A stable feature selection approach for optimizing traffic classification based on adaptive threshold

Wenbei Duan, Yuanli Wang

Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, Wuhan University of Technology, China

duanwenbei_dwb@163.com

Keywords: Traffic classification, Feature selection, ATFS, TRF, WSU.

Abstract. In recent years, machine learning algorithm has been widely studied in the field of traffic classification. However, most studies focus on performance improvement of classifier, pro-phase work of traffic classification - feature selection is ignored. Therefore, WSU is regarded as metric, an ATFS algorithm - (Adaptive threshold feature select) is designed on the basis. Namely, algorithm is based on precision autonomous selection threshold of classifier aiming at different datasets. Each dataset will generate a set of attribute subset eventually. Stable features are selected in different screened attribute subsets through TRF algorithm, thereby reaching the purpose of high precision. The experiment shows that the features finally selected in the algorithm can reach the precision of >96% on C4.5 classifier.

Introduction

Traffic identification classification is the basis of IP network management and network security monitoring. High-efficient recognition of traffic can assist many ISP (Internet Service Provider) and network administrators to deeply understand traffic composition, thereby controlling the traffic.

Traditional traffic identification algorithms are shown as follows: identification based on IP port, identification method based on applied payload signature[1-3] and identification method based on machine learning. Machine learning is realized based on traffic statistical features. It does not involve payload of message compared with other methods. However, it is still faced with some problems due to data imbalance and concept drift.

ATFS and TRF are proposed in the paper to ease data imbalance and concept drift problem. Innovation point in the paper lies in that (1) adaptive linear adjustment threshold is adopted on the basis of literature [4]; (2) an algorithm is proposed for screening attribute subset with stable performance.

Relevant work

In the field of flow recognition, machine learning has been widely used[5,6]. Traditional machine learning classification method is realized through optimization of machine learning algorithm. Suitable feature selection is not available in face of data distribution changing with time, a fixed and well-trained classifier can not be used for reaching high precision.

Data imbalance problem refers to imbalanced category distribution in test dataset. Namely, the proportion of all categories is greatly different in the dataset. The prediction result of the classifier is deviated to category with larger proportion in order to improve the precision [2]. The method of solving data imbalance problem can be basically divided into two categories: sampling method[6] and algorithm implementation[8].

Similarly, the concept drift also has great influence in the process of machine learning classification. Because a lot of network data flow is dynamic, as the change of time, those features based on data flow are also changing. Different network control strategies lead to great difference in P2P traffic at daytime and nighttime.

In the paper, two feature selection algorithms are proposed in order to optimize traffic classification, respectively ATFS and TRF. Firstly, ATFS is used to solve the problem of data imbalance. Stable attribute subset is further selected by TRF on the basis of features obtained from ATFS.

Realization method

Metric has very high influence in algorithm[7], and WSU in literature [4] is adopted by ATFS algorithm as metric after careful comparison. Suitable threshold is automatically selected aiming at each dataset algorithm, thereby realizing high precision of classifier. Finally, we use TRF algorithm to select stable features of attribute subsets in each dataset as results of the experiment.

Evaluation standard

Feature metric. Attribute dimension is higher in traffic classification. The correlation between attributes and features as well as the correlation among attributes is measured through WSU values. WSU value is higher, the correlation thereof is greater. Excellent attribute subset should have the following features: there is higher correlation between category attribute and all feature attributes in the subset. The redundancy is lower among all feature attributes.

In the experiment, categories are defined as X weights:

$$p_i = p(X = c_i) = 1\text{-}\ \frac{m_i}{m} \tag{1}$$

In the formula, m_i is the appearance frequency of category x_i in experiment data. The proportion of all categories can reach a relatively balanced state under the weighing of the weight. Therefore, the information entropy of category X is defined as follows:

$$H_w(X) = -\sum_i \sum_j p_i P(x_i, y_j) \log_2 P(x_i) \tag{2}$$

The condition information entropy of category X under condition Y is defined as follows:

$$H_w(X|Y) = -\sum_j P(y_j) \sum_i p_i P(x_i|y_j) \log_2 P(x_i|y_j) \tag{3}$$

Condition mutual information can be obtained on the basis of equations 2 and 3 as follows:

$$IG(X|Y) = H_w(X)\text{-}\ H_w(X|Y) \tag{4}$$

Finally, the weighted uncertainty WSU can be obtained as follows:

$$SU_w = 2\left[\frac{IG_w(X|Y)}{H_w(X) + H_w(Y)}\right] \tag{5}$$

ROC curve. The selected feature value is tested on the basis of classifier after threshold filtering, thereby reaching the best effect. AUC value is higher, the performance of the classifier is better. Compared with direct comparison of AUC values in each category or AUC average value of all categories, the weighed AUC value can alleviate the problem due to data imbalance.

