Towards Selecting Optimal Features for Flow Statistical Based Network Traffic Classification

Ming XU, Wenbo ZHU, Jian XU, Ning ZHENG
College of Computer
Hangzhou Dianzi University, Hangzhou, China
{mxu, 121050041, jian.xu, nzheng}@hdu.edu.cn

***Abstract*—The network traffic classification is one of the most fundamental work in the network measurement and management, and this problem is more and more impact as the network scale grows. Many methods are proposed by researchers, but methods based on flow statistics seem more popular than the others. In this paper, we proposed a novel method based on refined flow statistical features. The new statistics, skewness and kurtosis, and new flow statistical features, payload length, were introduced into raw feature set firstly. Then, with the consideration of efficiency in the classification stage, the feature selection was used on the raw feature set to get an optimal feature set and the feature selection are mainly based on the K-means clustering algorithm. The comparison experiment results show that the proposed optimal feature set reaches the same precision level with half time consuming and internal cluster distance when compared with the raw set.**

Keywords—network traffic classification; statistical feature; feature selection.

I. INTRODUCTION

The goal of Network traffic classification is identifying the applications that generate the network traffic. It is of great importance in network management and security, such as quality of service (QoS), intrusion detection and digital investigation. Traditional tools use the port numbers recorded in IANA list for traffic identification. Other classification methods are mainly based on payload content such as DPI method. They abandons the port number, but suffering from security and privacy problems. As the volume of network data is tremendous in daily practice, more researchers are fond of applying the machine learning to traffic classification recently.

The machine learning techniques can be generally categorized into supervised and unsupervised methods. In this paper, we mainly concern about the unsupervised methods, clustering. For a clustering method, the selection of features affects the result directly. Because we can not specify which one feature or collection of features is more efficient than the others. Besides, too many features will worsen the complexity of clustering. Moreover, there always are some features which do not make any effort in differentiating different flows. No work has given out a satisfactory answer to this problem.

In this work, our major contributions are as follows:

- We put forward a novel flow statistical features set based on existing works and our observation.
- We give out two approaches on evaluating the performance of feature set, the p_c and p_m.
- We attempt using correlation based filter and heuristic searching based wrapper to get a refined the feature set.

This work is supported by the Natural Science Foundation of China (Grant No. 61070212 and 61003195), and the State Key Program of Zhejiang Province Natural Science Foundation of China (Grant No. LZ15F020003).

II. RELATED WORK

Traffic classification using flow statistical features is proposed for years. In 2004, McGregor et al. [1] proposed a method to gather the traffic flows into several clusters using the EM algorithm by giving out the probability of each flow type. In this work, statistics features like packet size, interval time and byte counts are introduced in traffic classification.

In 2011, Wang et al. [2] proposed a semi-supervised approach for clustering the traffic. The author assumes that the flow f_1 and f_2 are sharing the same application protocol if they share a same IP address and port. Then they use the K-means algorithm to generate the flow clusters and mark the traffic in same cluster as the same type of flow.

In 2012, Zhang et al. [3] proposed a weighted symmetrical uncertainty metric for feature selection. They filter most of features using the metric and a wrapper method. Eventually, they reach a stable feature set, but the flow is confined to TCP.

In 2014, Fahad et al. [4] proposed an approach in feature selection for traffic classification. They firstly apply maximum entropy algorithm for refining. Then the most respective features are determined by random forest filter. Besides, they discussed little about unknown type of flow.

None of above works discuss thoroughly about the features in the classification. Why a feature should be kept, what feature performs a more significant role and is there a minimized feature set satisfies the precision requirement in classification. That is what we want to research about.

III. FLOW STATISTICAL FEATURES

A. Introduction of Higher order moments s

Firstly we would like to introduce the moment. In statistics, a moment is a specific quantitative measure about the shape of points set. For instance, the mean is a first order statistics, and the moment significance of variance is 2 and the standard deviation is also second order for it is derived from variance.

