International Journal of Future Generation Communication and Networking
Vol. 8, No. 2 (2015), pp. 213-224
http://dx.doi.org/10.14257/ijfgcn.2015.8.2.16

The Effectiveness Study of ML-based Methods for Protocol Identification in Different Network Environments

Zhang Luoshi[1], Xue Yibo[2] and Wang Dawei[3]

[1]*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*
[2]*Research Inst. of Info. & Tech., Tsinghua University, Beijing 100084, China*
[3]*Tsinghua National Lab for Information Sci. & Tech., Beijing 100084, China*
[3]*National Computer Network Emergency Response Technical Team / Coordination Center of China (CNCERT/CC), Beijing, 100029, China*
luoshi.zh@gmail.com, yiboxue@tsinghua.edu.cn, stonetools@yeah.net

Abstract

Due to the wide use of encrypted protocols and random ports, traditional methods that based on port number or packet payload have gradually lose their effectiveness. To address this issue, new methods that based on machine learning techniques become the research hotspots. With many further studies, some research institutions show that ML-based protocol identification methods can generally achieve over 95% accuracy. However, different from most research studies, industry claims that ML-based techniques are hardly to be deployed for practical use due to their high false positives and false negatives. In this paper, different Machine Learning techniques are evaluated for the actual accuracy under different network environments, and a variety of features are tested on different encrypted protocols. The results show that the identification accuracy will go down due to the changed network scale and network environment while the same ML-based models are used under different network environments, and the choices among different Machine Learning techniques, protocol types or statistical features are not critical.

Keywords: *Protocol Identification, Traffic Classification, Machine Learning, Accuracy*

1. Introduction

Regardless of network management, optimization or security, it is a key essential step to accurately identify the network traffic protocol. The traditional protocol identification methods are mainly categorized into port-based and payload-based ones. The former use the known port of IANA to determine the network traffic protocol. However, they are losing effect due to widely used technology such as random port and P2P [1]. The latter use the fixed field or pattern in the packet payload as the signature of the protocol and have higher accuracy. However, with the popularity of technology such as protocol encryption and traffic obfuscation, the fixed signature is hidden, and they are also losing effect. In order to effectively solve problems of protocol identification, machine learning methods have been introduced into the field of protocol identification.

The protocol identification method based on machine learning takes the network traffic information at packet level (such as packet size and packet arrival interval) and flow level (such as the number of all packets or the average length of packets in one flow) as the classification feature to map the network traffic into different protocol categories. In the past decade since 2003 when Early *et al.*, [2] for the first time used the decision tree algorithm to effectively distinguish various applications

ISSN: 2233-7857 IJFGCN

of network traffic, a large number of machine learning methods and theories have been applied to the field of protocol identification and intensively researched with a lot of findings achieved, and the accuracy has been increased from about 70% to 95% and above. Thus, generally speaking, the machine learning method is considered the most effective solution to identify the encryption protocol, proprietary protocol, P2P protocol and other complex protocols. Protocol identification technology based on machine learning method has become one of the hottest areas of research.

However, the existing protocol identification technology based on machine learning mainly has a premise that all kinds of samples in the training set are basically balanced in number, and the classifier trained on this premise has a better effect to classify the balanced test data in the laboratory environment. However, the real network traffic is imbalanced data. Therefore, as the classifier trained in balanced training set is used for real network, its accuracy drops greatly. In addition, the training samples are generally obtained from the fixed network environment, so the same classifier has lower ability to adapt to different network environment. It is susceptible to interference and obfuscation and has serious false positives and false negatives.

At present, most researches focus on the improvement of identification accuracy rather than the analysis of adaptability and accuracy of the machine learning method in different network environment, so they cannot solve the problem of actual deployment of the protocol identification method based on machine learning. To this end, this paper first analyzed the accuracy of protocol identification of different machine learning methods in the same network environment and then on this basis extended to the analysis of changes in the accuracy of protocol identification methods based on machine learning in different network environment. By studying changes in the accuracy of machine learning methods in laboratory network environment, campus network environment and backbone network environment, this paper analyzed the adaptability of machine learning methods in the real network environment and further analyzed the reasons for the changes. In order to ensure the credibility of the experimental results, the paper selected the most common network traffic features and the most commonly used machine learning methods in the field of protocol identification. Through the experiment, it is confirmed that with the existing protocol identification methods based on machine learning, the accuracy of protocol identification methods in the same network environment depends on the sources of training set and different machine learning methods, and there is small difference between different machine learning methods. But in different network environment, the accuracy of protocol identification of the same machine learning training model changes significantly as the network environment varies, and it is roughly inversely proportional to the size of the network environment.