Feature selection algorithm

ATFS algorithm. ATFS is a combined algorithm, and the following standard is adopted for evaluating δ performance.

$$w_i = \frac{n_i}{N} \tag{6}$$

$$M_s = \sum_i w_i{*}pre_i \tag{7}$$

Wherein, n_i represents the frequency of category x_i in the dataset. pre_i refers to precision of category x_i, namely prediction of actual positive proportion in positive results. $M_s$value is higher, it is represented that the classification effect is better after δ threshold is filtered, and the screened feature is used in the classifier.

ATFS algorithm realization flow is shown in figure 1.

Dataset D is a test data with M categories and N attributes. In addition, dataset undergoes algorithm discrete treatment of Kononenko in advance.

Step1:According to the calculated WSU value and most feature values filtered by δ, in the paper, the correlation $SU_w(F_i, C)$ between each attribute and category is calculated firstly, if it is larger than the threshold δ, all feature values consistent with conditions should be sequenced according to descending order on the basis of $SU_w(F_i, C)$ value. Then, $SU_w(F_p, F_q)$ of each pair of features is calculated,if $SU_w(F_p, F_q)$ is higher than $SU_w(F_i, C)$, it is judged that similarity is available between F_q and F_p, thereby deleting F_q. In order to avoid the defect in literature [4] that uniform threshold is

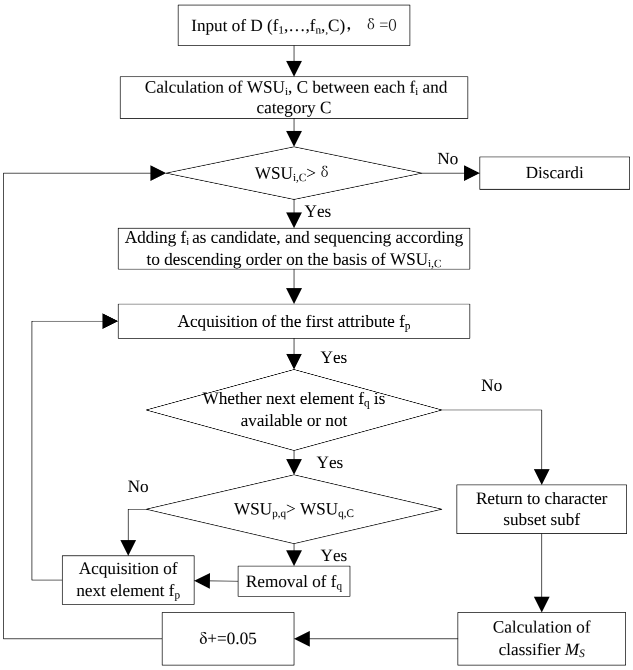
Input of D (f1,…,fn,,C)， δ =0
Calculation of WSUi, C between each fi and category C
WSUi,C> δ
No
Discardi
Yes
Adding fi as candidate, and sequencing according to descending order on the basis of WSUi,C
Acquisition of the first attribute fp
Yes
Whether next element fq is available or not
No
Yes
No
WSUp,q> WSUq,C
Return to character subset subf
Yes
Acquisition of next element fp
Removal of fq
δ+=0.05
Calculation of classifier MS

Fig 1. ATFS algorithm flowchart

used in all datasets, adaptive linear adjustment threshold δ with classification effect as indicator is proposed in the paper. One evaluation standard M_S is set. According to the standard, linear iteration is implemented according to fixed amplification ε=0.05 each time, appropriate value δ is selected.

Step 2: According to attribute subset trainer classification which is obtained in Step 1, seven remaining datasets are regarded as test sets for calculating AUC value under all categories. The weighed AUC value is calculated according to weight value. One feature is deleted from attribute subset to form a new attribute subset. Weighed AUC under new attribute subset are calculated in turn, they are compared with previous weighed AUC. If the value is lowered, the feature is retained, otherwise the feature is deleted. All attributes are traversed in turn. The attribute subset capable of obtaining the maximum AUC value can be screened.

TRF algorithm. Although ATFS algorithm can well solve the problem of data imbalance, when it is faced with different datasets, different attribute subsets still can be selected. Therefore, we need to select stable features as final attribute sets suitable for traffic classification.

All datasets are screened by ATFS; each dataset can produce a group of attribute subsets. The final subset is determined by appearance frequency of each feature and the correlation thereof with the category. Therefore, appearance frequency of all features and $WSU_{i,C}$ average value with categories should be obtained, and the dataset should be screened according to the following measurement standards:

$$\mu = P_i * \overline{WSU_{i,C}} \qquad (12)$$

Wherein, $P_i = \frac{f_i}{N}$, f is the appearance frequency of feature i. N is total appearance frequency of all features.

If μ is greater than the threshold, the feature is retained, and otherwise the feature is removed from the attribute subset.