The common higher order moments refer to the skewness and the kurtosis. The definitions of these to statistics are listed below as (1) and (2). The *E[X]* represents the expectation of *X*, the μ is the mean of *X* and the standard deviation of *X* is σ.

The skewness mainly describes the asymmetry of a distribution about the mean. If the distribution is symmetric, e.g., the Gaussian distribution, the skewness is zero. And the kurtosis is measuring the aggregation degree on its mean value of a certain distribution. The kurtosis of Gaussian distribution is 3 and the range of kurtosis is (0, $+\infty$).

$$Skewness(x) = E\left[\frac{(X-\mu)^3}{\sigma^3}\right] \quad (1)$$

$$Kurtosis(x) = E\left[\frac{(X-\mu)^4}{\sigma^4}\right] \quad (2)$$

The calculations of these two statistics are simplified in realization. For none negative numbers, when the mean and standard deviation are both zero, we would assign zero to these two higher order statistics representing they are invalid.

B. Novel statistical feature set

There is a prerequisite that makes all the statistics suitable for the classification work. That a certain type of flow has its pattern and the pattern could be quantified as a probability distribution. In order to depict the distribution precisely, we should use as many statistics as we can and we can make a speculation that the distribution cannot be easily depicted by only mean or standard deviation. So the higher order moments should be introduced to make a stronger feature set.

A statistical feature set proposed by Zhang et al. in [5] and other works of theirs. This set is independent with the network infrastructure and has a closer relationship with the application behavior. As a combination of this set and higher order moments, a novel statistical feature set is listed in TABLE I .

TABLE I THE NOVEL COMPLETE FEATURE SET (CFS)

Flow Features	Statistics
Packet Size (pkt_size)	Maximum (max), Minimum (min), Mean(mean), Standard Deviation (stde), Skewness (skew), Kurtosis (kurt).
Payload Size (pld_size)	
Time interval (tim_intv)	
Totals of the flow	Total packets count, Total bytes, Total payload.

IV. FEATURE SELECTION USING HEURISTIC SEARCHING METHODS

A. Feature selection methods and the evaluation function

Feature selection aims at picking out a subset from the complete set so that the classification could be simplified with the equivalent performance. Generally, there are two kinds of method for feature selection, the filter and the wrapper. And the performance of the set is judged by evaluation function.

The filter method mainly follows some given rules to decide whether one feature should be kept or removed. Common filter rules including the correlation of features and so on. This kind of rules is independent with the classifiers. It only concentrates on the characteristic of features themselves. The correlation filter is defined in (3), X and Y are two features:

$$r_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\cdot\mathrm{var}(Y)}} = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2\cdot\sum_{i=1}^{N}(Y_i-\bar{Y})^2}} \quad (3)$$

Then we talk about the wrapper and its evaluation function. The wrapper is a classifier and its evaluation function is the precision of classification result. In our problem, we must identify which flow is correctly classified and which is not. We employ the majority vote method for precision calculation as Fig. 1 illustrates. Firstly, for each flow in cluster C_i, there is a ground truth type T_t for every flow w_t, it votes for its type T. There must be at least one type is voted within a cluster. Secondly, we annotate the most voted type T_{most} as the type of cluster C_i. Finally, the precision of this cluster is defined as the proportion of flows who have voted for the T_{most} of all flows included in cluster C_i. This is shown in (4).

$$precision(C_i) = \frac{\mathrm{Count}(T_t = T_{most} \mid w_t \in C_i)}{\mathrm{Count}(w_t \in C_i)} \quad (4)$$

