FEATURE SELECTION FOR APPLICATION RECOGNITION IN COMMUNICATION NETWORKS

[a]MILAN ZELINA, [b]MILOŠ ORAVEC

Fakulta elektrotechniky a informatiky STU, Ilkovičova 3, 81219 Bratislava
email: [a]milan.zelina@stuba.sk, [b]milos.oravec@stuba.sk

Research described in this paper was done within the grant No. 1/0214/10 of the Slovak Grant Agency VEGA.

Abstract: Feature selection is important part of any machine learning process. In our work, we describe methods for feature selection, in order to perform classification of network traffic based on statistical features of the flow. For this purpose, we implemented Sequential forward selection method for feature subset selection. The features used for this method of traffic classification include packet sizes, port numbers, protocol, use of TCP flags and more. We try to identify the most important features and we evaluate generated feature subsets by Naïve Bayes classifier.

Keywords: feature selection, machine learning, traffic classification, traffic representation

1 INTRODUCTION

The research of traffic classification in communication networks is looking for techniques that do not rely on port numbers or packet payload examination. Recent work is emerging on the use of statistical traffic characteristics to assist in classification process. The application of machine learning techniques is very promising in this field [1]. When performing the classification using methods of machine learning, finding a good set of features is essential for the accuracy of classifier. In this paper we present methods for traffic classification and particularly feature selection to identify the most important of available features.

1.1 Classification using port numbers

First application classification practices used to rely on the use of transport-layer port numbers. In the case flow is classified assuming, that the most applications use "well known" TCP or UDP port numbers [1]. These port numbers are visible in TCP or UDP packet headers. In this approach, classifier only compares these obtained port numbers with IANA list [2] that is freely available. In Table 1 we can see port numbers for some popular applications. This approach could have been effective in the early days of internet, but currently it provides very limited information. The reason is, that many applications (intentionally or not) use inconsistent or random ports. Applications also might not be registered in IANA. This approach is also impossible to use for proprietary protocols. It was also reported that only 50 – 70 % of traffic was classifiable using IANA list [3]. Also, the proportion of encapsulated or encrypted traffic is increasing, what makes this approach less usable in the future.

Port number	Application
20	FTP - data
21	FTP - control
22	SSH
23	Telnet
25	SMTP (Simple Mail Transfer Protocol)
53	DNS
80	HTTP
110	POP3
194	IRC
443	HTTPS

Table 1: Port numbers for several popular applications

1.2 Payload based classification

Current methods for reliable traffic classification require examination of packet payload. Classifier simply looks for a application - specific string in a packet payload and compares it with a list of applications with their "signatures" [1]. In Table 2 we can see some application specific strings from packet payload. However, this is a very difficult task to perform. The first problem are privacy and legal issues, because this approach requires capture of whole packet payload and its examination. Another difficulty is the complexity of this operation. If we wanted to capture each packet on a high speed link, our computer would probably soon run out of space and memory. We also have to deal with problem that this signature can be on different place in packet for each application so we need to go through a large amount of data [4]. Though, main problem of the approach is its infeasibility if the packet payload is encrypted.

Application	String
eDonkey2000	0xe319010000
MSN messenger	"PNG"0x0d0a
IRC	"USERHOST"
NNTP	"ARTICLE"
SSH	"SSH"
BitTorrent	0x13BitTorrent
PPLive	0xE903

Table 2: Characteristic strings for several applications

1.3 Host behavior based classification

Host behavior-based approach was developed to capture social interaction observable between host computers. The BLINC [5] captures the profile of a host, based on other hosts it communicates with, applications the host is engaged in by comparing the captured profile (built in) with host behavior signatures of application servers, and then classifies traffic flows. This is done on 3 levels. At social level, behavior of a host indicated by its interactions with other hosts and popularity is captured. At functional level behavior at terms of its functional role in the network (provider, consumer, collaborative application) is captured. At application level, transport layer interactions between hosts are captured, with intent to identify the application of origin. Though, complete and reliable application using BLINC approach have not been developed yet.

