Indian Journal of Research in Applied Sciences Engineering (IJRASE)
Vol.2.No.1 2014pp1-9
available at: www.goniv.com
Paper Received :05-03-2014
Paper Published:28-03-2014
Paper Reviewed by: 1. John Arhter 2. Hendry Goyal
Editor : Prof. P.Muthukumar

EFFECTIVE INTERNET TRAFFIC CLASSIFICATION BY NAIVE BAYES PREDICTIONS AND C5.0

Saleembasah.N, Murugasen.M,
Dhanalakshmi college of Engineering, Chennai
Tamil Nadu, India
bashasaleem599@gmail.com

ABSTRACT

I presents a novel traffic classification scheme to improve classification performance when few trainingdata are available. In the proposed scheme, trafficflowsaredescribed using the discretized statistical features andhistorical dataflow correlation information is modeled by bag-of-flow (BoF). IsolvetheBoF-based traffic classification in a classifier combinationframework and theoretically analyze the performance benefit.Furthermore, a new BoF-based traffic classification of C5.0 method isproposed to aggregate the naive Bayes (NB) predictions of thecorrelatedflows. I also present an analysis on prediction errorsensitivity of the aggregation strategies. Finally, a large number ofexperiments are carried out on two large-scale real-world trafficdatasets to evaluate the proposed scheme and detect the attribute using correlated coefficient to detect unwanted large attribute . The experimentalresults show that the proposed scheme can achieve much betterclassification performance than existing state-of-the-art traffic using classification methods.

***Index Terms*—**Traffic classification, network security, naïveBayes,c5.0

I. INTRODUCTION

APPLICATION oriented traffic classification is a fundamental technology for modern network security. It is useful to tackle a number of network security problems including lawful interception and intrusion detection [1]. Forexample, traffic classification can be used to detect patternsindicative of denial of service attacks, worm propagation, in-trusions [2], and spam spread. In addition, traffic classificationalso plays an important role in modern network management,

such as quality of service (QoS) control. Many open source and commercial tools [3], [4] with traffic classification function have been deployed and there is an increasing demand on the development of modern traffic classification techniques [1], [5].

While traditional traffic classification technique marelyon the port numbers specified by different applications or the sig-nature strings in the payload of IP packets, modern techniques

Manuscript received December 06, 2011; revised April 29, 2012; acceptedOctober 03, 2012.Date of publication October 09, 2012; date of current versionDecember 26, 2012. This work was supported in part by ARC Discovery ProjectDP1095498, in part by ARC Linkage Project LP100100208, and in part by ARCLinkage Project LP120200266. The associate editor coordinating the review of this manuscript and approving it for publication was C.-C. Jay Kuo.The authors are with the School of Information Technology, DeakinUni-versity, Melbourne, 3125, Australia (e-mail: jun.zhang@deakin.edu.au; chao.chen@deakin.edu.ayangxiang@deakin.edu.au; wanlei@deakin.edu.au; yong.xiang@deakin.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org
Digital Object Identifier 10.1109/TIFS.2012.2223675

normally utilize host/network behavior analysis or flow level statistical features by taking emerging and encrypted applica- tions into account [6], [7].

Recently, substantial attention has been paid on the application of machine learning techniques to statistical features based traffic classification [1]. In the state-of-the-art traffic classification methods, Internet traffic is characterized by a set of flow statistical properties and machine learning techniques are applied to automatically search for structural pat terns. These methods can address the problems suffered from by the traditional methods, such as dynamic port numbers and userprivacy protection.

