Journal of Computational Information Systems 8: 24 (2012) 10429–10437
Available at http://www.Jofcis.com

Traffic Classification Model Based on Integration of Multiple Classifiers ⋆

Shi DONG [1,2,3,*], Dingding ZHOU [3], Wei DING [1,2]

[1] *Computer Science and Engineering, Southeast University, Nanjing 211189, China*

[2] *Eastern China (North) Regional Network Center, Nanjing 211189, China*

[3] *Zhoukou Normal University, Zhoukou 466001, China*

Abstract

This paper introduced the concept of multi_classifier fusion into traffic classification to improve the classification accuracy. And first define conception of preference degree and analyzes the impact on traffic protocol type, and proposed MCM (Multi_Classification_Model) and compared with traditional machine learning method, the results show MCM method will be able to increase identification accuracy and less be influenced by high sampling rate.

Keywords: Multi_classifier; Preference Degree; MCM; Sampling Rate

1 Introduction

With the increasing of network bandwidth, network behavior become more complex, produced a variety of new network applications, traffic identification and classification are becoming increasingly important for network administrators and designers are considered as a research focus by domestic and foreign researchers. In order to obtain more accurate identification accuracy and higher identification efficiency, before traffic classification we need to use feature selection method, which enable to choose quantitative and effective feature with great influence on classification accuracy. Currently there exists a lot of single classifier traffic identification methods, such as: BAYES neural network, SVM, C4.5 decision tree [1, 2, 3, 4, 5, 6, 7] and other methods, but because the algorithm for different samples exist different fitness. Thus in order to solve this, since the single classifier has one-sidedness problem, this paper presents a multi-classifier fusion based on the flow classification model. On the one hand can be overcome by the existence of a single classifier fitness defect, it can also improve the accuracy of classification and identification. At present, many researchers adopt a single classifier method, but performance improvement of a single classifier has been faced to a flat neck, and the study on flow identification of multi-classifier is relatively few. When different single classifiers deal with different data,

⋆Project supported by the National Nature Science Foundation of China (No. 2009CB320505).

*Corresponding author.

Email address: `njbsok@gmail.com` (Shi DONG).

December 15, 2012

so when faced with different network flow data, the noise data will make accuracy of a single weak classifier decline. But also the current problem that online traffic identification and classification impacted by the sampling is needed to solve, It has been applied to network intrusion detection and anomaly detection process [8]. However, the multiple classifiers in the research field of network traffic classification have just begun.this paper aims to propose a multi-classifier fusion model based on traffic classification, which can solve the identification results for impact by noise and sampling, and based on flow model we analyze the CERNET network data. The results show that: the model can effectively reduce the sampling jitter caused by the phenomenon of identification results, and can effectively deal with noise as the data caused by the instability of the phenomenon of identification result issues. And compared with a single classifier, we proposed multiple classifiers which have a higher identification accuracy and lower classification error rate.Integration and fusion of multiple classifiers is a method to improve the performance of the classifier, multi-classifier fusion is the combination of multiple classifiers, each classifier is also known as the base classifier. Base classifiers only perform classification tasks, while the combination of multi-classification phase is responsible for generating a number of different base classifiers, and the process of multiple classifier fusion is each base classifier which classify the sample data, and generate classification results, the integration of multiple classifiers using voting mechanism will assess the results of the classification, the final assessment of the results obtained according to the classification of the final results. Integration of multiple classifiers includes the most important points: 1: the generation of base classifier 2: Combination and evaluation of base classifier for classification results.multi-classifier fusion can be expressed as: C=$c_1, c_2, ..., c_n$ which states that the base classifier. Each base classifier $c_n(x) \rightarrow y$, y indicates the class label. The fusion of multiple classifiers is that the cue relationships $c_{1,2,...,n} \rightarrow y$. Only the integration of multiple classifiers is integrated mapping results obtained from each base classifier.

