FEATURE FILTERING TECHNIQUES APPLIED IN IP TRAFFIC CLASSIFICATION

Michael Taynnan Barros[*], Reinaldo Cezar Gomes[$], Marcelo Sampaio de Alencar[&] and Anderson Fabiano Costa[#]

[*] *Telecommunication Software and Systems Group/ Waterford Institute of Technology – Waterford, Ireland*
[$] *Computing and System Department/ Federal University of Campina Grande – Campina Grande, Brazil*
[&] *Institute for Advanced Studies in Communications/ Federal University of Campina Grande – Campina Grande, Brazil*
[#] *Federal Institute of Paraíba – Campina Grande, Brazil*

ABSTRACT

IP traffic classifiers using machine learning algorithms are promising techniques for many applications involving management and network security. One of such applications is Traffic Engineering, with regard to the QoS problem, which is important to map an application into a service level. To develop a model for IP traffic classifiers using machine learning the set of features from a traffic sample needs to be defined properly. This paper presents an analysis of such problem invoking filtering techniques in order to decrease the training time and system complexity. The results show a reduction of the number of features to five, instead of common used 37, keeping a level of accuracy of more than 90%.

KEYWORDS

Filtering, classification, IP Traffic, evaluation.

1. INTRODUCTION

Nowadays, the whole world can communicate by an infrastructure called Internet, which has caused economic and social effects, and its importance is unquestionable [Richards, 2002]. However, the Internet has technical problems, related to the establishment and guarantee of connections, causing a possible change in its current communication paradigm [Zittrain, 2008]. The increase in the number of users is a relevant concern, because reports indicate a number of five billion users in 2020, with more 50 billion entities (including sensors, security and control software) [Tania, 2010]. Then infrastructural problems can emerge, because the current Internet technology is not flexible enough to provide and guarantee the connections establishment and its quality. This scenario requires techniques such as Traffic Engineering (TE) and Quality of Service (QoS).

Briefly, TE establishes the parameter and the operational point for the three aspects of the network: the system of demands (traffic), the system of restrictions (interconnected network elements) and the system of response (network protocols and processes), in an operational context [Awduche, 99]. The most common TE technique is QoS, which guarantees reliability of information transmission according to the application requirements, as well as adequate performance levels [Kamienski, 2001]. Many obstacles have prevented a large scale implementation of QoS [Meddeb, 2010], and its maturity to handle commercial requirements. Nevertheless, there is still some research effort to improve QoS, specifically for mobile networks [Kim, 2011] [Natkaniec, 2011].

The mentioned techniques require an accurate identification of the traffic for the capability to process the traffic flows according to the network requirements. This shows the importance of IP traffic classification, which has gained recently attention of the network research community, focusing more on machine learning classifiers. Plus, traffic classification can be applied in other parts of the network, e.g. security, network managements and planning.

Machine learning based classifiers use the flow approach to identify applications based on statistical characteristics of flows. Works like [Moore, 2005] indicate a total of more than 260 characteristics of streams. This number can compromise the performance of the classifiers, since redundancy may increase the

development time of a classification model, increase the time and decrease the classification accuracy of classifiers. The number and order flows are directly related to the performance of classifiers based on the evaluation of flows [Li, 2007]. Works like [Zander, 2005][Li, 2007] point to the use of techniques filtering characteristics and present relevant data on gains in performance, but no filtering evaluation and comparison.

This paper presents an analysis of the performance of such filters, providing evidences for the importance of its usage in developing classification model with the machine learning classifiers. The evaluation consists of five filtering techniques (Correlation-based Feature Subset Selection, Chi-Squared Attribute Evaluation, Consistency Subset Evaluation, Gain Ratio Attribute Evaluation and OneR Attribute Evaluation) with two supervised machine learning classifiers (Decision tree and Bayesian Networks) and with five dataset from USA and Japan from 2002 to 2011.

The contributions of this work are: the process of filtering flows features increase the accuracy of machine learning classifiers; the number of flow features can be reduced and therefore the complexity and training time should be reduced as well; one can say that decision tree classifier is robust enough to not depend highly on the filtering process.

