Feature Selection Algorithm Based on Correlation between Muti Metric Network Traffic Flow Features

Yongfeng Cui[1,2], Shi Dong[1,2,3], and Wei Liu[2]
[1]School of Computer Science and Technology, Huazhong Universtiy of Science and Technology, China
[2]School of Computer Science and Technology, Zhoukou Normal University, China
[3]Department of Computer Science and Engineering, Washington University in St Louis, USA

Abstract: *Traffic identification is a hot issue in recent years, in order to overcome shortcomings of port-based and Deep Packet Inspection (DPI), machine learning algorithm has gained wide attention, but nowadays research focus on traffic identification based on full packets dataset, which would be a great challenge to identify online traffic flow. It is a way to overcome this shortcoming by considering the sampled flow records as identification object. In this paper, flow records NOC_SET is constructed as dataset, and inherent NETFLOW and extended flow metrics are regarded as features. This paper proposes feature selection algorithm MSAS to select features with high correlation. And classical machine learning algorithms are used to identify traffic. Experimental results show that machine learning flow identification algorithm based on sampled flow records has almost the same identification results as method based on full packets dataset, and the proposed feature selection algorithm MSAS can improve the result of application identification.*

Keywords: *Port identification, deep packet inspection, netflow flow, machine learning.*

Received February 5, 2014; accepted April 13, 2015

1. Introduction

With the increasing of network bandwidth, network behavior patterns become increasingly complex; produce a variety of new network applications, more and more attentions are being highly focused on network traffic identification in the fields of network management. Traditional methods of application identification can be categorized into three types: port-based [18], Deep Packet Inspection (DPI) [1] and machine learning [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15]. P2P applications generally use random dynamic port, so port-based method is not applicable for p2p applications, while DPI method needs to know application signatures based on full packet, it can't identify encrypted network applications, it is difficult to be applied to high-speed network. Machine learning methods adopt the flow behavior characteristics to identify network traffic. So it can identify encrypted network traffic without requiring packet information. However, nowadays many researchers still identify traffic by analyzing full payload packets information, it is a big obstacle to online traffic identification, so this paper studies flow record and extended flow records based on NETFLOW sampling method, and deeply analyzes some network flow characteristics and considers them as feature metrics. We take into account the impact of correlation between features and application categories on classification algorithm, propose MSAS feature selection algorithm, and use C4.5 and NBK to identify network traffic. Experiment results show that MSAS can improve the classification and identification results, and overall accuracy of flow based on extended NETFLOW flow records is almost same to using full packets data. Therefore, this method can be applied to the online network traffic identification, and get better identification results. The paper is structured as follows.

Section 2 provides the related work about ML method and how it can be applied in IP traffic identification, and introduces feature selection algorithms. Section 3 proposes the traffic identification model based on extended NETFLOW flow metrics. Section 4 introduces the multi metric correlation and proposes new MSAS feature selection algorithm. Sections 5 and 6 describes evaluation method and analyzes experiment results. Finally, conclusions are drawn in section 7.

2. The Related Works

The goal of ML method is to identify sample data and build a learning classifier, and then classify the testing samples through the constructed classifier. ML method is introduced to the field of network traffic identification. It can solve the solutions that DPI methods cannot identify the encrypted traffic. Many traffic identification and classification methods based on ML have been proposed to identify and classify the network traffic. We will introduce the methods as

follows.

2.1. Bayesian Methods

First, the Bayesian methods include NaivenBayes, BayesNet, etc., Moore and Papagiannaki [19] introduce the NaiveBayes algorithm to classify and identify network traffic, this algorithm has fast classification speed, but classification accuracy is lower. Later, Moore and Zuev [20] makes use of the Fast Correlation-Based Filter algorithm (FCBF) (feature selection algorithm) to filter the features, and adopt kernel estimation techniques (NaiveBayes with kernel density estimation) to improve classification. Experiment results show that the improved algorithm has been greatly improved in the overall accuracy of classification. Paper [9] extract the relevant feature characteristics, and use genetic algorithms to select features, adopts the Bayesian network to identify P2P traffic. Experiments show that the K2, Tree Augmented Naive Bayes (TAN) and Bayesian network Augmented Naive Bayes (BAN) have higher classification accuracy and faster classification speed. However, this method is a probability-based learning method, and too much dependence on the distribution of the sample space, which will produce potential instability.

2.2. Support Vector Machine (SVM)

Paper [22] propose a Support Vector Machine (SVM) algorithm, and compares with Naive Bayes (NB), Naive Bayes based on Kernel estimation (NBK), NB+FCBF, NBK+FCBF algorithm, the experiment results show that overall accuracy of SVM without feature selection algorithm is not only better than the NB, slightly better than NBK+FCBF algorithm using two kinds of optimization strategies. It can effectively avoid the impact caused by unstable factors, has an obvious advantage in dealing with the traffic classification problem. Literature [14] use the SVM method and feature selection strategy to divide flow into seven types, and experiment of biased samples and unbiased sample were carried out. The experiment results show that classification accuracy of unbiased sample is lower compared to the biased samples.

2.3. Decision Tree, Including C4.5, and Random Trees

Paper [16] evaluate 15 kinds of algorithms from the training time, test time and the overall accuracy of network traffic classification. Experimental results show that C4.5 is the best network traffic identification method. Alshammari and Zincir-Heywood [1] analyzes RIPPER and C4.5 classification algorithm. Results show that C4.5 has higher detection rate and lower false alarm rate. In addition, C4.5 is not affected by distribution of packet size. But it will not be able to achieve to be applied to the online traffic classification [7].