A necessary condition for accuracy of the integration of multiple classifiers is higher than any base classifier is: multiple classifiers meet the various members of the base classifier which have accuracy and diversity. The so-called right is the wrong division which is less than 0.5, while diversity refers to the error distribution of each classifier which is different, for example, there are three classifiers, one for this category were judge wrong, but the other two classifiers have made the right treatment, so that the final result is correct, and if the contrary, two have made the wrong judgments and the only one to make the right treatment, the outcome would be a mistake. If each base classifier error identification rate is larger than 0.5, when adopting the vote mechanism to deal with the final result, only more than half of the classifier to determine the error, the results will lead to error, so probability of classification error of the classifier fusion is $c_n^m p^m (1-p)^{n-m}$ Which, $m > n/2$ and the probability value is less than p. Therefore, the classifier fusion is greater than the accuracy of each base classifier. This document is organized as follows: Section 1 describes the related research, and relevant content for further analysis and discussion. Section 2 presents a multi-classifier fusion model based on flow-level. Section 3 proposed definition of preference and time degree.Section 4 introduced the algorithm of multi-classifier fusion.Section 5 proposed the metric feature based on extended netflow.Section 6, 7, 8 respectively, analyzed and evaluated the results of experiments and draw relevant conclusions and prospects.

2 Multi_Classification_Model of Traffic Classification

Logical structure of the flow classification integration model (see Figure 1 below):

This flow classification model is divided into three layers: the input layer, classification level, decision-making layer.

1) Input layer: The main statistical characteristics of the flow were considered as the input layer of the input data. This mainly were composed of the metric feature proposed in Section 5.

2) Classification layer: This layer is the core layer of the classification fusion model, mainly constituted by the base classifier.

3) Decision-making layer: which mainly was based on assess mechanisms to complete final choice, such as voting mechanism.

Model concepts and processes:

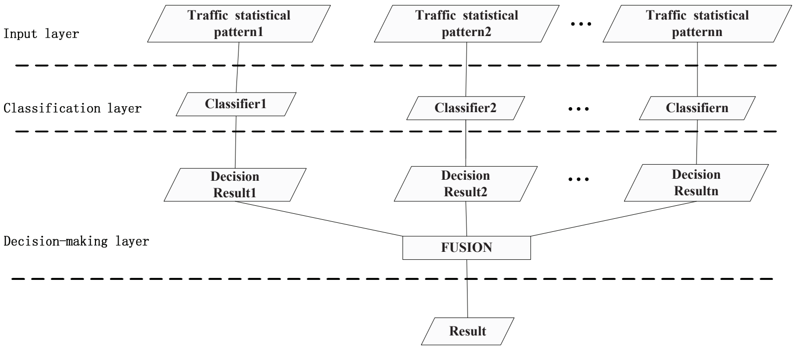

Fig. 1: The structure of multi-classifier model

Definition 1 ***Flows:*** *A typical flow definition uses the 5-tuple (source$_{IP}$, source$_{port}$, destination$_{IP}$, destination$_{port}$, transport-level protocol)*

Definition 2 ***Traffic statistics features****: It is composed of the flow records with characteristic, the so-called flow characteristics is calculated when TRACE data is composed into flows. Specific form of expression is:*

$$T_{nm} = \begin{pmatrix} T_{11} & T_{12} & \cdots & T_{1m} \\ \vdots & & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nm} \end{pmatrix} n > 0, 0 < m < 17 \quad (1)$$

Definition 3 ***Multi-classifier type:*** *Multiple classifiers are divided into two categories: Similar classifiers: the classifier with the same composition of mixed classification Diversity classifiers: the classifier is composed of different types of hybrid classifier*

Definition 4 ***Decisions types:*** *Decisions are composed with the mean, Bayesian decision making and voting.*

This paper adopts international standard classification which is divided into nine categories. Considering that the base classifier exist differences in classification efficiency and in preference for the samples of classification. This will need to consider in the decision-making process by adding the concept of weight, and this concept is based on the weight of the preferences of different classifiers. For example, classification of type i has good classification results, so we give higher weight. For example, the following section will detail introduce research on the classifier preference of traffic classification.

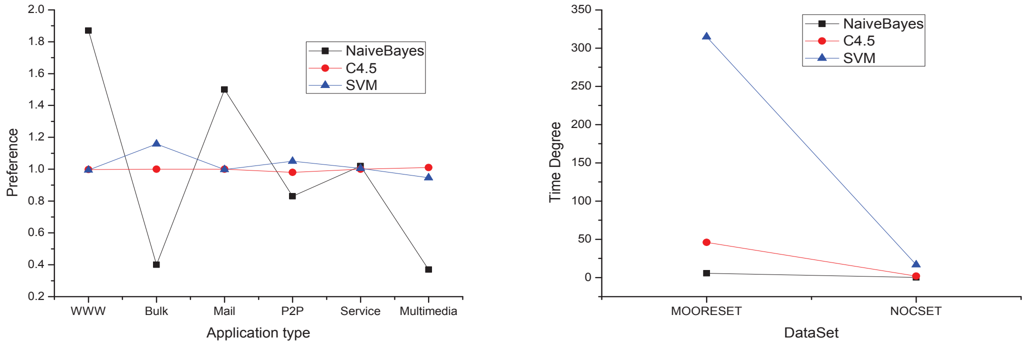

Fig. 2: preference and time degree

3 Preference and Time Degree of Traffic Classification

Definition 5 *In order to quantitatively analyze preference of traffic classifier for each category preference H are introduced.*

