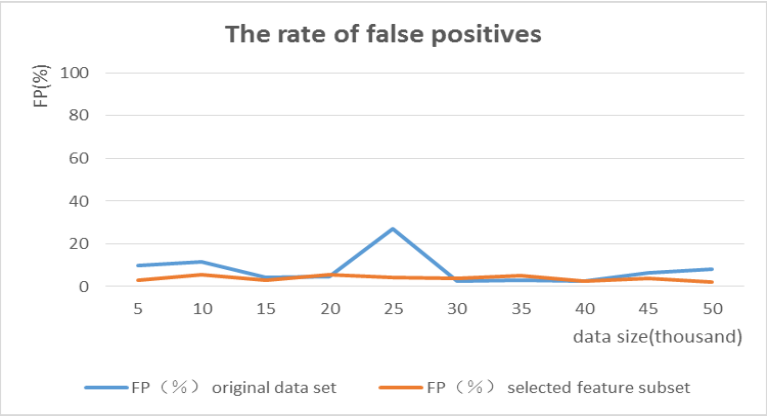


Figure 3. The Contrast Figure of the Rate of FP

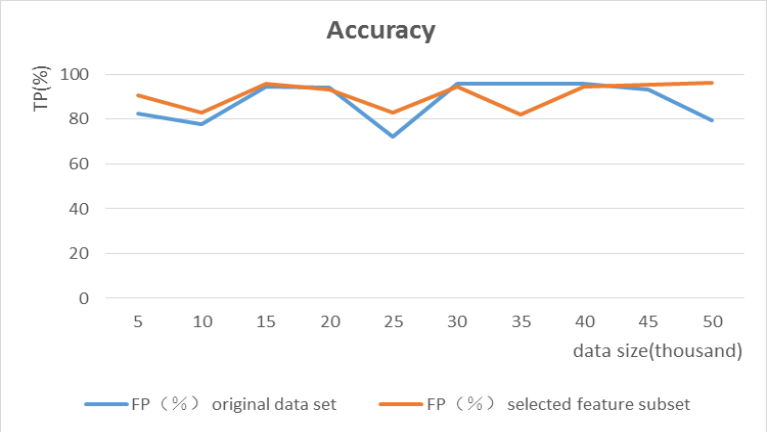


Figure 4. The Contrast Figure of the Rate of TP

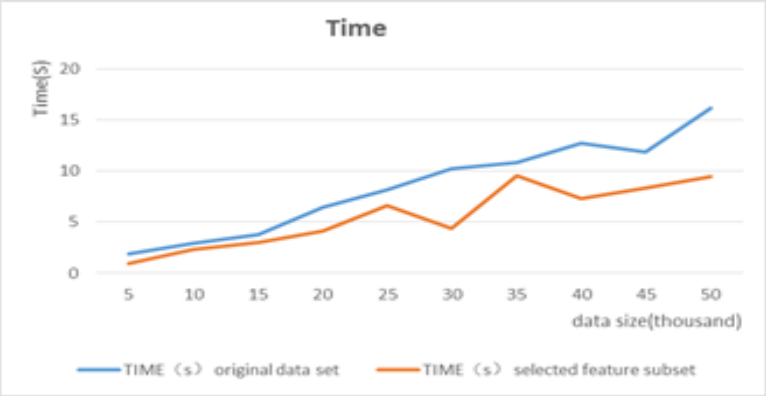


Figure 5. The Contrast Figure of Modeling Time

Besides the above analysis, we also compare the effect of selected sub feature set through different classifiers. And the results are shown as Table 4.

Table 4. Different Classifier for Feature Selection Test Results

classifier	feature selection before/after	Accuracy(%)	Time (s)
Naixe byes	before	92.2	0.73
	after	94	0.55
Lvq1	before	88	0.27
	after	96.1	0.19
J48	before	99.9	4.91
	after	99.9	1.77
PART	before	99.8	6.95
	after	99.9	3.27
Clonal	before	79.5	16.8
	after	78.4	13.84

We can see from the Table 4, when we test the data set after feature selection in different classifier, the detection precision is not lower too much, and in some classifiers, the precision of the subset is superior to that of the original data set. And at the same time, we can see from the table above, after feature selection, the data in the test model of reduce a lot of time, it is of great significance to improve the real-time performance of intrusion detection.

5. Conclusions

In this paper, we combines the weight measured by deviation minimum redundancy and the biggest correlation analysis between attributes and label attributes were selected, and using KDD CUP99 data set as test data sets, we analyze and get the optimal feature subset. Then, we respectively to detect the accuracy and the rate of false positives and detection time as evaluation standard, according to the experiment, we think that the subset is desirability. We further apply the subset of features which we choose to different classification, from the actual classification accuracy and time we come to the conclusion that the feature subset is optional.