2.4. Feature Selection Algorithm

Feature selection algorithm is roughly divided into two types: one is the filter model; the other is the wrapper model. Filter model mainly evaluates feature through evaluation function. Wrapper model considers classification error rate as the evaluation method. Evaluation function and the classifier of filter model are mutually independent. Filter model mainly includes ranking algorithm and subset search algorithm. Paper [6] compare the advantages and disadvantages of Symmetrical Uncertainty (SU), Relief (based on the Gini index) method and Minimum Description Length (MDL), and points out that the MDL method is the best when samples number is sufficient, and SU is most stable. Paper [17] propose a new calculation method of feature correlation, the maximum information compression index, however, its research is only limited to linear correlation, in order to further avoid the lack of linear correlation. Literature [3, 24, 25, 26, 27, 28] propose a fast feature selection method based on entropy and mutual information, and obtains better classification results. Paper [23] is also based on the entropy theory, points out that the decision variables, calculation method (including SU) of feature correlation can not accurately reflect the correlation, and proposes a calculation method of correlation when decision variables existing and applied to feature selection for network anomaly detection. Paper [4] evaluate Information Gain (IG), the Gain Rate (GR), SU, correlation characteristics (CFS), Support Vector Machine (SVM_RFE) to select the most important voice features, and points out that SVM_RFE is the less effective, results of the other four methods are similar. Paper [5] use algorithm for feature subset selection in order to eliminate the redundant and irrelevant ones. The best results have been achieved using optimal feature subset and MLP with an average rate of 94%. Paper [29] proposes the SRSF feature selection algorithm, and experimental results show that the method can achieve more than 94% flow accuracy and 80% bytes accuracy.

3. Traffic Identification Model based on Flow

Nowadays most researches focus on the data collected for the full payload packets in fields of traffic identification, so you can get more packets information, more accurate results of traffic identification and classification, while this method requires larger computational cost and higher calculating complexity. Online identification traffic is difficult. So we consider the inherent and extended NETFLOW flow statistical characteristics as flow

feature to identify traffic, which can reduce the pressure from the heavy traffic, but also can improve identification accuracy, the real online traffic identification will be realized. In view of this, consider NETFLOW and extended NETFLOW flow records as the study object and proposes traffic identification model as shown in Figure 1.

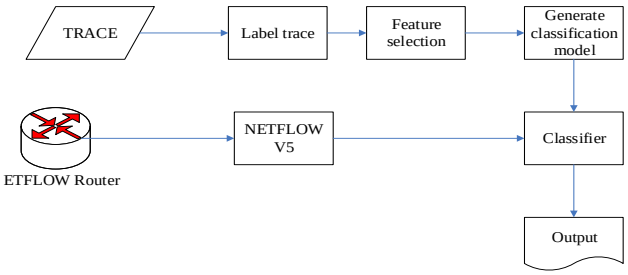

Figure 1. Traffic identification process-based on ML.

Before presenting the model, we will introduce definition of NETFLOW, extend flow records and application type:

- *Definition 1*. NETFLOW flow records and extended flow record; $X=\{x_1, x_2, ..., x_t\}$.
- *Definition 2*. *Application Type*: objective result of identification; $Y=F(x)=\{y_1, y_2,..., y_t\}$.

We can determine the parameters of function by training sample data, the classifier is function $F(X)$ itself.

Figure 1 depicts online traffic identification model based on extended NETFLOW flow. The model is divided into four phases which include data collection, feature selection, construction of traffic identification model, traffic identification.

Table 1. Metric feature.

Feature	Feature Description
Lport	Low Port Number
Hport	High Port Number
During	Flow During
Transproto	Transport Protocol Used(TCP/ UDP)
TCPflags1	TCP Header Flag, or (OR), Transport Layer Protocol is UDP, the Feature is 0
TCPflags2	TCP header flag, or (OR), Transport Layer Protocol is UDP, the Feature is 0
Pps	Packets/Duration
Bps	Bytes/Duration
Mean Packets Arrived Time	Duration/Packets
Biodirection Packets Ratio	Forward Packets/ Backward Packets
Biodirection Bytes Ratio	Forward Bytes/ Backward Bytes
Biodirection Packet Length Ratio	Biodirection Packets Length Ratio
Biodirection Packets	Forward Packets + Backward Packets
Biodirection Bytes	Forward Bytes + Backward Bytes
Tos	Biodirection TOS OR from NETFLOW
Mean Packet Length	Biodirection Bytes/Biodirection Packets

The data collection is process of collecting from full payload packets and uses DPI tools (l7filter) to label traffic. Process of feature selection is mainly to construct NETFLOW and extended NETFLOW flow metrics (detailed shown in experimental section). Traffic identification model is mainly to establish the appropriate classifier based on ML. Traffic identification mainly is to identify traffic by establishing classifier based on NETFLOW V5 format and extended NETFLOW records, and gets the corresponding identification results. Table 1 lists the feature metric used in this paper.

4. Feature Selection Algorithm

This paper introduces the SU based on filter model and FCBF search algorithm.

Entropy is the uncertainty of the random variable and it is metric of containing information. Let X be a random variable, its entropy is calculated by using Equation 1:

$$H(X)=-\sum_i P(x_i)\log_2(P(x_i)) \quad (1)$$

Where $P(x_i)$ value is a priori probability of random variable X, that is, $P(x_i) = P(X=x_i)$. $H(X)$ is the greater, entropy of variable X is the greater, that is, the uncertainty of X is the greater, and amount of information carried is the greater.

Determined the observed value of another random variable Y, the conditional entropy $H(X|Y)$ of the variable X is calculated by Equation 2:

$$H(X|Y)=-\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \quad (2)$$

Where $P(x_i|y_j)$ indicates probability of random variable X taking value x_i when the observed values of the random variable Y is y_j ,thus $P(x_i|y_j)$ is called as the posterior probability of the random variable X, $H(X)$ is regarded as the uncertainty of variable X, before Y value is known, while the $H(X|Y)$ shows that the uncertainty of variable X after random variable Y is known, then $H(X)$-$H(X|Y)$ is amount of X information provided by the random variable Y, it is called as the mutual information between X and Y in information theory, $I(X;Y)$ is expressed by Equation 3:

$$I(X;Y)=H(X)-H(X|Y)=H(X)+H(Y)-H(X,Y) \quad (3)$$

Where

$$H(X,Y)=-\sum_i\sum_j P(x_i,y_j)\log_2(P(x_i,y_j))$$

which is the joint entropy of two variables. $I(X;Y)$ shows that average information of variable X after the variable Y is obtained. It is also expressed as the degree of statistical constraints between two random variables. And the mutual information between the two variables has the symmetry. We can see from Equation 3: $I(X; Y)=I(Y; X)$.If the variable i and Y is uncorrelated, then $I(i; i)=0$; otherwise $I(X; Y)>0$, and $I(X; Y)$ is the greater, indicating that the stronger is the correlation between X and Y. If $I(X; Y)>I(Z; Y)$, then correlation of Y and X is stronger than the Y and Z. Therefore, we can use mutual information $I(X; Y)$ to quantitatively evaluate the correlation between two metrics. However, $I(X; Y)$ values are vulnerable to affect from variable values and