The main innovative points of this paper are as follows:

- It studies the accuracy of protocol identification of machine learning methods in the same network environment and analyzes the changes in the accuracy of protocol identification in the case of different machine learning methods, encryption protocols and statistical features;
- It studies the accuracy of machine learning methods in the network environment with different sizes, locations and application conditions and analyzes the changes in the identification accuracy of the encryption protocol based on the same training data in different network environment with different machine learning methods;
- It effectively analyzes the reasons for changes in the accuracy of the protocol identification method based on machine learning in different network

environment and explains the root causes of the changes in the accuracy in terms of network size, statistical features and training mode.

This paper is organized as follows: Section 2 describes the relate work of protocol identification technology based on machine learning; Section 3 contains the data source, experimental method and evaluation criteria; Section 4 shows the experimental results and the main findings; Section 5 summarizes and discusses the next research work.

2. Relate Work

As the traditional protocol identification methods based on port and payload gradually lost effect [1], machine learning methods were introduced to effectively solve the problem of encryption protocol identification. In 2003, Early et al. first started the research on protocol identification technology based on machine learning [2].

At present, the machine learning algorithms commonly used for protocol identification mainly contain supervised, semi-supervised and unsupervised types [3]. The unsupervised machine learning algorithm utilizes the similarity between network traffic features to cluster network traffic into different categories. Because these categories have not been marked in advance, it is difficult in practical use to determine all the real categorized protocols and their accuracy. Therefore, such algorithms have been seldom used. Supervised and semi-supervised machine learning algorithms are now popular, and their common feature is to rely on the marked protocol samples as the training set and generate a classifier with the machine learning method (Naive Bayes, SVM, C4.5, *etc*.) [4-5] to identify the protocol of the network traffic. Therefore, the coverage of the training set and effective selection of features are the basis for the accuracy of a model.

In order to further improve the accuracy of protocol identification methods based on machine learning, Moor *et al.*, [6] summarized 248 classes of features in the network traffic and identified some protocol from traffic by using Naive Bayes algorithm. Meanwhile, Williams *et. al*. made an effective evaluation on the performance and accuracy of five different machine learning algorithms in terms of network protocol identification [7].

Karagiannis *et. al*. [8] proposed the new method that called “BLINC”, which collected the statistical features from multiple levels used for making the traffic behavior of protocols. The TAGs [9] and mixed TAGs [10] methods has been proposed by Y. Jin, which analyzed communication pattern of multiple protocols in the large-scale network by using graph theory. Xi Liang *et. al.*, analyzed recent advance about immunity-based intrusion detection system (IIDS) and promoted the conversion of the theoretical fruits to applications [11].

Yang Baohua *et. al.* [12] proposed “SMILER” method and used the semi-supervised machine learning algorithm as well as the payload length of the first 5 packets to identify multiple protocols and achieved high accuracy. And based on active learning and SVM algorithms, Wang Yipeng *et. al*. [13] achieved the classification of unknown network protocol traffic only depending on the payload information in untreated network traffic. Dong Hui *et al.*, proposed a new method based on link homophile to classify application layer traffic without the payloads and properties at the flow level, and achieved 80% accuracy [14].

3. Experimental Method

The main purpose of this paper is to explain the changes in the accuracy of a classifier trained with the same machine learning algorithm and applied to different network environment. Therefore, the data sets representing different network sizes, the fixed algorithm configuration and the commonest statistical features were chosen in order to ensure the credibility and reproducibility of the results.

This section describes the data sets and statistical features used for accuracy evaluation, the machine learning algorithms used and relevant evaluation criteria.

3.1. Data Sources and Protocol Categories

In this paper, three data sets from different network environment (Trace-A, Trace-B and Trace-C) were mainly used.

The data set Trace-A contains 4G network traffic collected from the laboratory, including pure traffic of encryption protocols to be tested and background traffic for model training. It represents a small laboratory network.

The data set Trace-B contains 200G 12-hour network traffic obtained from a college gateway. It represents a campus LAN.