Recent research shows that flow statistical feature based traffic classification can be enhanced by feature discretization.Particularly, feature discretization is able to dramatically affect the performance of naive Bayes (NB). NB is one of the earliest classification methods applied in Internet traffic classification [7], which is a simple and effective probabilistic classifier employing the Bayes' theorem with naive feature independence assumptions [8]. Since independent features are assumed, an advantage of the NB classifier is that it only requires a small amount of training data to estimate the parameters of a classification model. However, the performance degradation of NB traffic classifier is reported in the existing works [5], [9]. Lim *et al.* found that the main reason for the underperformance of a number of traditional classifiers including NB is the lack of the feature discretization process [10]. For example, feature discretization can effectively improve the accuracies of the support vector machine (SVM) and -NN algorithms t the price of lower classification speed. More interestingly, NB with feature discretization demonstrates not only significantly higher accuracy but also much faster classification speed.

Considering complex network situation, a difficult question is that how to obtain a high-performance statistical feature basedtraffic classifier using a small set of training data. The solutions to this question are essential to address a number of difficult problems in the field of network security and management. For instance, in practice, we may only manually label very few samples as supervised training data since traffic labelling is time-consuming, especially for new applications and encrypted applications. Moreover, a big challenge for current network management is to handle a large number of emerging applications, where it is almost impossible to collect sufficient training samples in a limited time. These observations motivate our work.

In this paper, we provide a solution to effectively Improve NB-based traffic classifier with a small set of training samples

The idea is to seamlessly incorporate flow correlation [11] into the NB-based classification process with feature discretization.

Our major contributions are as follows

- We propose a new traffic classification scheme to utilize the information among the correlated traffic flows g enerated by an application. In the proposed scheme, bag-of-flow (BoF) is introduced for modelling correlated flows and the new BoF-based traffic classification is solved by aggregating correlated NB predictions.
- We provide a theoretical study on the proposed scheme.

  First, we explain why the proposed scheme does work in a theoretical framework of classifier combination. Second, we analyze the sensitivities to prediction errors of different aggregation rules employed in the proposed scheme.
- We present a comprehensive evaluation of the proposed scheme on two large scale real-world network datasets. The empirical study shows that the proposed scheme can effectively improve the traffic classification performancewith a small set of training data and it outperforms theexisting state-of-the-art traffic classification methods. Allcode and data related to this work will be available on request.

The remainder of the paper is structured as follows. Section II reviews some related works. The new traffic classificationscheme is proposed in Section III.Section IV presents the experimental results followed by a theoretical analysis on error sensitivity in Section V. Finally, the paper is concluded in Section VI.

II. Related Work

In the area of network traffic classification, the state-of-the-artmethods employ flow statisticalfeatures and machine learning techniques [1]. Many supervised classification algorithms and unsupervised clustering algorithms have been applied to categorize Internet traffic. In supervised traffic classification, the trafficclasses are predefined according torealapplications and a set of labelled training samples are also manually collected for classifier construction. In contrast, the clustering-based methods canautomatically group a set of unlabeled training samples and use the clustering results to train a traffic classifier. However, the number of clusters has to be set large enough to obtain useful and accurate traffic clusters, which results in a problem of mapping from a large number of traffic clusters to a small numberof real applications [12]–[16]. This problem is very difficult to solve without knowing any information about real applications.

A lot of effort has been made to develop effective supervised methods with the consideration of various network applications and situations. In early works, Moore and Zuev [7] applied the naive Bayes techniques to classify network trafficbased on the flow statistical features. Later, several well-known algorithms were also applied to traffic classification, such as Bayesian neural networks [17] and support vector machines [18]. Erman *et al.* [19] proposed to use unidirectional statistical features to facilitate traffic classification in the networkcore. Taking into account the real-time purpose, several supervised classification methods [20], [21] were proposed, which only used the first few packets. Other existing works include the Pearson's chi-Square test based technique [22], probability density function (PDF) based protocol fingerprints [23], and small time-windows based packet count [24]. Different methods may have their own advantages in different network situations.