units, so it needs further homogenization. (SU) [4] is expressed by Equation 4:

$$SU(X;Y)=SU(Y;X)=2\times\left[\frac{I(X;Y)}{H(X)+H(Y)}\right] \quad (4)$$

Range of SU and correlation coefficient ρ of pearson is the same as in the [0, 1], and which is monotonic increasing function of the mutual information $I(X; Y)$ [17], the greater is the value, the stronger is the correlation between two variables, and vice versa is weak. When correlation is 0, it represents that two variables are independent, if correlation is 1, it shows two variables have a strictly functional relationship. SU has a high accuracy and versatility; therefore it is commonly widely used as the analysis and evaluation of feature correlation. Wrapper model usually adopt search method to combine feature characteristic selected. Generally feature selection algorithm is a combination of filter model and wrapper model. FCBF feature selection algorithm proposed in this paper is a combination of filter model (SU algorithm) and wrapper model (FCBF search algorithm). The FCBF feature selection algorithm uses SU as evaluation metric. SU is greater; it indicates the feature characteristics have higher correlation. According to features correlation, select the best features subset. However, In currently FCBF method's can only evaluate the correlation between two metrics, and be short of the correlation evaluation method for multi-metric, so this paper proposed the MSAS feature selection algorithm based on the multi-metric correlation. Analysis of multi-metric correlation as pearson correlation coefficient cannot handle the complex correlation between feature metrics and the current is lack of effective evaluation method for the multi-metric correlation, this paper expands the definition of mutual information between the two variable in section 4 Equation 3 and gets the following proposition 1 shows that the mutual information between the random vector of arbitrary dimension, it means the information metric of the statistical correlation between multiple metrics:

- *Proposition 1*

$$I(\vec{X};\vec{Y})=H(\vec{X})+H(\vec{Y})-H(\vec{X},\vec{Y}) \quad (5)$$

Where $\vec{X}$, $\vec{Y}$ respectively is m, n-dimensional random vector.

Where

$$H(\vec{X})=-\sum_i P(\vec{x}_i)\log_2(P(\vec{x}_i))$$

$$H(\vec{X},\vec{Y})=-\sum_i\sum_j P(\vec{x}_i,\vec{y}_j)\log_2(P(\vec{x}_i,\vec{y}_j))$$

- *Proof 1*. the value of the joint values of all variables in the random vector map into another single random variable is derived equation

- *Theorem 1*. $I(\vec{X};\vec{Y})$ has the following recurrence relation:

$$I(\vec{X};\vec{Y})=\sum_{i=1}^{n} I(\vec{X};Y_i\mid Y_{i-1},Y_{i-2},...Y_1) \quad (6)$$

- *Proof 2*.

$$\begin{aligned}
&I(\bar{\mathbf{x}};\bar{\mathbf{Y}})=I\left[\bar{\mathbf{x}};(Y_1,Y_2,\cdots,Y_n)\right]=I\left[(Y_1,Y_2,\cdots,Y_n);\mathbf{x}\right]\\
&=H(Y_1,Y_2,\cdots,Y_n)-H(Y_1,Y_2,\cdots,Y_n\mid\mathbf{x})\\
&=\sum_{i=1}^{n}H(Y_i\mid Y_{i-1},\cdots,Y_1)-\sum_{i=1}^{n}H(Y_i\mid Y_{i-1},\cdots,Y_1,\mathbf{x})\\
&=\sum_{i=1}^{n}\left[H(Y_i\mid Y_{i-1},\cdots,Y_1)-H(Y_i\mid Y_{i-1},\cdots,Y_1,\mathbf{x})\right]\\
&=\sum_{i=1}^{n}I(\bar{\mathbf{x}};Y_i\mid Y_{i-1},Y_{i-2},\cdots Y_1)
\end{aligned}$$

In proof 1, recursive implies the average information amount of the random variables Y_1, ..., Y_n offers to $\bar{\mathbf{x}}$, it is equal to the average amount of information that Y_1 offers to $\bar{\mathbf{x}}$ + the average amount of information that Y_i offers to $\bar{\mathbf{x}}$,when $Y_1,\ldots,Y_{i-1}$ (i=2,…,n) known. So actual calculation of flow metric correlation we can calculate out the mutual information of multi-metric through less mutual information, to simplify the calculation process.

- *Theorem 2*. $I(\vec{X};\vec{Y})\geq I(\vec{X};Y_i)$, i=1, 2, …, n

the equation set up the conditions, if and only if all $(\vec{x},\vec{y})$ satisfied $P(\vec{x},\vec{y})>0$, where $P(\vec{x}\mid\vec{y})=P(\vec{x}\mid y_i)$.

- *Proof 2*. by Theorem 1,

$$\begin{aligned}
&I(\vec{X};\vec{Y})=\sum_{i=1}^{n}I(\vec{X};Y_i\mid Y_{i-1},Y_{i-2},...,Y_1)\\
&=I(\vec{X};Y_i)+I(\vec{X};Y_1\mid Y_i)+...+I(\vec{X};Y_{i-1}\mid Y_1,...,Y_{i-2})\\
&+I(\vec{X};Y_{i+1}\mid Y_1,...,Y_i)+...+I(\vec{X};Y_n\mid Y_1,...,Y_{n-1})\\
&\because I\geq 0\\
&\therefore I(\vec{X};\vec{Y})\geq I(\vec{X};Y_i)
\end{aligned}$$

And when the variables $Y_1,...,Y_{i-1},Y_{i+1},...,Y_n$ are independent, where conditions of mutual information values are only 0. $I(\vec{X};\vec{Y})=I(\vec{X};Y_i)$

- *Theorem 3*. the random vector $\vec{Y}$ provide the amount of information based on $\vec{x}$, which will not less than either the amount of information Y_i provided. That is, correlation between the random vector is stronger than any component.