The data set Trace-C contains 800G 1-hour network traffic collected from the national backbone network gateway. It represents an Internet WAN.

In order to identify the protocol, the network packet is usually divided into a number of different network traffic according to the five-tuple (<source IP, source port, destination IP, destination port, protocol>), and the bi-directional network traffic is divided into two unidirectional ones to extract the statistical features.

At present, the protocol identification methods based on machine learning mainly apply scenarios to identify the encryption protocol. However, the proprietary encryption protocol itself can not be identified with the traditional protocol identification method, so it is difficult to use non-machine learning methods to validate the results of data. Therefore, this paper selected the general SSL protocol (port 443) [15] with a fixed port number and strong load features as the research object.

3.2. Algorithms and Feature Selection

At present, the most commonly used supervised machine learning algorithms in the field of protocol identification mainly include Naive Bayes, SVM and Decision Tree Model.

This paper selected four algorithms: BayesNet, NaiveBayes, LibSVM and J48 of Weka (V3.7) [16] as the basis for research. In addition, all parameters were defaults.

For feature selection, the most commonly used features at present include packet length and the interval between two consecutive packets. There are a large number of papers demonstrating the effectiveness and feasibility of these two features. Therefore, this paper selected four different statistical features accordingly:

(1) The payload length of the first 5 consecutive packets in the same network flow;

(2) The payload length of the first 10 consecutive packets in the same network flow;

(3) The interval of the first 5 consecutive packets in the same network flow;

(4) The interval of the first 10 consecutive packets in the same network flow.

The load length refers to that of the data packet with IP header and TCP header removed, and the accuracy of the data packet interval is at the microsecond level in order to guarantee a strong discriminatory power.

3.3. Evaluation Criteria

The accuracy evaluation of the protocol identification method based on machine learning mainly depended on the following four indicators:

(1) Accuracy: the ratio of samples correctly classified by the classifier and total samples in the given test data set.

(2) Precision: the proportion of the correct samples in those correctly classified as a category in a given test data set, which is used to measure the model's precision capability, commonly represented with P.

(3) Recall: the proportion of the correctly classified samples in a particular category of correct samples in a given test data set, which is used to measure the model's recall capability, commonly represented with R.

(4) F1-Measure shows the weighted average of precision and recall and comprehensively measures the identification accuracy, as follows:

$$F_1 = \frac{2 * PR}{P + R} \tag{1}$$

These four indicators effectively evaluate the identification effect of different machine learning algorithms and different statistical features and better show the usability of algorithms.

4. Results and Analysis

Research institutions often select the same data set for model training and test and obtain higher accuracy. While they productize it, they are bound to apply the same machine learning model to different network environment. On the one hand, serious imbalance problem may be caused between the training flow set with class-balance and the actual traffic with class-imbalance. As a result, it affects the identification accuracy of specific protocol; on the other hand, the variances in the network traffic structure in different network environment may cause the situation of unmatched classifier.

Therefore, this section based on the above two cases designed and implemented two different experiments in order to demonstrate the accuracy of the protocol identification method based on machine learning and carried out research and analysis of the change in accuracy and reasons.

4.1. Experiment 1: Accuracy in the Same Data Set

Generally, k-fold cross-validation is used to evaluate the accuracy of the protocol identification method based on supervised machine learning.

K-fold cross-validation [17] is to divide the original data into k parts randomly, select one part each time as the test data and the remaining k-1 parts as the training data, and cycle for k times in order to solve the problem of insufficient data samples, and k is often 10.