Evaluation method

In the paper, we use quintuple form to define each flow. Quintuple form includes the follows: source IP, destination IP, source port, destination port and TCP protocol.

Dataset

The dataset adopted in the experiment is the data provided by Cambridge Lab, which is used most widely. The protocol distribution condition in the dataset is shown in Table 1.

Table 1. Protocol distribution condition in dataset

Application	Num of flow	Persent(%)
http	328091	86.91
Imap,PoP2/3,Smtp	28567	7.567
FTP	11539	3.056
Windows Media player, Real	1152	0.305
Oracle, Ingres, Postgres	2648	0.701
Dns,X11,Ntp	2099	0.556
Internet Worm And Virus Attacks	17393	0.475
SSH,Rlogin,Telnet	110	0.029
Kazaa,Bittorrent Cnutella	2094	0.555

Evaluation standards

In the paper, three metric standards are used for evaluating algorithm effects.

True positive rate (TPR) of each category is defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{13}$$

False positive rate(FPR) of each category is defined as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{14}$$

Precision of each category is defined as follows:

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{ALL}} \tag{15}$$

Experimental results

248 attribute features in literature [9] are regarded as foundation. We chooseCambridge dataset. It is finally proved that the generated attribute subset can effectively track TCP flow.

Firstly, the features selected by ATFS algorithm and WSU_AUC algorithm are shown in Table 2. Feature serial number is listed only due to space limitations. Concrete description of each feature can be checked in literature [9].

Table 2. Features selected by ATFS algorithm and WSU_AUC algorithm

Training set number	Thresholdδ	ATFS	Thresholdδ	WSU_AUC
01	0.3	96 95 89 60 2 1	0.55	65 1 95 96 66
02	0.35	163 118 111 96 95 93 83 60 46 45 1	0.55	1 96 45 2 224 4
03	0.6	96 95 60 1	0.55	1 96 46 93 2 111 216 9 206 218 3
04	0.65	96 95 86 1	0.55	1 96 60 89 2 218 36
05	0.66	186 96 95 83 1	0.55	1 96 88 60 170
06	0.6	186 95 83 1	0.55	1 60 46 36
07	0.35	180 156 113 96 95 93 90 84 83 60 46 1	0.55	1 83 96 162 93 232 36 2 121 209
08	0.3	185 180 163 162 114 112 96 95 93 90 84 83 60 45 2 1	0.55	1 113 162 90 93 112 114 2 36 111 209

Table 2 shows that fixed 0.55 is selected for WSU_AUC algorithm in the aspect of selecting threshold. Different threshold is selected for each dataset by ATFS algorithm in order to ensure classifier precision. The screened attribute subsets are also different.

After eight attribute subsets are obtained by ATFS and WSU_AUC, they are used for screening final attribute subset as input of TRF algorithm. TRF algorithm is affected by threshold. The influence of threshold change on average precision is discussed in Table 3.

Table 3. The influence of threshold changes on average precision

Threshold	Average precision(%)
0.048	0.964652284
0.044	0.964658547
0.020	0.964609788

It is obvious that when the threshold is 0.048, corresponding average precision is the highest, which is up to 96.4658%.

Therefore, the attribute subsets screened by ATFS algorithm and TRF algorithm on C4.5 classifier are shown in Table 4.

Table 4. Attribute subset screened by ATFS algorithm and TRF algorithm on C4.5 classifier

the robust features of ATFS (threshold = 0.044)	The robust features of WSU_AUC(threshold = 0.026)
1 96 60 2	1 83 95 96

Features screened by ATFS algorithm are 1, 60, 2 and 96, the features screened by WSU_AUC algorithm are 1, 83, 95 and 96.

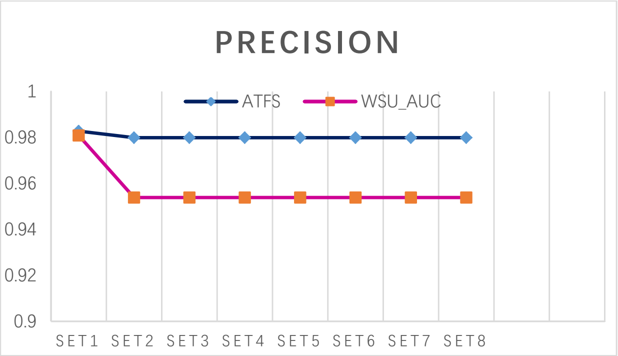

Fig 2. precision of all datasets

The two attribute subsets are respectively tested on eight datasets in order to verify the attribute subset screened by ATFS algorithm and agency algorithm. The precision on each dataset is calculated. Figure 2 shows the precision of attribute subset screened by ATFS algorithm and WSU_AUC algorithm on TRF algorithm on each dataset after classification. The results show that ATFS algorithm is remained above 98%. The precision of WSU_AUC algorithm on set 1 is equivalent to ATFS algorithm. However, the precision is maintained at about 95% on several remaining datasets. Therefore, attribute subset obtained by ATFS algorithm has better classification effect relatively.