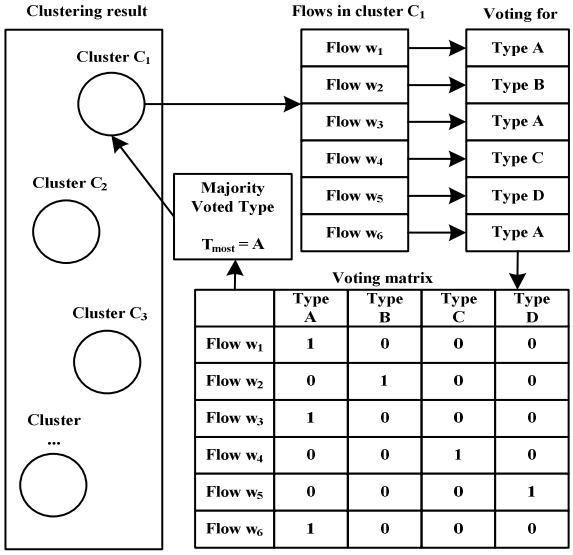

Fig. 1. The precision calculation employing majority vote

The overall precision can be defined in two ways. If there are n clusters after clustering stage, then the overall precision can be defined as the mean of all clusters' precision, noted as p_m. Besides, we can use the ratio of flows that correctly voted, noted as p_c. Correctly voted means one flow votes for the type that finally being selected as the T_{most} of the cluster. They are defined as (5) and (6). They are both proper for $f(S_k)$:

$$p_m = \frac{\sum_{i=1}^{n} precision(C_i)}{n} \quad (5)$$

$$p_c = \frac{\mathrm{Count}(correctly\ voted\ w)}{\mathrm{Count}(all\ flows)} \quad (6)$$

B. Heuristic searching and the SFS strategy

In order to find out the optimal features, we have to adjust the input of classifier carefully and inspect the classification result. We do not care about the details inside the classifier and we treat the classification stage as a black box.

In features selection, with the consideration of efficiency, researchers prefer using heuristic searching methods to the full search method. There are kinds of searching strategies and Sequential Forward Searching (SFS) is employed in our work.

Consider an n features selection situation, the feature set $E_n=\{F_1, F_2, \ldots, F_n\}$. SFS strategy starts with an empty set and adds one feature in each generation. We could define S_k as the local optimal feature set of k-th generation. We should run the classification for once using the S_k so that we could evaluate S_k by its classification result. We define $f(S_k)$ as the evaluation function of set S_k. Starts from S_k, we could get S_{k+1} following the rule that the feature which be added in should be the best performance one among its generation. The candidate set D_k comprised features included in E_n but excluded from S_k. Each feature in D_k will be combined with S_k. The feature in D_k which receives the highest precision with S_k, we name it as F_t, will be kept and then merged into S_k to generate the S_{k+1}. Given D_k, S_k and f, the S_{k+1} should meet the terms as (7):

$$\begin{aligned} f(S_{k+1}) &= f(S_k, F_t) \quad (t \in \{1,2,\ldots,n-k\}) \\ &= \max(f(S_k,F_1), f(S_k,F_2), \ldots, f(S_k,F_{n-k})) \end{aligned} \quad (7)$$

As introduced above, we can get a set $S'= \{S_1, S_2, \dots, S_n\}$ including feature sets of all generations. Given S' and f, the optimal feature set S_{opt} could be indicated as (8):

$$S_{opt} = \max(f(S_1), f(S_2), \dots, f(S_n)) \tag{8}$$

The global searching procedure is time consuming and the searching finishes when $k=n$, i.e., $S_n=E_n$. A good termination condition should interrupt the searching at a proper occasion to avoid useless works and a practical condition is when all the feature set of one generation cannot reach higher performance than the optimal set of previous generation S_k. Then, $S_{opt}=S_k$. Given the S_k, D_k and f, the stop condition describe above is (9):

$$\begin{aligned} f(S_k) &> f(S_{k+1}) \\ &= f(S_k, F_t) \quad (t \in \{1, 2, \dots, n-k\}) \\ &= \max(f(S_k, F_1), f(S_k, F_2), \dots, f(S_k, F_t)) \end{aligned} \tag{9}$$

Assume we have a feature selection problem on a feature set comprised of 5 features. Given the $E_5 = \{F_1, F_2, F_3, F_4, F_5\}$, the f and the stop condition mention in (9), then the whole selection procedure by using SFS is illustrated in TABLE II .