This paper is organized as follows: Section II discusses the problem statement, Section III presents the filtering techniques used as solutions, Section IV discusses the experimentation process, and is followed by the results in Section V. Finally, the conclusions are presented in Section VI.

2. PROBLEM STATEMENT

The supervised learning creates knowledge structures that support the task of classifying new instances of predefined classes [Reich, 1991]. The output of the learning process is a classification model that is built by analyzing and generalizing from samples previously provided.

The supervised learning focuses on the input/output of the relationships modeling. Its goal is to identify a mapping from input features to an output class. The obtained knowledge (e.g., points in common between members of the same class and differences among competitors) can be presented as a flowchart, a decision tree classification rules, etc., and used later to classify a new sample [Nguyen, 2008]. There are two stages (stages) on supervised learning, which are shown in Figure 1:

Training: learning phase which examines the datum samples (called the set of training data) together with the construction of a classification model. Tests (also known as classification): The model that was built during the training phase is used to classify new cases.

In the training process, the datum samples should have characteristics needed to build the classification model and they need to be enough in number to differ in each data flow. The common thinking is that if there are more characteristics the classification model will be more accurate. This paper analyzes this argument by simulation, to see with filtering techniques, responsible for eliminating some redundant characteristics, has positive effect in the system, increasing the accuracy. One can argue, as well, that the high numbers of characteristics have direct impact in time performance, by increasing the system complexity. Applying filtering techniques may, also, decrease the system complexity and also decrease the time for training and testing.

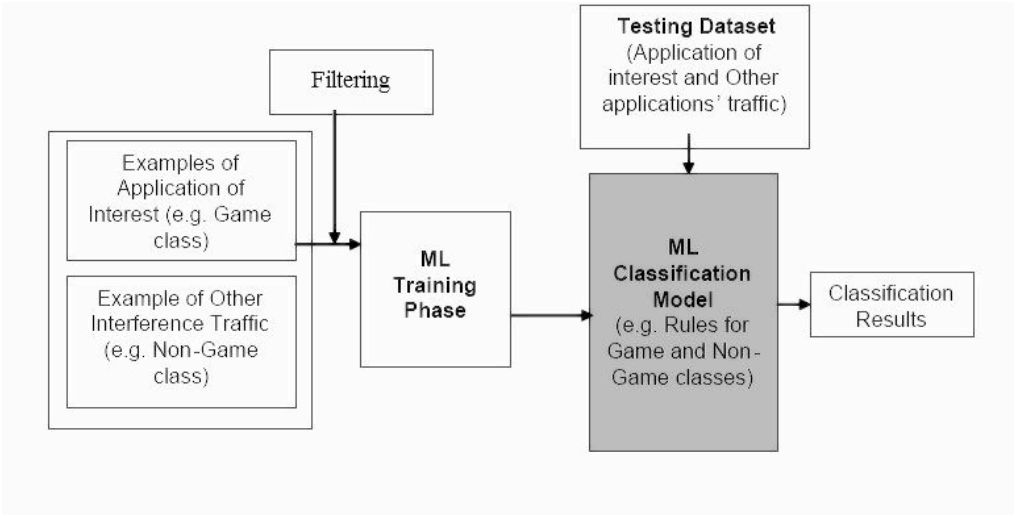

Figure 1. Logic view of the both traffic classifier and conditioner [Nguyen, 2008].

3. FILTERS

3.1 Correlation-based Feature Subset Selection - CfsSubsetEval

The CfsSubsetEval evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [Hall, 1998]. The following equation gives the merit of a feature subset consisting of features:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}.$$

Here, $\overline{r_{cf}}$ is the mean of all feature to classification correlations, and $\overline{r_{ff}}$ is the mean of all feature to feature correlations. The CfsSubsetEval criterion is defined as:

$$Cfs = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \cdots + r_{f_if_j} + \cdots + r_{f_kf_1})}} \right]$$

The r_{cfi} and r_{fifj} variables are referred to as correlations, but are not necessarily Pearson's correlation coefficient or Spearman's ρ.