$$H = \frac{precision}{recall} \tag{2}$$

Where, definition of Precision and Recall is showed in section 6. In pattern recognition and information retrieval, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. Through a large number of experiments show that the relationship between the precision and recall is opposite dependence, if precision is increased, the recall will be reduced and vice versa.

Definition 6 *In order to take into account the flow can be used to identify in high-speed environment, so the degree of time is introduced. On the same sample data, different classifiers have different time efficiency, when we design the classifier fusion, this factor need to be considered. As can be seen from Figure 2, different classifiers have different timeliness.*

$$T = Training\_time + Classification\_time \tag{3}$$

Time degree T consists of two parts, which is summation of training time and classification time. No matter which time all will have influence of T

4 Classification Algorithm of the Multiclassifier Integration

Multi_classifier introduce voting mechanism in decision-making layer, the classification level of the base classifier are expansible, set BC_1, BC_2, ..., BC_n as the base classifier. Adopting the parallel strategy, we firstly proposed the preference of flow training sample, and by adopting different singer classifier to train traffic flow. The preference of classifiers were obtained. Such as BC_1 has a high preference for the WWW, BC_2 has high preference for P2P, BC_3 has good preference for

Bulk. Then consider to adopt BC_1, BC_2, BC_3 respectively to train WWW, P2P and Bulk, to set weights and then use majority voting method in decision-making. Majority voting method is a thought which decision-making obtained from different classifier and consider max decision-making number for application type as classification result. In formula 4, C_m represents the classification categories. P_{ij} shows the probability of the flow identified as C_m.

$$C_m = MAX_{j=1}^{m} \sum_{i=1}^{n} p_{ij} \tag{4}$$

The algorithm description is as follows:

Input: flow with a specific flow feature

Output: flow with label

The base classifier identify the application type according to their preference, each base classifier generate c sub-classifier. Firstly compute the preference of different classifiers. Sort the preference order by dsc. and select $(\lfloor n/2 \rfloor + 1) * c$ the classifier as classifier set to adopt application of maximum votes as the application type of flow.

Algorithm 1: Multi_Classication_Model

```
// Initialize in the network (often randomly)
for each $i \in application\_list, i = 1, 2, ..., c$ do
    for each $n \in classifiers, n = 1, 2, ...$ do
        // gernerated i*n based classification $BC_{ni}$ ,n for classifier type
        $BC_{ni}$=Get_BC(training_samples,$app_i$);
        // After generated based classification,and start training process of
           multi-classification .using multi_vote method to mark label
        if $BC_{ni}$!=null then
            $Mi'$ = Feature_Metric_Selection(training_samples, M, $app_i$);
            $Mi''$ = Redundant_Metric_Delection(training_samples, $Mi'$);
            $H_{ni}$=Get_Preference(training_samples,$|Mi''|$,$app_i$);
            $T_{ni}$=Get_Timedegree(training_samples,$|Mi''|$,$app_i$);
            $O_{ni}$=Get_O(training_samples,$BC_{ni}$,$H_{ni}$,$T_{ni}$);
            $BC_{Oi}$ = Multi_ Classication_Model (training_samples, $|Mi''|$,$O_{ni}$);
            // Add based classification model $BC_{Oi}$ into MCM algorithm model lib and label
               as <app, $app_i$>,where $o = \lfloor n/2 \rfloor + 1$
            $Put\_BC_{Oi}$(app, $app_i$, $BC_{Oi}$);
            // The classification of the fuzzy sets to form a new data set and use the
               sample proportion and choice of fuzzy set preference to mark the end of the
               label as <app, $app_i$>
            VOTE_BC(app, $app_i$,$BC_{Oi}$);
        else
            Goto exit
```

5 Metric Feature

Most of the current study used data collected for the whole packet of data, so you can get more information, more accurate identification and classification. However, these classifications method can only collect the current offline data, and then online identification, such identification efficiency is relatively low. For the emergence of NETFLOW flow, we consider the use of NETFLOW inherent properties to achieve the traffic flow identification, which can greatly reduce the load pressure from traffic, but also can improve the identification efficiency, truly online traffic identification. In view of this, this paper made use of NETFLOW flow records and extend flow records as research object, we introduced some following concepts:

Definition 7 *NETFLOW flow records and extend flow record are described: x = (x1, x2, x3, ...xt);*

Definition 8 *The target set of application types are described: Y=F(x)=(y1, y2, y3, ...); function parameters can be determined by training Sample data, and the classifier is a function F(x) itself.*

Flow metrics We adopted are 16 metrics (low port number; high port number; Flow duration; Transport protocol used (TCP/UDP); TCPflags1; TCPflags2; pps; bps; Mean packets arrived time; Biodirection Packets ratio; Biodirection Bytes ratio; Biodirection packets; Biodirection bytes; tos; Mean packet length).

6 Evaluation

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

TP (true positive): is the number of the samples that actually have type i among all those correctly classified as type i by the classification model.The flows of application i are classified as i correctly, which is a correct result for the classification;

FP (false positive): is the number of the samples that actually have type i among all those classified as another types by the classification model.The flows not in i are misclassified as i. For example, a non-P2P flow is misclassified as a P2P flow. FP will produce false warnings for the classification system;

FN (false negative): is the number of the samples that do not have type i among all those misclassified as type i by the classification model.

The calculating methods are as follows:

Precision:

The percentage of samples classified as i that are really in class i

$$Precision = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

Overall accuracy:

The percentage of samples that are correctly classified. It is the proportion of the all instances that are correctly classified as truly types in whole samples

$$Overallaccuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \tag{6}$$

The experiments with K-fold method ensures that when the limited available data, can effectively obtain precision estimation.

Table 1: *NOC_SET* dataset

AppID	Application	Protocal	Flow number	Proportion
1	WWW	HTTP	4943	64.6
2	Bulk	FTP	39	0.5
3	Mail	IMAP, POP3, SMTP	91	1.19
4	P2P	BitTorrent, eDonkey, Gnutella, XunLei	1414	18.5
5	Service	DNS, NTP	433	5.7
6	Interactive	SSH, CVS, pcAnywhere	6	0.08
7	Multimedia	RTSP, Real	20	0.3
8	Voice	SIP, Skype	276	3.6
9	Others	games, attacks	431	5.6

7 Experiment

Experimental data: (1) NOC_SET: End systems capture data (.Pcap file) and generated 16 feature metrics.

7.1 Local data capture

This paper studies several applications: WWW, Bulk, Mail, P2P, Service, Interactive, Multimedia, and Voice. Eight kinds of applications are captured based on packet. Use L7-filter [9] software to deal with these data and label type, finally which was composed into flow, and generated NOC_SET data set. Shown in Table 1.

7.2 Experiment results and analysis

In this paper, the data were based on NOC_SET and classification algorithm is adopted for four experiments, 1: A multiclassifier model algorithm (MCM) 2: The three single classical machine learning classifier (C4.5, Naivebayes, and SVM). Assessment and validation adopt ten folds cross-validation of data cross-validation. Ten-fold cross-validation method is commonly used by precision test methods, and its basic idea is that the data set is divided into 10 parts, one of nine is considered as the training data, only one as the test data, Each experiment will produce the appropriate accuracy, and consider the average of correct rate of 10 times as the estimated accuracy of the algorithm. We adopt cross-validation method to deal with data in this article, and after an assessment of the classification algorithm, the assessment results as shown in the figure below. To analyze influence of the packet sampling on multiple classifier fusion models, respectively, through five different sampling rates to observe the overall accuracy of the final impact of the results as shown figure 4: It can be seen from Figure 3 and table 2; multi-classifier model has higher identification than the traditional single-classifier algorithm, from Table 2, the overall accuracy of MCM classificaiton and identification found is highest in four methods. Figure 4 show MCM method is less impacted by different sampling rate and overall classification accuracy rate is relatively stable, while the traditional single classifier algorithm's overall accuracy appeared to increase with the sampling rate increasing. Thus we can apply the MCM mehod to network management system with sampling strategy (such as Netflow).