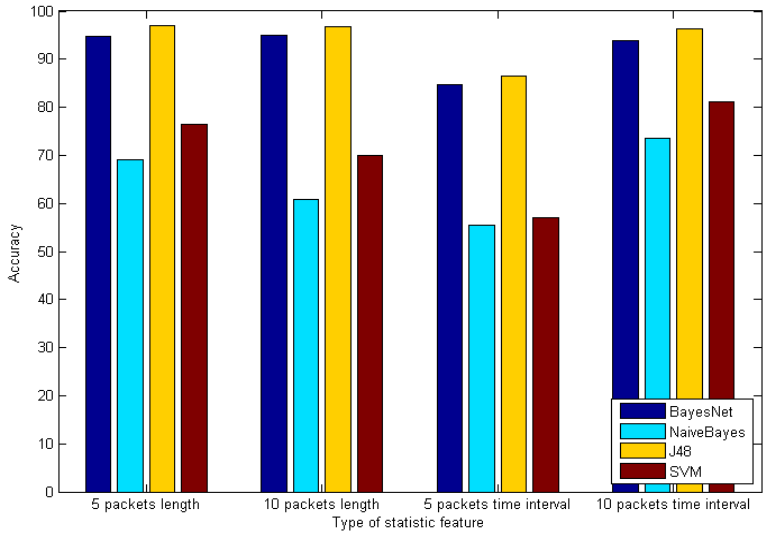
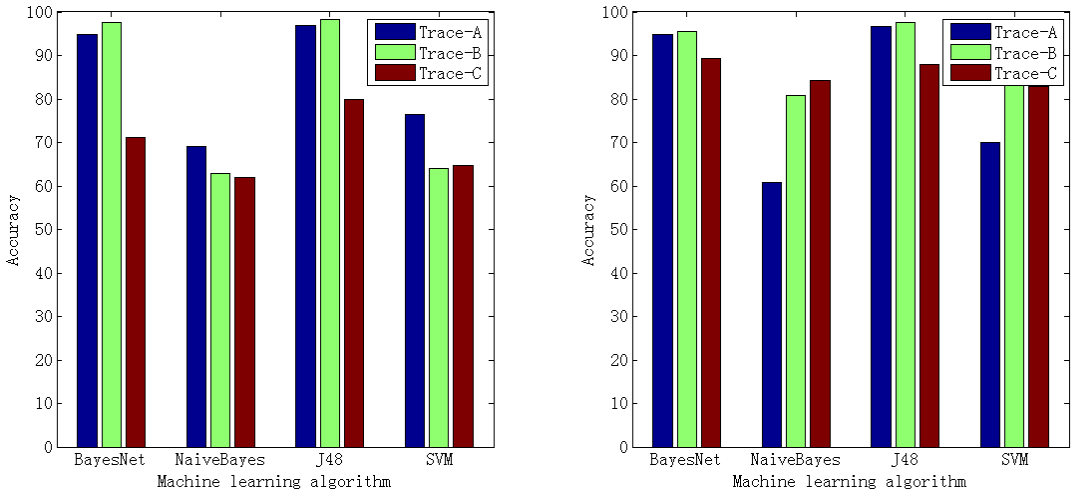


Figure 1. The Comparison of Identification Accuracy (Trace-A)

Figure 1 shows the result of 10-fold cross-validation of SSL protocols in Trace-A, and the training set and the test set are from the same data set. As can be seen from the figure, different machine learning methods and statistical features are used for the same data set with high accuracy both. The accuracy of BayesNet and J48 algorithms is even about 90%.



(a) The length of the first 5 packets (b) The length of the first 10 packets

Figure 2. The Comparison of Identification Accuracy between Different Datasets

Similarly, three data sets representing different network environment are used for 10-fold cross-validation and both the training set and the test set each time are from the same data set. As shown in Figure 2, despite different statistical features, as the same machine learning method is used, the accuracy of protocol identification of the three data sets is substantially unchanged.

This result is basically the same as those of most papers. It proves that when training set samples and test set samples are from the same data set, the protocol identification methods based on machine learning have high identification accuracy and depend on the adopted machine learning algorithm and statistical features.

4.2. Experiment 2: Accuracy in Different Data Sets

In the first experiment, the data samples are selected from the same network environment for training and test of machine learning models, while in this experiment, the changes in accuracy of the same machine learning model applied to different network environment are analyzed and studied.