Similarly, in order to overcome the impact of variable unit on metric correlation, we consider the homogenization treatment of $I(\vec{X};\vec{Y})$ according to Equation 2, get the definition of symmetric uncertainty based on the expansion of any dimension statistical correlation:

- *Definition 3*.

$$SU(\vec{X};\vec{Y})=2\times\left[\frac{I(\vec{X};\vec{Y})}{H(\vec{X})+H(\vec{Y})}\right] \quad (7)$$

$\vec{X}$, $\vec{Y}$ is m, n-dimensional random vector。

Defined by Proposition 1 shows that the symmetry uncertainty between the multi-metric is also in [0, 1], and the results value are higher, the correlation between the random vector $\vec{X}$, $\vec{Y}$ is stronger. In

particular, when m=1, n=2, *SU(X;YZ)* can be expressed by Equation 8:

$$SU(X;YZ)=2\times\left[\frac{I(X;YZ)}{H(X)+H(YZ)}\right]=2\times\left[1-\frac{H(XYZ)}{H(X)+H(YZ)}\right] \quad (8)$$

It represents a statistical multiple correlation relationship between the three measures.
And:

$$SU(X;YZ)\geq SU(X;Y)$$
$$SU(X;YZ)\geq SU(X;Z)$$

4.1. FCBF Algorithm Description

- *Definition 4*. Flow features set {A_1, A_2, ..., A_i, ..., A_n}; flow class set {B_1, B_2, ...,B_k, ..., B_m}.
  First, calculate all the SU(A_i, B_k) between the features and class features, then according to the size of the threshold delta, features that SU(A_i, B_k)< δ will be deleted. The remaining features are arranged in the order of SU(A_i, B_k) from large to small, and respectively the SU(A_i, A_j) values between the features are calculated, if feature A_i is located at front of the feature Bi and SU(A_i, A_j)>SU(A_i, B_k), then feature B_k is redundancy and removed from the queue. The final feature obtained is the set of features reduction.

4.2. MSAS Attribute Selection Algorithm

Introduce the following concept to evaluate the characteristics of the flow features, especially Defined flow classification entropy is expressed as follows:

- *Definition 5*. Information Entropy:

$$Entropy(F)=-p\log p \quad (9)$$

- *Definition 6*. Information Gain:

$$InfoGain(F,A)=Entropy(F)-\sum_{t}\frac{|F_t|}{|F|}Entropy(F_t) \quad (10)$$

Where value A is set of all characteristic feature A, F_t is a set of value of feature A equal to t in flow F.

- *Definition 7*. The expected cross-entropy flow in ECE:

$$ECE(t)=P(t)\sum_i P(c_i\mid t)\log\frac{P(c_i\mid t)}{P(c_i)} \quad (11)$$

Where $P(c_i \mid t)$ indicates probability of the application type is c_i when flow $F \in t$ feature metric. $P(t)$ represents the probability of the feature t, $P(c_i)$ shows the probability of type c_i. $P(c_i \mid t)$ is greater ,then it will prove that the correlation between t and type c_i is stronger. ECE reflects the distance between the classification probability distribution of the c_i and the probability distribution of feature metric t known. Expectation entropy of characteristics features is the greater, and then it will prove that impact of distribution on flow classification is the greater.

- *Definition 8. Flow Distinguish between Features MSAS:*

$$IGECE=InfoGain(F,A)*ECE(t) \quad (12)$$

The Basic Idea of the MSAS Algorithm

Input: With all the features of full flow records.

Output: Features queue selected.

Set C(n,k) is the number of SU combinations between the all multi-metric, k is the feature number which is selected from n the feature metrics.

First we delete features of SU (A_1, …, A_k)<threshold given, the SU of remaining feature is arranged in ascending order, to calculate out the maximum value of SU (A_1,…A_k), and calculate out the minimum value of IGECE(F,A_t). If At is the same feature to A_k, then we will put it into the deleted features queue. Otherwise add it into the selected features queue.

MSAS algorithm we proposed for traffic identification is illustrated in Algorithm 1. InfoGain, ECE and IGECE are listed in Table 2.

- *Theorem 2*. MSAS attribute algorithm time complexity is no more than O $(k*t)^2$
- *Proof* . Let k represents the number of feature sets, t is the type number, d shows number of features queue deleted, s is number of features queue selected, then d+s=k, according to the complexity of SU(A_i,A_j) is O(i*j),where 0<i<k,0<j<k. Complexity of MSAS algorithm is O(i)=O(k*t)*O(i*j)<O$(k*t)^2$

In summary, the time complexity of the algorithm does not exceed O$(k*t)^2$. Is proved.

Algorithm 1: MSAS algorithm.

// Preprocessing stage
Initialize flow feature;
Generate new flow F;
while $F \in A$ *feature do*
compute Entropy(F);
for t do 1 to i
compute InfoGain(F,A_t);
compute ECE(t);
compute InfoGain(F, A_t);
if SU($A_{c(n,0)}$, …,$A_{c(n,k)}$) $> \delta$ *then*
add feature to select feature queue;
compute max(SU(A_i,A_j));
if A_i=A_j then
deleted feature queue←A_i;
Else
selcted feature queue←A_i
end while;
return A;

Table 2. InfoGain, ECE and IGECE of each feature.

ID	Metric	InfoGain	ECE	IGECE
1	Bidirectional pkt	3.08825606703015	0.161845690953003	0.499821
2	Bidirectional Bytes	4.03752409896933	0.447603842491633	1.807211
3	Mean Packet Length	4.15668159567677	0.483473823208809	2.009647
4	Flow During	4.07234788125259	0.458086845521357	1.865489
5	TOS	2.93468392213458	0.115615868840978	0.339296
6	TCPFLAGS1	3.23022474599294	0.204582521765590	0.660848
7	TCPFLAGS2	3.27027924911432	0.216640128666544	0.708474
8	Transproto	2.91725337461152	0.110368751195693	0.321974
9	Low Port Number	3.89946697982442	0.406044508514030	1.583357
10	High Port Number	3.98905799871266	0.433014092541501	1.727318
11	PPS	4.12144779038396	0.472867390954274	1.948898
12	BPS	4.14071941185435	0.478668727081951	1.982033
13	Mean Packets Arrived Time	4.11716015374125	0.471576683714321	1.941557
14	Biodirection Packets Ratio	2.73129901065091	0.0543909098189299	0.148558
15	Biodirection Bytes Ratio	4.17705770529182	0.489607643397876	2.045119
16	Biodirection Packet Length Ratio	4.16410942575108	0.485709822863870	2.022549

5. Evaluations

In this paper, we use the routine evaluation standard for verifying the effectiveness of our classification algorithm. The effectiveness of the current flow identification algorithm has the following three evaluation criteria. And the concepts are as follows:

- *True Positive* (*TP*): The flows of application A are classified as A correctly, which is a correct result for the classification.
- *False Positive* (*FP*): The flows not in A are misclassified as A. For example, a non-P2P flow is misclassified as a P2P flow. *FP* will produce false warnings for the classification system.
- *False Negative* (*FN*): The flows in A are misclassified as some other category. For example, a true P2P flow is not identified as P2P. *FN* will result in classification accuracy loss.