Acknowledgments

This paper is a revised and expanded version of a paper entitled “A Hybrid Feature Selection Algorithm” presented at AITS 2015, Harbin, China, August 21-23, 2015. This work was funded by the National Natural Science Foundation of China (61373134, 61402234), and by the Industrial Strategic Technology Development Program (10041740) funded by the Ministry of Trade, Industry and Energy (MOTIE) Korea. It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (No.KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). Prof. Jin Wang is the corresponding author.

References

[1] E. Backer and J. Schipper, "On the max-min approach for feature ordering and selection", The Seminar on Pattern Recognition, **(1977)**.
[2] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection", Pattern recognition letters, vol. 10, no. 5, **(1989)**.
[3] C. Yin, "Towards Accurate Node-Based Detection of P2P Botnets", The Scientific World Journal, 2014, **(2014)**.
[4] C. Yin, M. Zou, D. Iko and J. Wang, "Botnet Detection Based on Correlation of Malicious Behaviors", International Journal of Hybrid Information Technology, vol. 6, no. 6, **(2013)**.
[5] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman and S. Li, "Incremental learning for ν-Support Vector Regression", Neural Networks, vol. 67, **(2015)**.
[6] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano and S. Li, "Incremental Support Vector Learning for Ordinal Regression", **(2014)**.
[7] U. EHSC, "An Efficient and Intelligent Intrusion Detection and Response System using Virtual Private Networks", Firewalls and Packet Filters, Channels, 200.
[8] X. C. Yuan and C. M. Pun, "Geometrically invariant image watermarking based on feature extraction and Zernike transform", Int J Secur Appl, vol. 6, no. 2, **(2012)**.
[9] S. Suthaharan and T. Panchagnula, "Relevance feature selection with data cleaning for intrusion detection system", Southeastcon, 2012 Proceedings of IEEE, **(2012)**.
[10] F. Zhang and D. Wang, "An effective feature selection approach for network intrusion detection, Networking", Architecture and Storage (NAS), 2013 IEEE Eighth International Conference on, **(2013)**.
[11] Y. Cui and N. Xie, "A Intrusion Detection Method Based on Feature Selection", Jilin University Journals: Neo-confucianism Edition, vol. 53, no. 1, **(2015)**.
[12] G. H. John, R. Kohavi and K. Pfleger, "Irrelevant features and the subset selection problem", Machine Learning: Proceedings of the Eleventh International Conference, **(1994)**.
[13] D. Koller and M. Sahami, "Toward optimal feature selection", **(1996)**.
[14] M. Dash and H. Liu, "Feature selection for classification, Intelligent data analysis", vol. 1, no. 1, **(1997)**.
[15] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm", AAAI, **(1992)**.
[16] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection", Computers, IEEE Transactions on, vol. 100, no. 9, **(1977)**.
[17] J. Wang, J. U. Kim, L. Shu, Y. Niu and S. Lee, “A distance-based energy aware routing algorithm for wireless sensor networks”, Sensors, vol. 10, no. 10, **(2000)**.

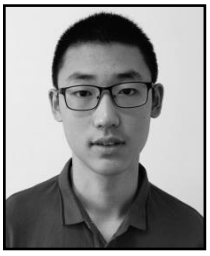
[18] J. Wang, Y. Yin, J. Zhang, S. Lee and R. S. Sherratt, "Mobility based energy efficient and multi-sink algorithms for consumer home networks", IEEE Transactions on Consumer Electronics, vol. 59, no. 1, **(2013)**.

Authors

Chunyong Yin, Dr. Chunyong Yin is currently an associate Professor and Dean with the Nanjing University of Information Science & Technology, China. He received his Bachelor (SDUT, China, 1998), Master (GZU, China, 2005), PhD (GZU, 2008) and was Post-doctoral associate (University of New Brunswick, 2010).He has authored or coauthored more than twenty journal and conference papers. His current research interests include privacy preserving and network security.

Luyu Ma, She received her BE degree in network engineering from Nanjing University of Information Science & Technology, China, in 2013. Currently she is a graduate student at the School of Computer and Software of Nanjing University of Information Science & Technology. Her research interests are in network security and intrusion detection.

Lu Feng, received his bachelor degree in 2013 from Nanjing University of Information Science & Technology. His research interests are data-stream classification and feature extraction algorithm.



Zhichao Yin, is studying in Nanjing No.1 Middle School. His current research interests include network security and mathematical modeling.

Jin Wang, received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor in the Computer and Software Institute, Nanjing University of Information Science and technology. His research interests mainly include routing method and algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.