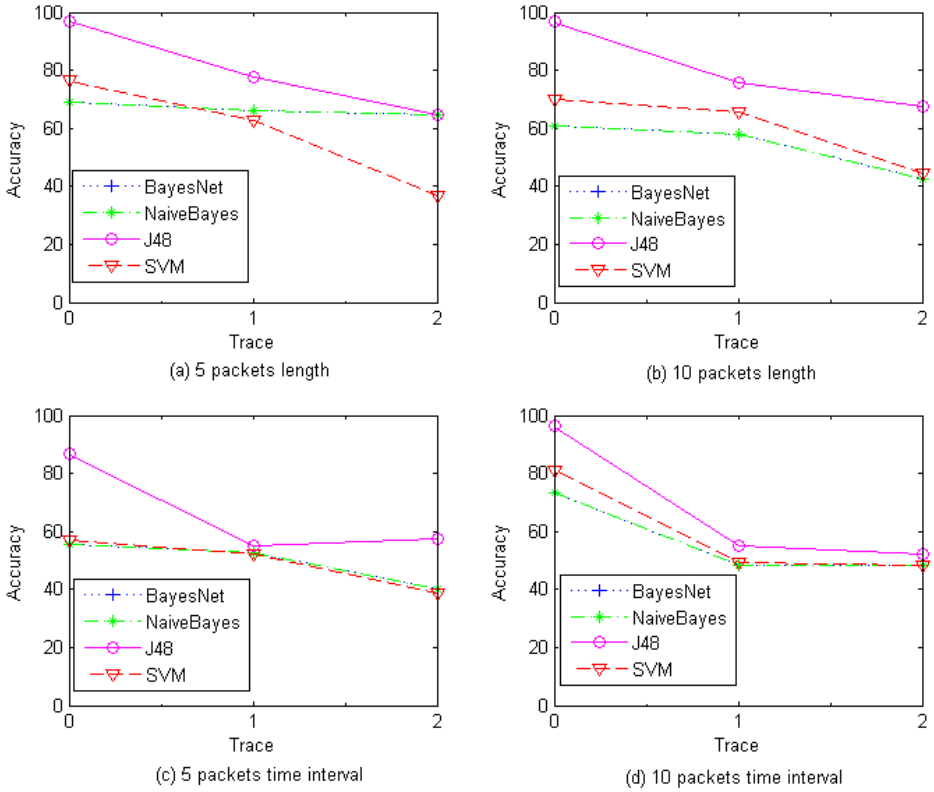


Figure 3. The Comparison of Accuracy with Different Datasets in the Same Model (1)

Figure 3 shows the changes in accuracy of the machine learning model trained in Trace-A and used in the other two data sets. The abscissa values 0, 1 and 2 are the small network represented by Trace-A, the campus LAN represented by Trace-B and the backbone network represented by Trace-C, and the number of protocol samples to be tested in the test set is substantially the same as that of samples of background traffic.

It can be found from the figure that regardless of any algorithm or statistical feature, as the same machine learning model is used for other network environment, the accuracy decreases to some extent. But it can be further found in the comparison of identification accuracy of Trace-B and Trace-C that the accuracy changes slightly. This indicates that in the case of the same proportion of positive and negative examples, the changes in identification accuracy mostly result from the differences between samples of test set and training set.