The calculating methods are expressed by Equations13, 14, and 15):

- *Precision*: The percentage of samples classified as A that are really in class A.

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

- *Recall*: The percentage of samples in class A that are correctly classified as A.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

- *Overall Accuracy*: The percentage of samples that are correctly classified.

$$Overall\ Accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \quad (15)$$

6. Experiment Results and Analysis

The main purpose of this section is to evaluate the performance of algorithm. For this, we adopt two dataset and compare MSAS with well-known algorithm (FCBF).

6.1. Dataset

NOC-SET: As shown from Table 3 NOC-SET Data is collected at southeast university, and L7_filter_modify software is used to label the flow.

Table 3. NOC_SET dataset.

AppID	Application	Protocol	Flow Number
1	WWW	HTTP, https, etc.,	904572
2	Bulk	FTP	5483
3	Mail	Pop3, Imap, Smtp	385
4	P2P	BitTorrent, eDonkey, Xunlei, etc.,	11186
5	Service	DNS, NTP	3035
6	Interactive	SSH, CVS, pcAnywhere, etc.,	6
7	Multimedia	RTSP, Real, etc.,	20
8	Voice	SIP, Skype, etc.,	276
9	Others	Games, attacks, etc.,	26500

L7_filter_modify is developed based on L7filter [13]. Finally, NOC_SET dataset is built. A basic requirement of traffic classification is that the flow types are correctly identified.

Table 3 also shows the frequently used application classes of the data sets used in this study. An application class may contain different kinds of data, for example, the class mail includes IMAP, SMTP and POP3.TCP/IP traffic flows are the fundamental objects for classification, which is represented as a flow of one or more packets between two hosts of a network using network communication protocols. The flow is clarified by the IP five-tuple consisting of the source-IP, destination-IP, source-port, destination-port and the protocol type. In order to focus on the traffic classification process itself, the semantically complete TCP connections are selected to make up the training sets and testing sets, where semantically complete TCP flow is defined as: a bi-directional flow for which one can observe the complete connection set-up (SYN-ACK) and another complete connection tear-down (FIN-ACK).

- *MOORE-SET*: The dataset described originally by Moore *et al.* [21] was used for the experiment. This data was randomly sampled in several different periods from one node on the internet. This site was shared by about 1,000 researchers, technicians and management staff of three research institutions, and connected to the Internet through a full-duplex gigabit Ethernet link. All full-duplex traffic on this connection was captured in a full 24hours period, so the original traffic-set contained all full duplex traffic connected the node in both link directions. The original traffic-set contained all full duplex traffic connected the node in both link directions. Since the original traffic-set is too large, Moore divides it into ten subsets by a random sampling method. The sampling times of each subset are almost the same (approximately 1680seconds each) and the non-overlapping random samples are uniformly distributed over the 24hour interval. The number of flows and the proportion of the various types of network traffic are shown in Table 4.

Table 4. MOORE_SET dataset.

AppID	Application	Protocol	Flow Number	Proportion(%)
1	WWW	HTTP, https	328091	86.91
2	BULK	FTP	11539	3.056
3	MAIL	Pop3, Imap, Smtp	28567	7.567
4	DB	Sqlnet, Oracle	2648	0.701
5	SERV	DNS, NTP, Ldap	2099	0.556
6	P2P	Kazaa, Bittorrent, Gnutella	2094	0.555
7	ATTACK	Worm, virus Attacks	1793	0.475
8	MULT	MediaPlayer, Real	1152	0.305
9	INT	Ssh, klogin, Telnet	110	0.029
10	GAME	Halflife	8	0.002

6.2. Experiment

The paper adopts C4.5 and NBK algorithms which are integrated in the WEKA suite, and modified L7-filter tools to label traffic. Experimental platform includes: two PC computers with the Intel Core 2 Duo CPU 2.80GHZ, System of labelling traffic are used Linux systems, System of classification and identification is Windows XP system.

6.3. Experimental Results and Analysis

In this paper, two kinds of experiment are done based on NOC_SET and MOORE_SET data, Symmetrical UncertAttributeSetEval (SU) evaluation methods and FCBF search method, and this paper proposes MSAS feature selection method. In addition, the classification algorithm used C4.5 and the NBK classification algorithm and we adopt 10-fold cross-validation method to evaluate algorithm.10-fold cross validation method is commonly used in precision, its basic idea is that the data set is divided into 10 parts, 9 parts are considered as the training data while only 1 part as the test data. Each experiment will achieve the accuracy, and finally the average of 10 times accuracy of the algorithm is considered as the accuracy. Thus we use feature selection algorithm to reduce features, and adopt cross validation to evaluate classifier, the experiment results as shown in Figures 2, 3, 4, and 5.

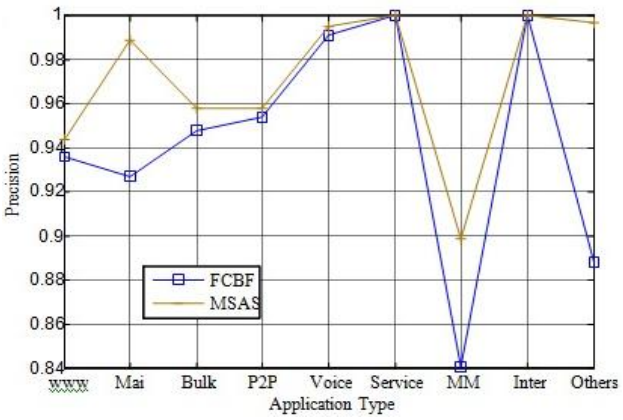

Figure 2. Precision of traffic identification based on C4.5.(NOC_SET).