1.4 Classification based on flow statistical properties

Because of limitations of previous methods, we treat the problem of classification as a statistical problem. In this approach, we assume that traffic at the network layer has statistical properties which contain enough variance to distinguish among certain classes of applications. This enables us to detect applications in which we are interested. We try to find some different properties of flows that are based on observations and distributions of many various flows. These might include statistics either of individual packets or whole flows. Our goal is to choose statistics that provide us relevant information about flow and are (ideally) different for each application. For each flow packet size distribution and packet inter – arrival times exhibit high amount of variation [6] so they are among statistics commonly used for traffic classification [7, 8, 9]. There are other statistics which are used for traffic classification and these include flow size and duration [9, 10], TCP-specific values (e.g. total payload bytes transmitted, total number of PUSHED packets, total number of ACK packets carrying SACK information etc.) [7], effective bandwidth or entropy based [7, 11]. There are of course many other statistics that can help us to recognize the applications. The extensive list of flow features can be found for example at [12].

2 DATA USED IN ANALYSIS

While some problems solved with machine learning have “standard” dataset (e.g. FERET for faces) on which we can test and train machine learning algorithm, in traffic classification there is no such dataset. In practice, we have to either capture our own traffic, or deal with a limited quantity (and quality) of datasets. There is also problem, that these datasets are mostly not pre – labeled, so for training with supervised learning algorithms we need to put aside a part of dataset and classify it either manually of with a help of some software. However, this approach does not guarantee that our training data is well labeled. Own traffic can be obtained using any packet capturing software. We have to keep in mind that properties of traffic vary at different places in network, or at different time. In this paper we work with part of UNIBS dataset [13]. It is anonymized, payload stripped dataset collected at University of Brescia, which is relatively recent, because it was collected in October/November 2009. The workstations operated the Ground Truth (GT) system which is used for associating accurate ground truth information with internet traffic traces. By probing the monitored host’s kernel to obtain information on open internet sessions, GT gathers guaranteed truth at the application level. UNIBS dataset also provides outcome of the DPI (Deep Packet Inspection) packet payload analysis considering first 200B of data for each packet and signature patterns. Based on this information we are able to infer the application of origin or at least class of application, that generated each network flow. We selected 6 classes of applications (SSH, TELNET, HTTP, POP3, FTP, HTTPS).

3 FEATURES OF NETWORK FLOWS

3.1 Feature extraction

When using statistical methods for network flow classification, we need to obtain features for each flow, which enable us to recognize the application of interest. Authors in paper [12] provide 249 features (also called discriminators) that can be obtained from each flow. Many of these features can be redundant, but many of currently used features is subset of features used here. The features selected for traffic classification usually include: packet sizes, packet inter-arrival times, number of transmitted packets (or bytes), count (or ratio) of packets with TCP flag set, port numbers and protocol. Most of these features should be intact even after encryption, so it might be possible to classify also encrypted traffic. List of features used in our tests is in Table 3.

Feature number	Description
1	Packet sizes
2	Source port
3	Destination port
4	Sent / received packets ratio
5	Duration of the flow
6	SYN flag packet ratio
7	RST flag packet ratio
8	FIN flag packet ratio
9	Sent packet size variance
10	Received packet size variance
11	Protocol
12	#packets in the flow

Table 3: List of tested features

3.2 Feature selection

Feature selection is important part of data processing, because it removes redundant or irrelevant features from our dataset. If we can select the most important features of network traffic we not only increase the precision of classifier, but also significantly reduce its complexity. Identification of the smallest possible set of features is a key part of classifier’s design [14].

Quality of feature set is essential for performance of ML algorithm. Including too many unneeded features has negative impact on precision of most ML algorithms. Also, the number of data that needs to be processed grows with dimensionality of features.