3.2 Chi-Squared Attribute Evaluation - ChiSquaredAttributeEval

The ChiSquaredAttributeEval evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.

3.3 Consistency Subset Evaluation - ConsistencySubsetEval

The ConsistencySubsetEval evaluates the worth of a subset of attributes by the consistency in the class values when training instances are projected onto the subset of attributes [Liu, 1996]. Consistency of any subset can never be lower than that of the full attributes set, so, one must use this filtering technique in conjunction with a Random or Exhaustive search which looks for the smallest subset with consistency equal to that of the full set of attributes.

3.4 Gain Ratio Attribute Evaluation – GainRatioAttributeEval

The GainRatioAttributeEval evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

$$GainR(Class, Attribute) = \frac{(H(Class) - H(Class|Attribute))}{H(Attribute)}.$$

3.5 OneR Attribute Evaluation - OneRAttributeEval

The OneRAttributeEval evaluates the worth of an attribute by using the OneR classifier. OneR takes as input a set of samples, each with loads of attributes and classes. It aims to deduct a rule that predicts the class given the values of the attributes. The OneR algorithm chooses the attribute with more information and bases the rule itself on this particular attribute [Nevill-Manning, 95]. It also assumes that the attributes are discrete. If not, then they must be discretized. Missing values are handled in the algorithm by treating them as a separate value in the enumeration of an attribute [Holte, 1989].

4. SIMULATION SET-UP

In academic works, as in [Zander, 2005][Li,2007], it is presented evidences that the quantity and combination of statistical characteristics of flows change the outcome of the classification. Considering this, it is necessary to better evaluate the techniques that filter characteristics from a flow set, eliminating redundancy in some cases, benefiting performance and defining the most appropriate number and order of such characteristics.

For this evaluation are used 37 features for unidirectional flow: protocol, source and destination ports, the number of packets, transferred bytes, the number of packets without Layer 4 payload, start time, end time, duration, average packet throughput and byte throughput, max/min/average/standard deviation of packet sizes and inter-arrival times, number of TCP packets with FIN, SYN, RSTS, PUSH, ACK, URG (Urgent), CWE (Congestion Window Reduced), and ECE (Explicit Congestion Notification Echo) flags set (all zero for UDP packets), and the size of the first ten packets.

We use two independent variables that are the classifier and the filtering technique of flow characteristics. The classifiers chosen are Decision Trees and Bayesian Networks.

They have been chosen within a bigger set, starting with 20 techniques, that can be found in the tool used in this study (WEKA). In this set some techniques showed no appropriate convergence of the results and they were excluded. The classifiers were evaluated by means of accuracy, precision, recall and f-measure. To classify a given application one single packet is not enough, you need a grouping of packets into flows, defined by a tuple of five elements: IP source address, IP destination address, source port number, destination port number and protocol of the transport layer.

Classifiers that use machine learning identify applications based on statistical characteristics of the flows. For this work will be computed 37 features unidirectional flow already presented.

Five samples of Internet IP traffic were collected at the edge network backbone and can be found in [CAIDA, 2011] and [WIDE, 2011]. They have temporal and geographic diversity, and belonging to a period from 2002 to 2011 and located in the United States and Japan. The samples are: Link CAIDA OC48 U.S. in 2002 and 2003, CAIDA Backbone Network Chicago USA in 2010, Backbone Network are CAIDA Francisco-USA Network in 2011 and BackBone WIDE-JP (M) in 2011. These data sets were selected to allow us to have a temporal and geographical distribution of data, making data more heterogeneous and, consequently, considering the diversity expected from networks around the world.

The classification of applications is done by identifying traffic flow IP in 11 categories. Table I presents all applications and their categories for port-based classification.