The attribute subset screened by ATFS algorithm is respectively tested on C4.5 and NBK classifier in order to find a classifier with better classification effect on the attribute subset.

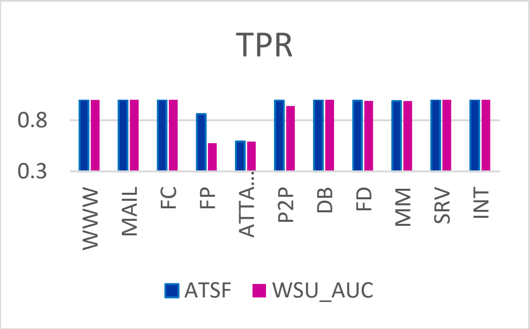

Fig 3. TPR under each category

TPR of two attribute subsets under all categories of C4.5 decision tree classifier is analyzed in Figure 3. The category with higher proportion, two algorithms reach higher TPR. ATFS algorithm TPR is prominently higher than WSU_AUC aiming at other category with low proportion under FP and P2P categories. Figure 4 shows FPR corresponding to two attribute subsets under each category. ATSF algorithm FPR is prominently lower than WSU_AUC on WWW and FD category. However, the WSU_AUC FPR under FP and P2P categories is slightly higher than ATFS algorithm.

Future 4 feature selection algorithm should focus on search mode, evaluation method, etc. In the paper, the evaluation method is optimized only. In addition, we will study how to test dynamic data flow change and how to reconstruct corresponding model to adapt to the changes deeply.

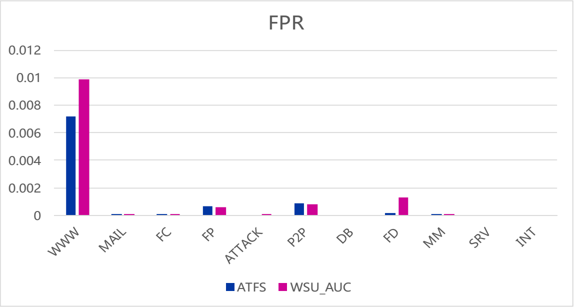

Fig 4. FPR under each category

Conclusion

In the paper, two algorithms of ATFS and TRF are proposed. In the high-dimensional unbalanced dataset, ATFS algorithm can independently set appropriate δ in order to adapt to different data distribution. Suitable algorithm is selected in TRF algorithm for practical application aiming at attribute subset selected on the basis of ATFS. In the paper, validation is implemented on Cambridge dataset. Experiments show that ATFS algorithm will not be limited to a fixed value compared with WSU_AUC algorithm, the precision of classifier is used as standard to choose appropriate δ. The correlation between attribute appearance frequency and category can be comprehensive considered in TRF algorithm. The selected feature can achieve higher precision on C4.5 classifier. However, algorithm running time is scarified due to threshold iteration. In addition, C4.5 decision tree classification effect is more satisfactory than NBK classification effect.

References

[1] S. Sen, Oliver Spatscheck, Dongmei Wang, Accurate, scalable in network identification of P2P traffic using application signatures, in: Proceedings of the 13th International Conference on World Wide Web, New York, USA, May, 2004, pp. 512-521.

[2] Thomas Karagiannis, Application-specific payload bit strings, http://www.cs.ucr.edu/

[3] H. Patrick, S. Subhabrata, O. Spatschek, D. Wang, ACAS: automated construction of application signatures, in: Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data. Philadelphia, Pennsylvania, USA, 2005, pp. 197-202.

[4] Hongli Zhang. Gang Lu. Feature selection for optimizing traffic classification. Computer Communications 35 (2012) 1457–1471.

[5] L. Peng, H. Zhang, B. Yang, Y.H. Chen, M.T. Qassrawi, G. Lu, Traffic identification using flexible neural trees, in: Proceedings of the 18th International Workshop on Quality of Service (IWQoS), Beijing, China, June 2010, pp. 1–5.

[6] T. Auld, A.W. Moore, S.F. Gull, Bayesian neural networks for internet traffic classification, IEEE Trans. Neural Networks 18 (1) (2007) 223–238.

[7] Adil Fahada. Zahir Tari. An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion. Future Generation Computer Systems 36 (2014) 156–169.

[8] D.A. Cieslak, N.V. Chawla, A. Striegel, Combating imbalance in networkintrusion datasets, in: IEEE International conference on Granular Computing,Athens, Georgia, May, 2006, pp.732–737.

[9] A. Moore, D. Zuev, M. Crogan, Discriminators for use in flow-based classification, Technical Report, University of Cambridge, Computer Laboratory.