Generally, feature selection consists of several steps as we can see in Figure 1. The first step is generation of candidate subsets. Each subset is evaluated and when the stopping criterion is met, the validity of selected subset is tested. We can divide feature selection methods to 2 classes [14]: filter a wrapper methods, which we are going to discuss later in this paper. These methods usually require algorithm for searching the subspace of features. Number of subset search algorithms can be used, which include [1]: Greedy search, Best first search, Sequential forward selection or use of genetic algorithms.

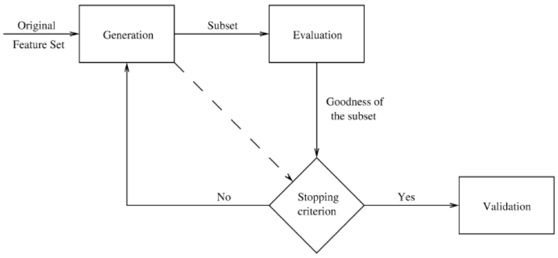

Figure 1 [15]: General scheme of feature selection

3.3 Filter methods for feature selection

Many filter methods for feature selection can be used e.g. Correlation based feature selection (CFS) and Consistency based feature selection. CFS algorithm examines the importance of features by measuring the mutual correlation. The method is based on the hypothesis that good feature subsets contain features highly correlated with the class, yet uncorrelated with each other.

The consistency based approach looks for inconsistent features in the dataset. The pattern is considered inconsistent if there exist at least two instances such that they match all but their class labels. The inconsistency rate of a feature subset is calculated as a ratio of inconsistent patterns from all patterns in the set. Based on this rate, the new feature subsets are selected [14].

3.4 Wrapper methods for feature selection

Wrapper methods select the best feature subset by ML algorithm that will be used for classification. Subset with the highest overall precision is considered to be the best. Because this method needs to execute the ML algorithm for each subset, they might be impractical and complex for big feature sets or slow ML algorithms. The accuracy of classification is of course the highest, if the algorithm for feature selection and classification is the same [15]. Subsets of features generated by wrapper methods, are the upper limit for classification accuracy, but they are not comparable to results obtained by filter methods (because the features are different).

3.5 Common features for traffic classification

Feature selection is usually treated with great attention. In some papers authors focus on classification using just packet sizes. They are usually divided to several intervals (bins). Of course, for classification can be used wide variety of features. For example, authors in [16] select their features from 37 and use

CFS to select the most important features using best first search method. They state the packet size, protocol, port numbers and TCP flag information are the most important features for their method of classification. Use of this selected subset degrades the accuracy of classifier by 0.1 – 1.4 % while the training is 10 times faster. Packet inter – arrival times were not selected to any subset, probably because on their strong dependence of used link.

4 RESULTS

We implement wrapper feature selection method for traffic classification. The traffic is classified into 6 classes : SSH, TELNET, HTTP, POP3, FTP, HTTPS. From each class 200 flows were selected. The Machine learning algorithm for evaluation of feature subsets is Naïve Bayes. We consider the set of features from Table 3. The packet sizes were divided into 4 intervals. We take direction of the packets into account (from client / from server), so each flow is by 8 bins.

For selecting the features to each subset we chose the algorithm of sequential forward selection [17]. The feature which has the maximal classification accuracy is selected as the first component of the resulting feature vector. Then we calculate accuracy for all pairs of features having the first feature fixed. The pair manifesting minimum total classification error is chosen as a starting point for the next step. In similar way, we proceed further for triples, n-tuples, etc., unless all the features are processed. The results in Table 4 show combinations of features with their accuracy. For convenience, we only use the sequence number to represent the features in Table 3. The highest accuracy is printed in bold.