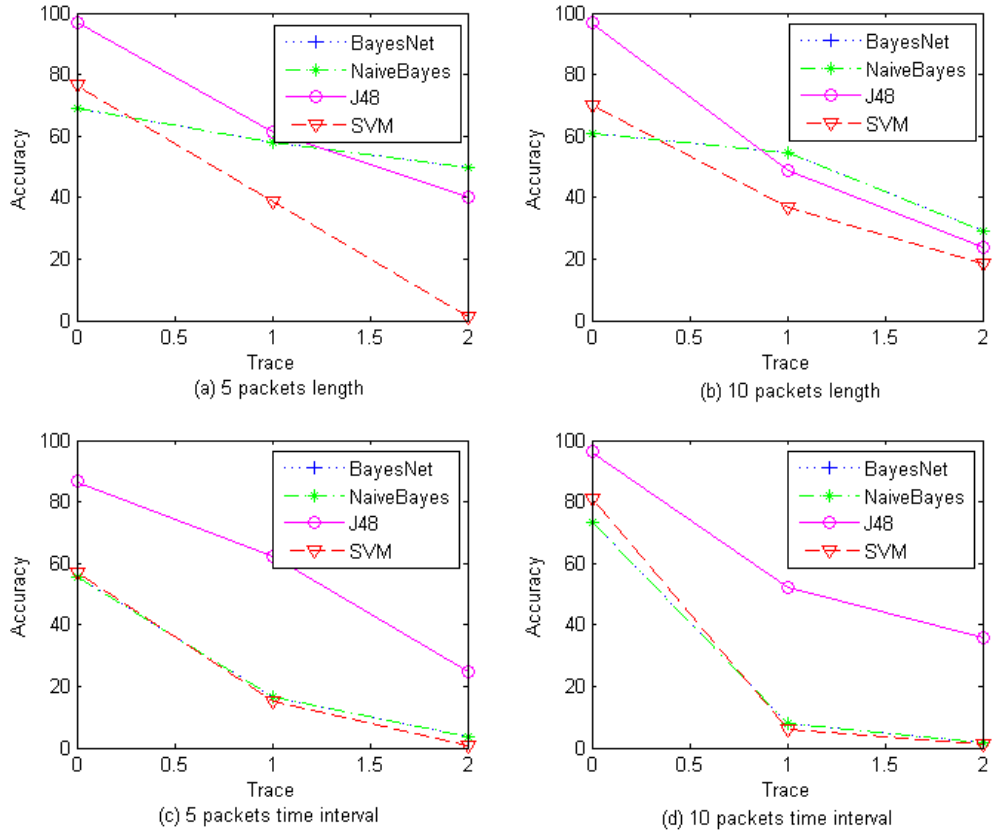


Figure 4. The Comparison of Accuracy with Different Datasets in the Same Model (2)

Furthermore, if the number of samples of background traffic is not restricted, as shown in Figure 4 when the machine learning model trained in Trace-A is used for other network environment, the accuracy decreases significantly as the network size increases. Even in the backbone network, the accuracy of some machine learning algorithms has dropped below 5%. This indicates that the identification accuracy of the same machine learning model in different network environment is largely associated with the network size.

4.3. Analysis of Experimental Results

As can be seen from the results of both experiments, when the data sources of training and test sets belong to the same network environment and the data with class-balanced are selected, the accuracy is high, but when one of the data sets is used as the training set and another data set as test set, the accuracy declines significantly and even affects the usability of the algorithm.

Based on further analysis of the data and results, it is believed that the reasons for the decrease in accuracy are as follows:

(1) The increasing number of false positives caused by the increasing size of background traffic

Although the decrease in accuracy in the experiment is different on the basis of different machine learning methods and statistical features, the decrease trend is unchanged. Therefore, it proves that the decrease in accuracy will not occur due to changes in algorithms and the statistical features.

Meanwhile, the evaluation criteria for the results of both experiments is accuracy, that is, the correctly classified samples include test protocol samples as positive

examples and background samples as negative examples. Table 1 shows the comparison of accuracy of the identification model trained in Trace-A to respectively identify specific protocols in Trace-C and Trace-A. The value in brackets indicates the identification result of k-fold cross-validation of Trace-A.

Table 1. Identify Results for SSL Protocol with Different Algorithm

Algorithm	Accuracy	Precision	Recall
BayesNet	42.56% (94.77%)	0.1% (94.5%)	74.6% (95%)
NaiveBayes	49.64% (69.04%)	0.1% (79.5%)	57.1% (50.3%)
J48	39.97% (96.99%)	0.1% (96.8%)	73.2% (97.1%)
SVM	1.23% (76.48%)	0.0004% (74.7%)	95% (79.3%)

It can be found from the result that when the accuracy drops significantly, the precision decreases more greatly, but the recall shows no significant change. This indicates that the decrease in accuracy is mainly due to the sharp increase in the number of false positives samples.

Different network environment have significant differences in the composition of network traffic protocol. As the network size increases, the number of network protocols increases sharply. It can seriously compress the proportion of each protocol in the total traffic. As a result, the background traffic has an increasing size. If the data is not filtered but all taken as the test set, significant false positives will produce.

A false positive usually refers to the credibility of the results tested as positive, which can be calculated with Bayesian inference, as follows:

$$P(A|B) = P(A)\frac{P(B|A)}{P(B|A)P(A) + P\left(B\middle|\overline{A}\right)P\left(\overline{A}\right)} \quad (2)$$

where, P (A) represents the proportion of positive examples in the total samples of the test set, P(B|A) represents the value of TP Rate, *i.e.*, the proportion of positive samples determined as positive examples in the test set, and represents the value of FP Rate, *i.e.*, the proportion of negative samples determined as positive examples in the test set, as shown in Table 2.

Table 2. The Identify Result with Different Proportion of Positive and Negative Cases

Positive : Negative	TP Rate	FP Rate
1:1	0.732	0.408
1:10	0.732	0.619
1:100	0.732	0.631
1:1000	0.732	0.626

Figure 5 shows the credibility of the result obtained from Formula 2, F1-Measure and accuracy comparison when the machine learning model trained in the same data set is used and the proportion of positive and negative training examples in the test set changes.