TABLE II A EXAMPLE OF WORKING PROCEDURE OF SFS METHOD

Ge[a]	Feature Set	$f(S_k)$	Note
0	***Empty Set***	**0**	S_0, no feature is selected.
1	$\{F_1\}$	35	
	$\{F_2\}$	40	
	$\{F_3\}$	***45***	S_1, feature F_3 is selected.
	$\{F_4\}$	30	
	$\{F_5\}$	33	
2	$\{F_1, F_3\}$	25	
	$\{F_2, F_3\}$	57	
	$\{F_3, F_4\}$	***60***	S_2, feature F_4 is selected.
	$\{F_3, F_5\}$	40	
3	$\{F_1, F_3, F_4\}$	73	
	$\{F_2, F_3, F_4\}$	***78***	S_3, feature F_2 is selected.
	$\{F_3, F_4, F_5\}$	45	
4	$\{F_1, F_2, F_3, F_4\}$	72	None set receives better performance,
	$\{F_2, F_3, F_4, F_5\}$	55	so the search stops.
Op[b]	$\{F_2, F_3, F_4\}$	***78***	The $f(S_k)$ reaches the max when k=3

[a] Ge is short for generation.

[b] Op is short for the Optimal Set.

C. The selection procedure of proposed method

The proposed procedure of feature selection combines both filter method and wrapper method. We starts from the CFS in TABLE I . As it contains more than 20 features, we consider applying the filter method upon CFS to obtain the MFS (Middle Feature Set). Then we put the MFS into a wrapper, judge the performance of subsets and find out the OFS (Optimal Feature Set). By combining these two stage, we could reduce the complexity of wrapper and avoid the problem of distinguish correlated features in wrapper stage. The specifications of the two stages are as below.

Correlation based filter stage:

(1) Divide the features of CFS into several related groups.

(2) Calculate correlation of features within certain group.

(3) Remove features which receive high correlation.

Heuristic searching wrapper stage (also see TABLE II):

(1) Select the highest scored feature from candidate set.

(2) Add the feature into optimal set and remove it from candidate set at the same time to renew both set.

(3) Repeat step (1) and (2) until the stop condition is reached, the previous optimal set is the final optimal set.

(4) Repeat step (1), (2) and (3) under different conditions to receive stable optimal set.

V. EXPERIMENT AND PERFORMANCE EVALUATION

A. Dataset and the ground truth of flow type

We use the WAND dataset and its corresponding DPI tools, libprotoident [6], to generate the ground truth, which is capable of ground-truth-constructing for its high accuracy [7]. The details about the datasets are listed in TABLE III .

TABLE III THE INFORMATION OF FRAGMENTS IN WAND

Fragment	Starting time(GMT)	Duration	Fragment Size(MB)	Flow count
030946.dsl	2010-01-06 03:09:46	0:20:14	1,818	148K
033000.dsl	2010-01-06 03:30:00	0:30:00	2,675	207K
040000.dsl	2010-01-06 03:04:00	0:30:00	2,763	211K
043000.dsl	2010-01-06 03:04:30	0:30:00	2,691	220K
050000.dsl	2010-01-06 03:05:00	0:30:00	2,655	207K
053000.dsl	2010-01-06 03:05:30	0:30:00	2,749	218K
060000.dsl	2010-01-06 03:06:00	0:30:00	2,670	217K
063000.dsl	2010-01-06 03:06:30	0:30:00	2,797	215K

B. Using filter feature selection on the CFS

As the filter method concerns the internal relations of features, we think the result extracts from one fragment is applicable for others. So we apply filter method only on the data fragment 030946.dsl.

The features in CFS could be grouped into time interval and size by its background meaning. The correlation of the two group is illustrated as TABLE IV and TABLE V .