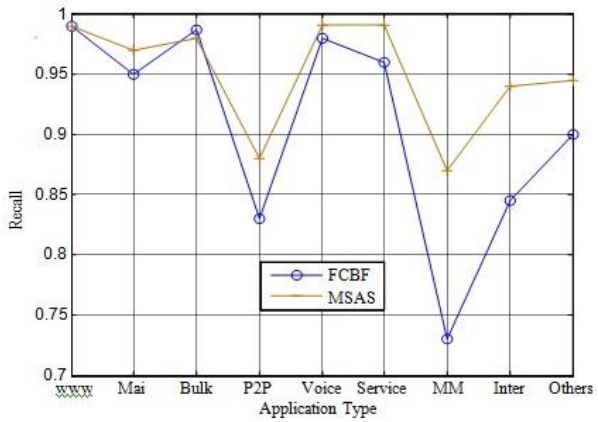

Figure 3. Recall of traffic identification based on C4.5.(NOC_SET).

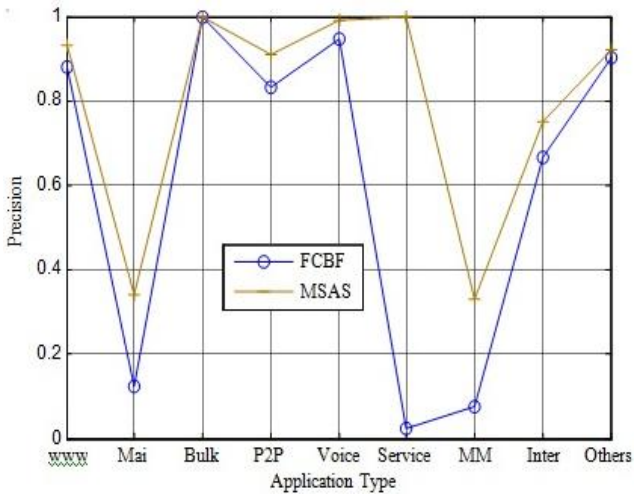

Figure 4. Precision of traffic identification based on NBK. (NOC_SET).

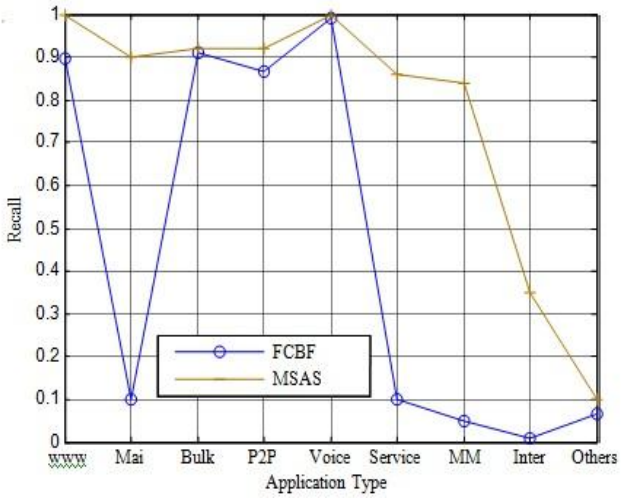

Figure 5. Recall of traffic identification based on NBK. (NOC_SET)

It can be seen from Figures 2, 3, 4, 5, 6, 7, 8, and 9 we use the MSAS feature selection algorithm to sort MSAS value and compare smallest the MSAS value of the feature column and maximum of SU(A_i, A_j) feature column, if the value is the same, it will be deleted. Only the larger feature column of the MSAS will be preserved, and then the MSAS algorithms optimize and combine the features column. Theoretically, these features have larger impact on the overall accuracy of flow F. Facts show that we respectively adopt the C4.5 and the NBK classification algorithm for traffic identification, the identification results are evaluated by the precision and recall. Experiment results show that feature selection algorithm MSAS can improve precision, recall of identification. Www, voice and service have reached more than 97% in precision, while multimedia relatively is lower in precision. From the perspective analysis of samples, we find out that number of www, voice, service sufficient is higher. While multimedia relatively is small. From the view of ECE, $P(c_i|t)$ represents that t to c_i value of the Multimedia is the minimum, resulting in the classification results imbalance.

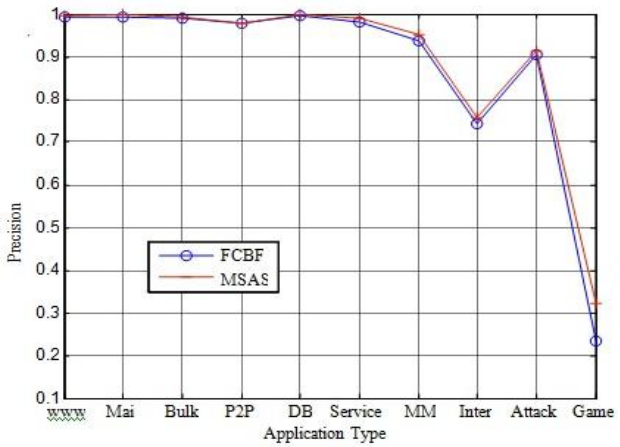

Figure 6. Precision of traffic identification based on C4.5.(MOORE_SET).

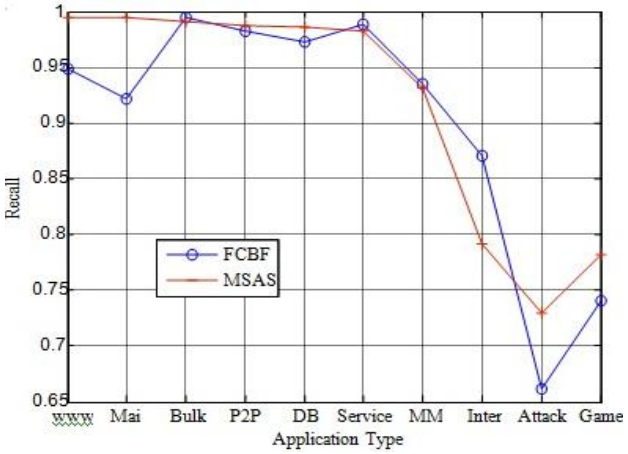

Figure 7. Recall of traffic identification based on C4.5.(MOORE_SET)