Table 1. Application Categories

Category	Application/Protocol
WEB	HTTP, HTTP
P2P	FastTrack, eDonkey, BitTorrent, Ares,Direct Connect, Gnutella, WinMX, OpenNap, MP2P, SoulSeek, FileBEE, GoBoogy, Soribada, PeerEnabler, Napster Blubster, FileGuri, FilePia IMESH, ROMNET, HotLine, Waste
FTP	FTP
DNS	DNS
Mail/News	BIFF, SMTP, POP, IMAP, IDENTD, NNTP
Streaming	MMS(WMP), Real, Quicktime, Shoutcast, Vbrick Strmg, Logitech Video IM,Backbone Radio, PointCast, ABACast
NetOp	Netbios, SMB, SNMP, NTP, SpamAssasin, GoToMyPc, RIP,ICMP, BGP, Bootp, Traceroute
Encryption	SSH, SSL, Kerberos, IPSec, ISAKMP
Games	Quake, HalfLife, Age of Empires, DOOM, WOW, Star Sieze, Everquest, Startcraft, Asherons,
	Battle Feld Vietnam, HALO
Chat	AIM, IRC, MSN Messenger, Yahoo messenger, IChat, QNext, MS Netmeet, PGPFone, TALK.
Unknown	-

Figure 2 presents all the flows quantities in their respective categories mapping applications by color, beings the ground-truth of the samples used in the article. The purpose of this figure is to show the percentage of applications in each sample in favor of fairer conclusions, based on the available number of flows in each application for classification.

To increase the validity of the results, we use the full factorial experimental design to combine the variables. Based on cited data, the test can formulate like: *5 samples x 5 Filtering techniques x 2 classifiers = 50 test*.

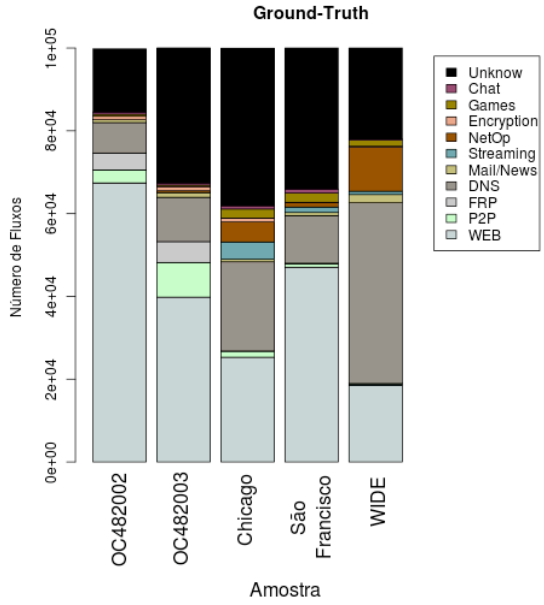

Figure 2. Ground Truth mapping number of flows, samples, and application

The randomness of the experiment is found in five samples of Internet traffic, a number that is meaningful to draw valid conclusions [Barros, 2012]. The experiments consist of three blocks. The first block is to capture, analyze and pre-processing of the samples, including the filtering process. The second block carries out inferences on all samples following units experimental treatments. Finally, in third block, the results obtained with the measurement of the dependent variables are statistically analyzed and conclusions are obtained.

We used in each test the 10-fold-cross-validation technique to train and construct the model for the flow classification of each sample. In Tables II, III, IV, V and VI, we can find the techniques set up used in this paper.

Table 2. CfsSubsetEval Set Up

Parameter	Value
locallyPredictive	True
missingSeparate	False
search	BestFirst -D 1 -N 5

Table 3. ChiSquaredAttributeEval Set Up

Parameter	Value
binarizeNumAttributes	False
missingMerg	True
search	Ranker -T 2.0 -N -1

Table 4. ConsistencySubsetEval Set Up

Parameter	Value
search	BestFirst -D 1 -N 5

Table 5. GainRatioAttributeEval Set Up

Parameter	Value
search	Ranker -T 2.0 -N -1

Table 6. OneRAttributeEval Set Up

Parameter	Value
evalUsingTrainingData	False
folds	10
minimumBucketSize	6
search	Ranker -T 2.0 -N -1

5. RESULTS AND DISCUSSION

In Figure 3, one can find the full results of each filter for Accuracy, Precision, Coverage and F-measure. It can be seen that CfsSubsetEval presents the highest values for all metrics. Note also that the performance of ConsistencySubsetEval is relevant, however, due to the variation of the results, one can not guarantee that it outperforms ChiSquareAttributeEval and without filters.