TABLE IV CORRELATION OF TIME INTERVAL GROUP

F1 / *F2*	(I) max time intv	(II) min time intv	(III) mean time intv	(IV) stde time intv	V skew time intv	VI kurt time intv
(I)						
(II)	0.0845					
(III)	0.6830	0.4805				
(IV)	***0.9393***	0.0327	***0.7728***			
(V)	0.1037	-0.0553	-0.1379	-0.0270		
(VI)	-0.0021	-0.0523	-0.0216	0.0007	-0.0077	

We set a threshold $r = 0.75$ and any correlation coefficient value above r is highlighted and at least one feature will be removed. Then we can receive the MFS shown in TABLE VI .

C. Using wrapper feature selection on the MFS

We then apply the wrapper method on the MFS to get the OFS. We employ the simple K-means that wrapped in the feature selection and the value of K is 50. We only apply the SFS searching on the first two fragment due to time limit. And we use both p_m and p_c as the evaluation function. The searching stops as is mentioned above. The searching results are shown in TABLE VII and it shows that we can receive feature sets with similar component and length. Based on these sets, we give out the global OFS in TABLE VIII with the descending order on times of shown and ascending order on generation. Besides, the *A* means absence.

D. The performance of optimal feature set

We firstly give out the comparison of precision using CFS, MFS and OFS in Fig. 2. We can see that the precision is not severely affected by the reduction of feature counts while some subsets even reach higher precision level. Then we give out the time consumption of classification stage in Fig. 3 and all data came from a fixed machine under same working condition.

At last, we give out the breakdown of each data fragment in TABLE IX for validation of our work.

TABLE V CORRELATION OF DATA LENGTH GROUP (MEASURE: 0.001)

F2 \ F1		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	max_pkt_size														
2	min_pkt_size	-30													
3	mean_pkt_size	***761***	47												
4	stde_pkt_size	***875***	-79	666											
5	skew_pkt_size	-310	-46	-583	-305										
6	kurt_pkt_size	7	5	1	-80	-58									
7	total_pkt_size	49	73	89	-24	-282	378								
8	max_pld_size	***1000***	-30	***762***	***875***	-311	7	48							
9	min_pld_size	-90	***988***	-9	-130	-26	5	68	-88						
10	mean_pld_size	-6	0	-3	-5	-5	0	0	-7	3					
11	stde_pld_size	-5	-1	-2	-4	-5	0	1	-10	3	***976***				
12	skew_pld_size	-286	-43	-536	-285	***950***	-5	-253	-288	-24	3	13			
13	kurt_pld_size	-2	3	-7	-69	17	***803***	281	-3	3	0	7	227		
14	total_pld_size	50	73	92	-22	-291	374	***999***	50	69	0	1	-264	271	

TABLE VI MIDDLE FEATURE SET(MFS) AND THE NOTATIONS

Flow Features	Counts	Statistics
Packet Size (pkt_size)	7	max (A), min (B), mean (C), stde (D), skew (E), kurt (F), Total packet size (G).
Payload Size (pld_size)	1	Mean (mean, H).
Time interval (tim_intv)	5	max (I), min (J), mean (K), skew (L), kurt (M).
Total	1	Total packets count (N).

TABLE VII THE SEARCHING RESULTS AFTER APPLYING SFS

k	030946.dsl				033000.dsl			
	F_k	p_m	F_k	p_c	F_k	p_m	F_k	p_c
1	C	0.51	A	0.56	C	0.51	A	0.56
2	A	0.62	C	0.64	D	0.59	C	0.66
3	E	0.67	D	0.66	A	0.63	E	0.69
4	D	0.71	J	0.69	E	0.66	D	0.71
5	J	0.74	E	0.71	J	0.71	B	0.74
6	B	0.75	L	0.73	B	0.75	F	0.76
7			F	0.73	G	0.76		
8					L	0.77		

TABLE VIII GLOBAL OPTIMAL FEATURE SET (OFS)