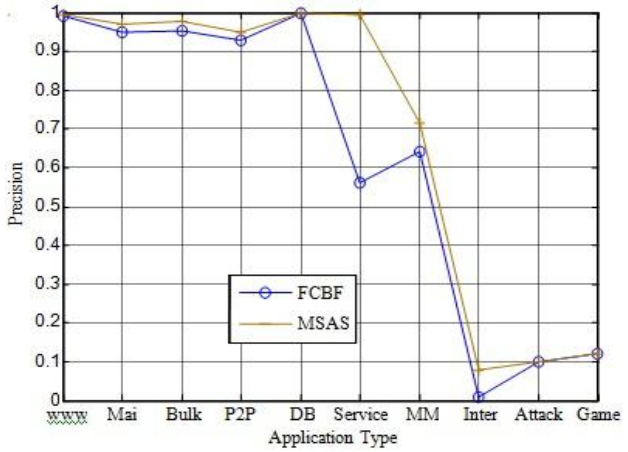

Figure 8. Precision of traffic identification based on C4.5.(MOORE_SET).

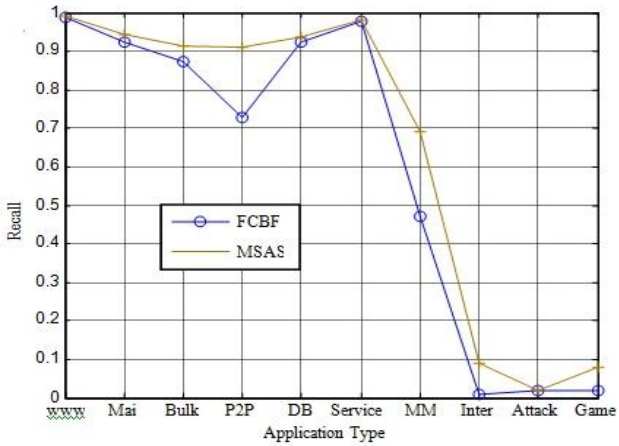

Figure 9. Recall of traffic identification based on C4.5.(MOORE_SET).

As is shown in Tables 5 and 6 we use two different feature selection algorithms to optimize and filter the feature, respectively, adopt C4.5 and the NBK classification algorithm to classify traffic, the overall identification rate results show that the overall accuracy of the MSAS feature selection algorithm has the highest accuracy rate while the C4.5 algorithm with FCBF to carry out feature selection will be the worst results. It is mainly due to using this algorithm to select three features(feature number 9, 12, 14) and using the MSAS to select four features(feature number 3, 12, 15, 16) from 16 features in the NOC_SET, select the best eight features (feature number 4, 72, 108, 91, 155, 202, 113, 50) and the best six features(feature number 20, 31, 50, 108, 200, 212) from 248 features in MOORE_SET, detailed information is shown in Tables 5 and 6. The C4.5 algorithm itself has selection strategy which may have conflicts and contradictions with FCBF selection method, and thus make the classification results even worse. The MSAS overcome the problem that FCBF attribute selection algorithm is vulnerable to the impact of the classification algorithm, and the classifier itself will become more independence.

Table 5. Best features for MSAS (NOCSET).

NOCSET		
ID	**Feature Abbreviation**	**NIGECE**
15	Biodirecation Bytes Ratio	1.0000
16	Biodirecation Packets Length Ratio	0.9881
3	Mean Packets Length	0.9813
12	Bps	0.9667

Table 6. Best features for MSAS(MOORESET).

MOORESET		
ID	**Feature Abbreviation**	**NIGECE**
31	Total Packets a b	1.0000
20	Mean Data ip	0.9634
108	Initial Window-Bytes b-a	0.9012
50	Sack Pkts Sent b-a	0.8934
212	Duration	0.8875
200	Max-IAT_ a b	0.8786

All the analysis shows that this paper proposes the MSAS feature selection algorithm based on flow records and adopt C4.5 and NBK to classify traffic,

classification results show that: MSAS feature selection algorithm whether in precisions or in the recall is higher than FCBF algorithm, and the overall accuracy also has been verified. Traffic identification based on flow records only has few features, but the identification results almost are same to full packets data, so this will provide a good way for online traffic classification. By adding a small amount of the above-mentioned metric features to NETFLOW and build new metric features, it can achieve better classification results, but also can improve online classification and identification.

7. Conclusions

Traffic identification is one of the core issues of the network traffic planning and management, this paper acquires network data from the network boundary of Jiangsu Province and local network, and adopts L7-filter to label the data, construct a baseline data set NOC_SET. The MSAS feature selection algorithm to reduce the dimension of multidimensional features and common FCBF algorithm were compared, the results show that the feature selection algorithm proposed in this paper will get better classification results, and the higher identification rate.

The innovations of this paper are: construction of NOC_SET standard data set from Jiangsu Province-based network boundary data; some features based on the flow are proposed in this paper, and MSAS feature selection algorithms are introduced.

Based on these studies, we will do further research on the flow data and flow measure features and provide data support for future research, but also improve the feature selection algorithm to select the better features and propose some better measure features.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grant No.U1504602), China Postdoctoral Science Foundation (Grant No.2015M572141), Science and Technology Plan Projects of Henan Province (Grant No.162102310147), Education Department of Henan Province Science and Technology Key Project Funding (14A520065), the Development Plan of Science and technology of Henan of China(NO.142300410402), Scientific Research Fund of Henan Provincial Education Department of China(NO.14B520057). School-based Project of Zhoukou Normal University (ZKNUB115101).

References

[1] Alshammari R. and Zincir-Heywood A., "Investigating Two Different Approaches for Encrypted Traffic Classification," *in Proceeding of Sixth Annual Conference on Privacy, Security and Trust*, New Brunswick, pp. 156-166, 2008.

[2] Bell D. and Wang H., "A Formalism for Relevance and Its Application in Feature Subset Selection," *Machine learning*, vol. 41, no. 2, pp. 175-195, 2000.

[3] Flannery B., Press W., Teukolsky S., and Vetterling W., *Numerical recipes in c*, Cambridge University Press, 1992.