It shows that the theoretical calculation result has striking similarity to the real F1-Measure in terms of credibility. It indicates that the decrease in accuracy of the protocol identification method based on machine learning in different networks is largely associated with the proportion of positive and negative examples in the test set: as the proportion of positive and negative examples in the test set increases gradually, a serious problem of false positives will reduce the credibility and usability of the identification result and show a sharp decline in accuracy and precision.

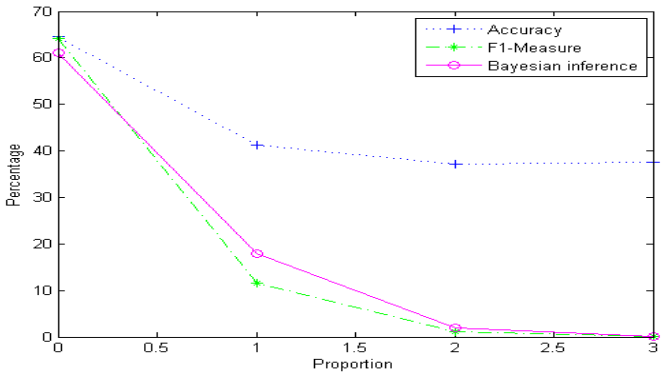


Figure 5. The Comparison of F1-Measue and the Result of Bayesian Inference

(2) The features used result in a problem of feature coverage.

However, it still has to be noted that the protocol identification method based on port or load features shows no significant false positives as the proportion of positive and negative examples increases. This is not consistent with the result based on Bayesian inference.

In fact, the protocol identification features most commonly used at present are usually from the protocol specification or special regulations on protocol design. Therefore, its effective discriminatory power from other protocols is taken into account in design, which indirectly ensures the strong discriminatory power of identification features.

However, the statistical features of the protocol are not a pre-set but produced based on the access content, the network environment and even the user's habits in the actual use of protocol. This makes it difficult to evaluate the discriminatory power of statistical features.

On the one hand, because the machine learning method is mainly to solve the identification problem of encryption protocols, and the training samples of such protocols often cannot be automatically or artificially extracted from the real network environment, limited training samples can only be obtained with repeated and frequent use of identification protocols in the pure environment. But this automated feature extraction method may lead to repeatability of traffic features and cannot reflect the true scope of statistical features of protocol, which may reduce the feature coverage with the changing network environment, as shown in Figures 3 and 4.

On the other hand, with changing network size, number of users, access habits and protocol composition, the network structure has significant changes. The current statistical features have limited coverage. For example, the maximum feature range

of the packet length is [0, 1480], which tends to cause more feature conflicts with the increasing size of background traffic, leading to an increase in the number of false positives.

Thus, it can be seen from the above experiments and analysis: as the classifier trained with the same machine learning algorithm is used for different network environment, the identification accuracy of a particular protocol will decline sharply with the increase in the network size. This is slightly associated with the selected machine learning method and statistical features but largely associated with the proportion of positive and negative examples in the test set and has something to do with the collection method and coverage of statistical features.

5. Conclusion

Protocol identification is the basis for network security and network management. However, with the full use of encryption protocols and proprietary protocols, the traditional protocol identification method in the face of new problems appears to be inadequate. In order to effectively solve the problems, the protocol identification method based on machine learning gradually becomes a research focus with a wealth of research results. But, it encounters significant decrease in accuracy when shifting from the laboratory research to practical application. Therefore, this paper shows the case of a decrease in accuracy and summarized three reasons as follows:

(1) False positives caused by data with imbalanced categories;

(2) Feature coverage caused by random under-sampling in the collection of training set;

(3) Feature conflict caused by the complexity of the network environment.

In future work, solutions will be further researched based on the reasons for the decline in accuracy found in this paper to improve the usability of machine learning methods in the field of protocol identification so as to overcome the problems of industrialization of identification methods based on machine learning.

Acknowledgements

This work was supported by the National Key Technology R&D Program of China under Grant No.2012BAH46B04.

References

[1] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos, “Is p2p dying or just hiding?”, Proceedings of Global Telecommunications Conference, vol. 29, no. 3, **(2004)** November- December, Dallas, USA.