	Feature	Times of Shown	Generation of shown			
C	mean_pkt_size	4	1	2	1	2
A	max_pkt_size	4	2	1	3	1
D	stde_pkt_size	4	3	5	4	3
E	skew_pkt_size	4	4	3	2	4
J	min_tim_intv	3	5	4	5	*A*
B	min_pkt_size	3	6	*A*	6	5
L	skew_tim_intv	2	*A*	6	7	*A*
F	kurt_pkt_size	2	*A*	7	*A*	6
G	total_pkt_size	1	*A*	*A*	7	*A*

TABLE IX THE BREAKDOWN OF DATA FRAGMENTS

Type \ Frag	Un-known	HTTP/ HTTPS	SSL/ MAIL/ OTHER	DNS	P2P
030946	33.88	18.85	5.81	1.63	39.83
033000	34.82	18.05	6.99	1.68	38.45
040000	34.26	17.62	7.10	1.66	39.36
043000	34.90	17.66	6.52	1.68	39.25
050000	35.72	16.70	6.65	1.57	39.37
053000	35.24	18.15	6.91	1.42	38.29
060000	32.33	18.14	6.53	1.47	41.53
063000	31.07	17.73	7.04	1.03	43.12

VI. CONCLUSION

We proposed a novel feature set with some new statistical features and apply a serial of refinement on this set. These works meet our expectation as the experiments illustrate. In the future, we may do some pre-processing on the flows, use some other feature selection methods or employ a filter as the evaluation function of feature selection.

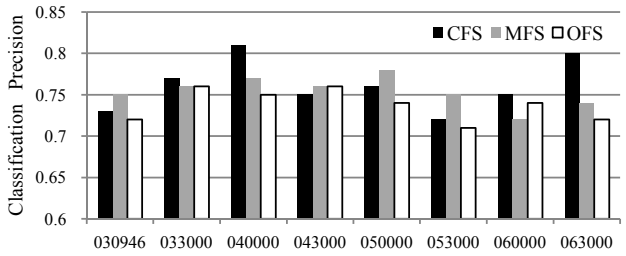

Fig. 2. The classification precision of three feature sets

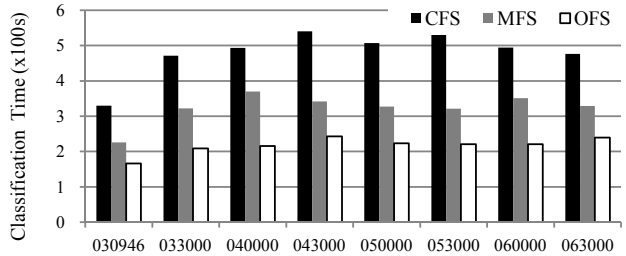

Fig. 3. Fig.3 Time consumption of three feature sets

REFERENCES

[1] Anthony McGregor, Mark Hall, Perry Lorier, et al. Flow Clustering Using Machine Learning Techniques[C]. Springer Berlin Heidelberg, 2004: 205-214.

[2] Wang Yu, Xiang Yang, Zhang Jun, et al. A novel semi-supervised approach for network traffic clustering[C]. 5th International Conference on Network and System Security. IEEE, 2011: 169-175.

[3] Zhang H, Lu G, Qassrawi M T, et al. Feature selection for optimizing traffic classification[J]. Computer Communications, 2012, 35(12):1457–1471.

[4] Adil Fahad, Zahir Tari, Ibrahim Khalil, et al. An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion[J]. Future Generation Computer Systems, 2014, 0(36): 156-169.

[5] Zhang Jun, Chen Chao, Xiang Yang, et al. Classification of Correlated Internet Traffic Flows[C]. IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, 2012: 490-496.

[6] WAND Network Resarch Group[EB/OL]. 2013. http://wand.net.nz/

[7] Valentín Carela-Español, Tomasz Bujlow, Pere Barlet-Ros. Is Our Ground-Truth for Traffic Classification Reliable?[C]. Springer International Publishing, 2014: 98-108.