Selected feature	Accuracy
1	68,80%
1, 2	85,50%
1, 2, 4	87,20%
1, 2, 4, 11	86,70%
1, 2, 4, 11, 3	88,10%
1, 2, 4, 11, 3, 6	**88,70%**
1, 2, 4, 11, 3, 6, 8	87,50%
1, 2, 4, 11, 3, 6, 8, 10	87,40%
1, 2, 4, 11, 3, 6, 8, 10, 12	88,00%
1, 2, 4, 11, 3, 6, 8, 10, 12, 9	87,80%
1, 2, 4, 11, 3, 6, 8, 10, 12, 9, 5	87,40%
1, 2, 4, 11, 3, 6, 8, 10, 12, 9, 5, 7	87,60%

Table 4: Features selected by sequential forward selection

5 CONCLUSION

Since there is a great amount of features we can obtain from statistical properties of the flow, feature selection plays a vital role in classification process. In this paper we proposed a wrapper – based method for feature selection using sequential forward selection. The results show that among the most important features in our dataset are packet sizes, source / destination ports, ratio of sent and received packets, protocol and ratio of packets with SYN flag set. Using these features we can obtain classification accuracy 88,7% using simple Naïve Bayes classifier. With a limited set of 3 features, classifier is still able to achieve accuracy just 1,5% lower.

This paper presents only our preliminary results. In future we plan to implement more feature selection methods to identify the most important features of the flow. Also use of other Machine learning algorithms might increase the classification accuracy.

Literature:

1. Nguyen, T. T., Armitage, G.: *A Survey of Techniques for Internet Traffic Classification using Machine Learning*. IEEE Communications Surveys and Tutorials, 2008.
2. Internet Assigned Numbers Authority (IANA). 2010, http://www.iana.org/assignments/port-numbers, last revision 20th November, 2010.
3. Moore, A., Papagiannaki, D.: *Toward the accurate identification of network applications*. Proceedings of the Sixth Passive and Active Measurement Workshop, Springer-Verlag, 2005.
4. Cascarano, N.: *Application Layer Traffic Classification*. Facolta di Ingegneria, Politecnico Di Torino, 2007.
5. Karagiannis, T., Papagiannaki, K., Faloutsos, M.: *Blinc: Multilevel traffic classification in the dark.* In Proceedings of the Special Interest Group on Data Communication conference, 2005.
6. Trivedi, C., Chow, M., Nilsson, A., Trussel, H.J.: *Classification of Internet Traffic using Artificial Neural Networks*. Eusipco2005, 2005.
7. Auld, T., Moore, A. W., Gull, S. F.: *Bayesian neural networks for Internet traffic classification.* IEEE Transactions on Neural Networks, Volume 1. Pages 223–239, 2007.
8. Crotti, M., Dusi, M., Gringoli, F., Salgarelli, L.: *Traffic classification through simple statistical fingerprinting.* SIGCOMM Communications Revue, Volume 37, No. 1, Pages 5–16, 2007.
9. McGregor, A., Hall, M., Lorie, P., Brunskill, J.: *Flow clustering using machine learning techniques*, Juan-les-Pins, France: Proceedings of Passive and Active Measurement Workshop, 2004.
10. Park, J., Tyan, H.R.: *Internet traffic classification for scalable QoS provision*, Toronto, IEEE International Conference on Multimedia, 2006.
11. Moore, A., Zuev, D.: *Internet traffic classification using Bayesian analysis techniques*. Alberta, Canada, ACM International Conference on Measurement and Modeling of Computer Systems, 2005.
12. Moore, A., Zuev, D.: *Discriminators for use in flow – based classification.* Queen Mary University of London, 2005.
13. UNIBS dataset, http://www.ing.unibs.it/ntw/tools/traces/ last revision: 20th november 2010.
14. Witten I.H., Eibe F.: *Data Mining - practical machine learning tools and techniques*. Morgan Kaufman, 2005. ISBN: 0-12-088407-0.
15. Dash, M., Liu, H.: *Consistency – based search in feature selection*, Artificial Intelligence 151, 2003.
16. Kim, H., et al.: *Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices*. ACM CoNEXT, 2008.
17. Ververidis, D., Kotropoulos, C.: *Sequential forward feature selection with low computational cost*. Department of Informatics, Aristotle University of Thessaloniki, 2004.

Primary Paper Section: I

Secondary Paper Section: D, G, L