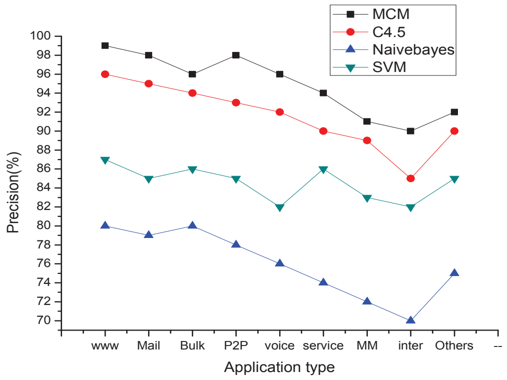

Fig. 3: Precision of multi-classifier and ML method

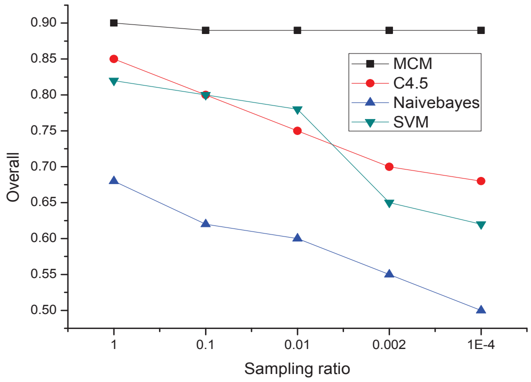

Fig. 4: Impact of sampling on multi-classifier and ML method

Table 2: Overall accuracy of multi-classifier and ML method

Classification method	Overall accuracy(%)
MCM	97.36
C4.5	96.65
Naivebayes	58.5
SVM	92.5

In summary, this paper introduces the definitions of preference into the multi-classifier model based on flow, the identification results showed that: multi-classifier model algorithm both the precision and recall are higher than traditional single-classifier algorithm, while the overall accuracy has also been verified. And from the results, when adopting flow records to identify, although very few relevant features, but the identification result almost can be achieved the same classification results with full-packet data set, so that it can be used for online traffic classification to provide a good way. We can add a small amount of the above features into NETFLOW existing features and it will be able to achieve very good classification results, but also improve the efficiency of online classification.

8 Conclusion

This paper builds ground truth NOC_SET. By adopting combination method of multiple classifiers to solve the current network traffic classification problems, and compared with the usual machine learning algorithms method, the results show that the proposed multi-classifier model can be better identified, higher is identification accuracy, and which is no influence of sampling. Innovation of this paper is: (1) Data based on the border of Jiangsu Province build NOC_SET network standard data set; (2) proposed several metrics based on flow features. (3) Proposed a multiple classifiers model for the flow classification. Based on this research, the next step is mainly: based on this flow data and the measure of flow properties proposed, to provide data to support the further study, is also able to improve the identification of multi-classifier model to better meet the online traffic identification.

Acknowledgment

This paper is supported by National 973 Plan Projects (2009CB320505) and National Science and Technology Plan Projects (2008BAH37B04).

References

[1] T. T. T. Nguyen and G. Armitage. "A survey of techniques for internet traffic classification using machine learning", *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 4, pp. 56 – 76, 2008.

[2] H. Trussell, A. Nilsson, P. Patel, and Y. Wang. "Estimation and detection of network traffic", in *Digital Signal Processing Workshop, 2004 and the 3rd IEEE Signal Processing Education Workshop. 2004 IEEE 11th.* IEEE, 2004, pp. 246 – 248.

[3] A. McGregor, M. Hall, P. Lorier, and J. Brunskill. "Flow clustering using machine learning techniques", *Passive and Active Network Measurement*, pp. 205 – 214, 2004.

[4] S. Zander, T. Nguyen, and G. Armitage. "Self-learning ip traffic classification based on statistical flow characteristics", *Passive and Active Network Measurement*, pp. 325 – 328, 2005.

[5] "Automated traffic classification and application identification using machine learning", in *Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on.* IEEE, 2005, pp. 250 – 257.

[6] K. Lan and J. Heidemann. "A measurement study of correlations of internet flow characteristics", *Computer Networks*, vol. 50, no. 1, pp. 46 – 62, 2006.

[7] C. GU, S. ZHANG, X. CHEN, and A. DU. "Realtime traffic classification based on semi-supervised learning", *Journal of Computational Information Systems*, vol. 7, no. 7, pp. 2347 – 2355, 2011.

[8] I. Corona, G. Giacinto, C. Mazzariello, F. Roli, and C. Sansone. "Information fusion for computer security: State of the art and open issues", *Information Fusion*, vol. 10, no. 4, pp. 274 – 284, 2009.

[9] J. Levandoski, E. Sommer, M. Strait *et al.* "Application layer packet classifier for linux", 2008.