It is noteworthy that the number of features obtained by the technique CfsSubsetEval is five, which is smaller than that of the 37 previously chosen. One can appreciate that the relation of the higher number of characteristics the better is classification performance is not true, indicating that the number of characteristics is not itself a variable very important, but a certain number with a certain order of characteristics.

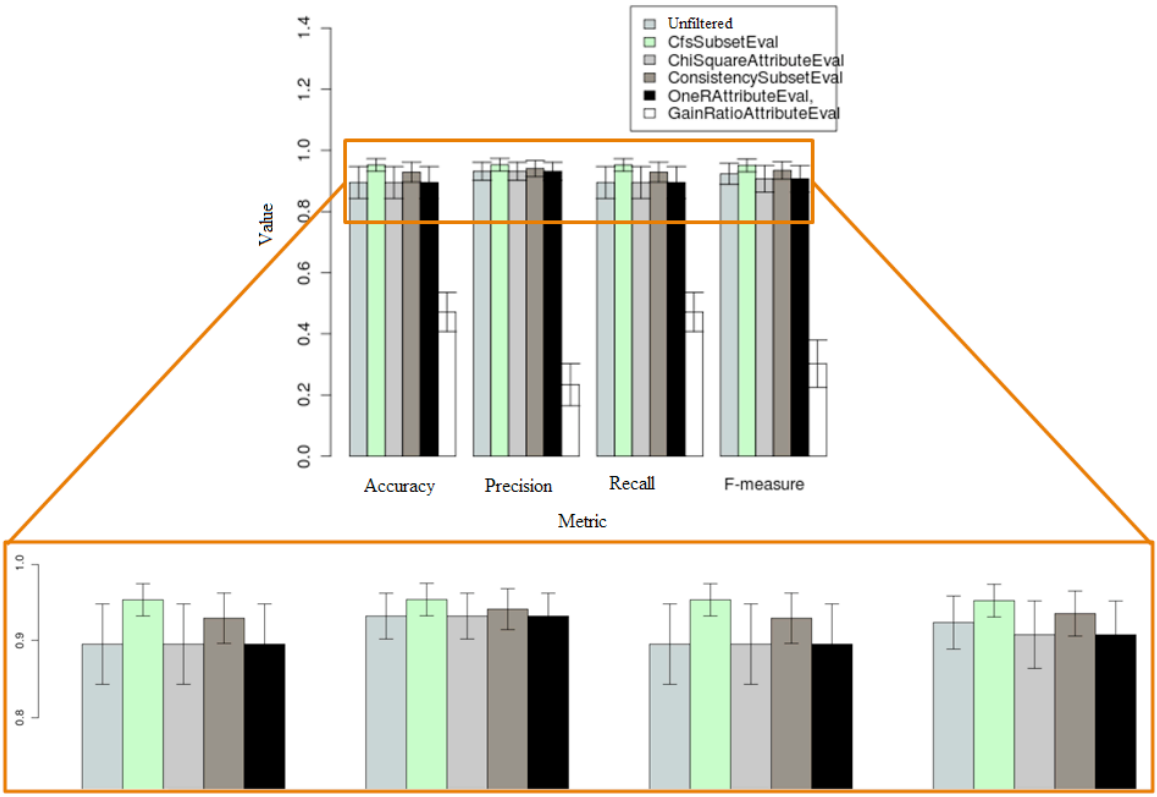

Figure 3. Total performance of the filters.

With the results presented, indicating the use of the filter CfsSubsetEval by presenting a gain in performance of the classifiers and to have a low number of flow characteristics, which reflects the complexity of the performance of classifiers.

6. CONCLUSION

This paper presents a performance evaluation of the filtering algorithm of flow characteristics of dataset for IP traffic classifiers based on machine learning algorithms. The evaluation considers five filtering techniques (CfsSubsetEval, ChiSquareAttributeEval, ConsistencySubsetEval, GainRatioAttributeEval, OneRAttributeEval) with five dataset from USA and Japan from 2002 to 2011. The results indicate a gain in performance with the CfsSubsetEval that indicates the use of only five characteristics to achieve a level of more than 90% of accuracy.