[2] E. James, C. Brodley, and C. Rosenberg, “Behavioral authentication of server flows”, Proceedings of the 19th Annual Computer Security Applications Conference, **(2003)** December 8-12, Las Vegas, NV, USA.

[3] T. M. Mitchell, “Machine Learning”, McGraw-Hill Education, **(1997)**.

[4] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, “Flow clustering using machine learning techniques,” Proceedings of Passive and Active Measurement Workshop, **(2004)** April 19-20, Antibes Juan-les-Pins, France.

[5] T. Auld, A. W. Moore and S. F. Gull, “Bayesian neural networks for Internet traffic classification”, IEEE Trans. Neural Networks, vol. 1, no. 223, **(2007)**.

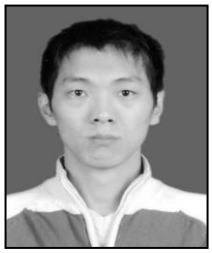
[6] W. M. Andrew and D. Zuev, “Internet Traffic Classification Using Bayesian Analysis Techniques,” Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, **(2005)** June 6-10, New York, NY, USA.

[7] N. Williams, S. Zander, and G. Armitage, “A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification, ACM SIGCOMM Computer Communication Review, vol. 5, no. 36, **(2006)**.

[8] K. Thomas, K. Papagiannaki, and M. Faloutsos, “BLINC: multilevel traffic classification in the dark, Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications, **(2005)** August 21-26, Philadelphia, PA, USA.

[9] J. Yu, E. Sharafuddin, and Z. Zhang, "Unveiling core network-wide communication patterns through application traffic activity graph decomposition", ACM SIGMETRICS Performance Evaluation Review, vol. 1, no. 37, **(2009)**.
[10] Y. Jin, N. Duffield, P. Haffner, S. Sen, and Z. Zhang, "Can't See Forest through the Trees?" Understanding Mixed Network Traffic Graphs from Application Class Distribution. Proceedings of the 9th Workshop on Mining and Learning with Graphs, **(2011)** August 20-21, San Diego, CA.
[11] X. Liang and Z. Fengbin, "Recent Advances and Prospects of Immunity-based Intrusion Detection Systems", Journal of Harbin University of Science and Technology, vol. 2, no. 19, **(2014)**.
[12] B. Yang, G. Hou, L. Ruan, Y. Xue, and J. Li, "SMILER: towards practical online traffic classification", Proceedings of the 2011 ACM/IEEE Seventh Symposium on Architectures for Networking and Communications Systems, **(2011)** October 3-4, Brooklyn, NY, USA.
[13] W. YiPeng, Y. XiaoChun, Z. YongZheng and L. ShuHao, "Network protocol identification based on active learning and SVM algorithm", Journal of Communications, vol. 10, **(2003)**.
[14] D. Hui, S. Guanglu, L. Dandan, and X. Feng, "Application Layer Traffic Classification Based on Link Homophily", Journal of Harbin University of Science and Technology, vol. 4, no. 18, **(2013)**.
[15] IANA, http://www.iana.org/
[16] Weka, http://www.cs.waikato.ac.nz/ml/weka/
[17] R. D. Juan, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation", Pattern Analysis and Machine Intelligence, vol. 3, no. 32, **(2010)**.

Authors



Zhang Luoshi was born in 1983. He is a Ph.D. candidate at School of Computer Science and Technology, Harbin University of Science and Technology. His research interests include networking security, protocol identification and traffic management, etc.

Xue Yibo was born in 1967. He received his M.S. and Bachelor degree in computer science and engineering at Harbin Institute of Technology in 1992 and 1989 respectively, and Ph.D. degree in computer architecture at Institute of Computing Technology, Chinese Academy of Sciences in 1995. Now he is a professor at Research Institute of Information Technology, Tsinghua University. He is a senior member of CCF and a member of IEEE/ACM. His research interests include computer network and information security, parallel processing and distributed system. He has published more than 100 papers in journals and conferences and applied for more than 30 patents

Wang Dawei was born in 1982. He received his B.S. degree in Computer Science from Nanjing University of Post and Telecommunication in 2005, and got his M.S and Ph.D. degree in Computer Science from Harbin University of Science and Technology in 2007 and 2010, respectively. Now he is an engineer of National Computer network Emergency Response technical Team/Coordination Center of China. His research interests include intrusion detection, traffic management, statistical pattern recognition and artificial immune system.