[4] Ganchev T., Zervas P., Fakotakis N., and Kokkinakis G., "Benchmarking Feature Selection Techniques on the Speaker Verification Task," *in Proceeding of 5th International Symposium on Communication Systems, Networks and Digital Signal Processing*, Patras, pp. 314-318, 2006.

[5] Gazzah S. and Amara N., "Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script," *The International Arab Journal of Information Technology*, vol. 5, no. 1, pp. 92-101, 2008.

[6] Hall M., *Correlation-Based Feature Selection for Machine Learning*, PhD Thesis, the University of Waikato, 1999.

[7] Hirvonen M. and Laulajainen J., "Two-Phased Network Traffic Classification Method for Quality of Service Management," *in Proceeding of IEEE 13th International Symposium on Consumer Electronics*, Kyoto, pp. 962-966, 2009.

[8] Iliofotou M., Kim H., Faloutsos M., Mitzenmacher M., Pappu P., and Varghese G., "Graph-Based p2p Traffic Classification at the Internet Backbone," *in Proceeding of INFOCOM Workshops*, Rio de janeiro, pp. 1-6, 2009.

[9] Jun L. and Shunyi Z., "Peer-To-Peer Traffic Identification Using Bayesian Networks," *Journal of Applied Sciences*, vol. 27, no. 2, pp. 124-130, 2005.

[10] Karagiannis T., Broido A., Claffy K., and Faloutsos M., "Transport Layer Identification of P2P Traffic," *in Proceeding of the 4th ACM SIGCOMM Conference on Internet Measurement*, Taormina, pp. 121-134, 2004.

[11] Karagiannis T., Papagiannaki K., and Faloutsos M., "Blinc: Multilevel Traffic Classification in the Dark," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 4, pp. 229-240, 2005.

[12] Kiziloren T. and Germen E., "Network Traffic Classification with Self Organizing Maps," *in Proceeding of 22nd International Symposium on Computer and Information Sciences*, Ankara, pp. 1-5, 2007.

[13] Levandoski J., Sommer E., and Strait M., Application Layer Packet Classifier for Linux, 2008.

[14] Li Z., Yuan R., and Guan X., "Accurate Classification of the Internet Traffic Based on the SVM Method," *in Proceeding of IEEE*

International Conference on Communications, Glasgow, pp. 1373-1378, 2007.
[15] Lim Y., Kim H., Jeong J., Kim C., Kwon T., and Choi Y., "Internet Traffic Classification Demystified: on the Sources of the Discriminative Power," *in Proceeding of the 6th International Conference*, Philadelphia, 2010.
[16] Ma Y., Qian Z., Shou G., and Hu Y., "Study on Preliminary Performance of Algorithms for Network Traffic Identification," *in Proceeding of International Conference on Computer Science and Software Engineering*, Wuhan, pp. 629-633, 2008.
[17] Mitra P., Murthy C., and Pal S., "Unsupervised Feature Selection Using Feature Similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301-312, 2002.
[18] Moore D., Keys K., Koga R., Lagache E., and Claffy K., "The Coralreef Software Suite As a Tool for System and Network Administrators," *in Proceeding of the 15th USENIX Conference on System Administration*, San Diego, pp. 133-144, 2001.
[19] Moore A. and Papagiannaki K., "Toward the Accurate Identification of Network Applications," *in Proceeding of the 6th International Conference on Passive and Active Network Measurement*, Boston, pp. 41-54, 2005.
[20] Moore A. and Zuev D., "Internet Traffic Classification Using Bayesian Analysis Techniques," *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1, pp. 50-60, 2005.
[21] Moore A., Zuev D., and Crogan M., "Discriminators for use in Flow-Based Classification," *University of London*, pp. 1-14, 2005.
[22] Peng X., Qiong L., and Sen L., "Internet Traffic Classification Using Support Vector Machine," *Journal of Computer Research and Development*, vol. 46, no. 3, pp. 407-414, 2009.
[23] Qu G., Hariri S., and Yousif M., "A New Dependency and Correlation Analysis for Features," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1199-1207, 2005.
[24] Teufl P., Payer U., Amling M., Godec M., Ruff S., Scheikl G., and Walzl G., "Infect Network Traffic Classification," *in Proceeding of 7th International Conference on Networking*, Cancun, pp. 439-444, 2008.
[25] Valenti S., Rossi D., Meo M., Mellia M., and Bermolen P., "Accurate, Fine-Grained Classification of P2P-TV Applications by Simply Counting Packets," *in Proceeding of Traffic Monitoring and Analysis First International Workshop*, Aachen, pp. 84-92, 2009.
[26] Xu K., Zhang Z., and Bhattacharyya S., "Profiling Internet Backbone Traffic: Behavior Models and Applications," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 4, pp. 169-180, 2005.
[27] Yu L. and Liu H., "Efficient Feature Selection via Analysis of Relevance and Redundancy," *The Journal of Machine Learning Research*, vol. 5, no. 5, pp. 1205-1224, 2004.
[28] Yuan J., Li Z., and Yuan R., "Information Entropy Based Clustering Method for Unsupervised Internet Traffic Classification," *in Proceeding of IEEE International Conference on Communications*, Beijing, pp. 1588-1592, 2008.
[29] Zhang H., Lu G., Qassrawi M., Zhang Y., and Yu X., "Feature Selection for Optimizing Traffic Classification," *Computer Communications*, vol. 35, no. 12, pp. 1457-1471, 2012.

Yongfeng Cui received the BS degree in Computer Science and Technology from Henan Normal University and the MS degree in Computer Application Technology from Huazhong University of Science and Technology, China in 2000 and 2007 respectively. He is currently researching on Computer Application Technology (CAT).

Shi Dong received the M.E. degree in computer application technology from University of Electronic Science and Technology of China in 2009 and the PhD in computer application technology from Southeast University in 2013. Currently, he is an associate professor in the School of Computer Science and Technology at Zhoukou Normal University and he also works as post doctor researcher in Huazhong University of Science and Technology. He is member of China Computer Federation and a visiting scholar in Washington University in St. Louis. His research interests include distributed computing, network management.

Wei Liu received the BS degree in computer science and technology from Henan Normal University and the MS degree in Computer Technology from The PLA Information Engineering University, China in 1996 and 2006 respectively. He is currently researching on Network Technology (NT) and Computer Application Technology (CAT).