From the results one can argue that there is a dependency between the dataset and the flow feature results, this drives to the conclusion that the use of filtering techniques should be the answer, not a specific dataset with specific characteristics. This conclusion is good but leads to another questions, which is how to apply these filtering techniques on on-line classifiers. One can say, as well, that decision tree classifier is robust enough to not depend highly on the filtering process, but again, this conclusion leads to another question, which is if this behavior stays the same for the other machine learning traffic classifiers.

Future works include re-evaluation of machine learning classifiers, because the filtering techniques have effect in some classifiers performance, having the possibility to improve their results. This could lead to more fare conclusions. Also, more analysis could be done, like evaluating the performance of on-line traffic classifier with these filtering algorithms.

ACKNOWLEDGEMENT

Thanks to Iecom for infrastructure support, and CAPES for financial support.

REFERENCES

S. Richards, FutureNet: The Past, Present, and Future of the Internet as Told by Its Creators and Visionaries, Wiley, Ed. Wiley, 2002.

J. Zittrain, The Future of the Internet–And How to Stop It, Caravan, Ed. Caravan, 2008.

D. O. Awduche, "MPLS and traffic engineering in IP networks," *IEEE Communications Magazine*, vol. 37, pp. 42–47, 1999.

C. A. Kamienski e Djamel Sadok, "Chameleon: Uma arquitetura para serviços fim a fim na Internet," in *Anais do Simpsio Brasileiro de Redes de Computadores e Sistemas Distribudos*, 2001.

A. Meddeb, "Internet QoS: pieces of the puzzle," *Comm. Mag.*, vol. 48, pp. 86–94, January 2010.

K. Kim, "A distributed channel assignment control for QoS support in mobile ad hoc networks," *J. Parallel Distrib. Comput.,* vol. 71, pp. 335–342, March 2011.

M. Natkaniec, K. Kosek-Szott, S. Szott, J. Gozdecki, A. Glowacz, and S. Sargento, "Supporting QoS in integrated ad-hoc networks," *Wirel. Pers. Commun.*, vol. 56, pp. 183–206, January 2011.

A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," *in ACM SIGMETRICS international conference on Measurement and modeling of computer systems, ser. SIGMETRICS '05*. New York, NY, USA: ACM, 2005, pp. 50–60.

Z. Li, R. Yuan, and X. Guan, "Accurate classification of the Internet traffic based on the SVM method," *in IEEE International Conference on Communications*, june 2007, pp. 1373 –1378.

S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," *in The IEEE Conference on Local Computer Networks*, nov. 2005, pp. 250 –257.

Y. Reich and S. J. Fenves, "The formation and use of abstract concepts in design," *in Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann, 1991, pp. 323–353.

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

T. Tronco, New Network Architectures: The path to the future Internet, Springer, Ed. Springer, 2010.

M. A. Hall, "Correlation-based feature subset selection for machine learning," *Ph.D. dissertation, University of Waikato, Hamilton*, New Zealand, 1998.

R. S. H. Liu, "A probabilistic approach to feature selection - a filter solution," *in In: 13th International Conference on Machine Learning*, 1996.

C. Nevill-Manning, G. Holmes, and I. H. Witten, "The development of holte's 1r classifier," 1995.

R. C. Holte, L. E. Acker, and B. W. Porter, "Concept learning and the problem of small disjuncts," *in Proceedings of the 11th international joint conference on Artificial intelligence - Volume 1, ser. IJCAI'89*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 813–818.

CAIDA, http://www.caida.org/home/., Accessed in August 2011.

WIDE, http://mawi.wide.ad.jp/mawi/., Accessed in August 2011.

M. Barros, R. de Morais Gomes, M. Alencar, P. J´unior, and A. Costa, "Avaliação de classificação de tráfego IP baseado em aprendizagem de máquina restrita à arquitetura de serviços diferenciados," *Revista de Tecnologia da Informação e Comunicação – (RTIC),* vol. 1, 2012, pp. 10–20.