Some empirical study [25], [9], [5], [26] evaluated the traffic classification performance of different methods for practical usage. Roughan*et al.* [25] have tested NN and LDA methods for traffic classification using five categories of statistical features. Williams *et al.* [9] compared the supervised algorithms including naive Bayes with discretization, naive Bayes with kernel density estimation, C4.5 decision tree, Bayesian network and naive Bayes tree. Kim *et al.* [5] extensively evaluated ports-based CorelReef method, host behavior-based BLINC method and seven common statistical feature based methods using supervised algorithms on seven different traffic traces. A recent research finding is that feature discretization is critical and essential for Internet traffic classification [10]. By investigating the reasons for C4.5 performing very well under any circumstances, Lim *et al.* discovered that feature discretization can substantially improve the classification accuracy of every tested machine learning algorithm [10].

Since the performance of supervised methods is Sensitive to the size of training data, some proposals tried to address this problem. Erman*et al.* [27] proposed to use a set of supervised training data in an unsupervised approach to address the problem of mapping from flow clusters to real applications. However, the mapping method will produce a large proportion of 'unknown' clusters, especially when the supervised training data is very small. Another recent research finding is that flow correlation can be beneficial to traffic classification. Ma *et al.* [11] proposed a payload-based clustering method for protocol inference, in which they grouped flows into equivalence clusters using a 3-tuple heuristic, i.e., the flows sharing the same destination IP, destination port and transport layer protocol are generated by the same application. Canini*et al.* [28] tested the correctness of the 3-tuple heuristic with real-world traces. In our previous work [29], we applied the heuristic to improve unsupervised traffic clustering. However, it is unclear why flow correlation is helpful to traffic classification and how to apply flow correlation in the supervised classification approach. The problem of how to effectively classify network traffic using a small set of training data, is still to be solved.

III. Proposed Classification Scheme

NaiveyBaiyes is one of the earliest classification methods applied in Internet traffic classification which is a simple and effective probabilistic classifieremployingthe Bayes' theorem with naive feature independence assumptions .It assumes independent features. NaiveyBaiyes classifier is that it only requires a small amount of training data to estimate the parameters of a classification model NaiveyBaiyes with feature discretization demonstrates not only significantly higher accuracy but also much faster classification speed.

NB effectively improves the accuracies of the support vector machine (SVM) and –N algorithms at the price of lower classification speed.

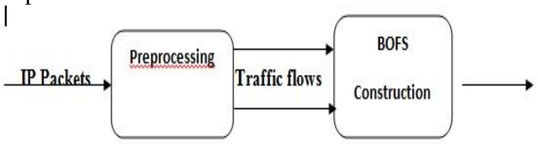
NB-based traffic classifier improves classification with a small set of training samples.

1 Analyzing the Data set

A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values will normally all be of the same kinHowever, there may also be "missing values", it need to be inditiy the miss.

2 Classification Process:

It is based on a flow-level traffic classification. The system captures IP packets crossing a target network and constructs traffic flows by checking the headers of IP packets Itisflow-level traffic classification. A flow consists of successive IP packets with the same 5-tuple: source IP, source port, destination IP, destination port, and transport layer protocol. It uses heuristic way to determine the correlated flows and model them. If the flows observed in a certain period of time share the same destination IP, destination port, and transport layer protocol, they are determined as correlated flows and form a BoF. For the classification purpose, a set of flow statistical features are extracted and discretized to represent traffic flows

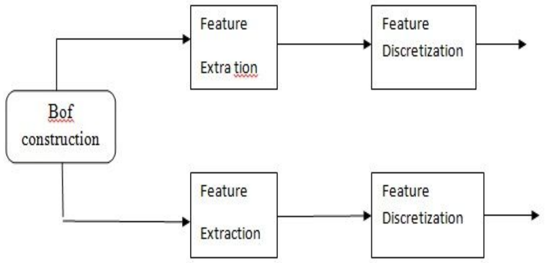


.

Figure Classification Process

A BoF-Based Classification Framework:

In this a set of correlated flows are generated by the same application, which is modeled using a bag of flows BoF. A novel approach is proposed for traffic classification, namely aggregation of correlated NB predictions, which consists of two steps. In the first step, the single NB predictor produces the posteriori class-conditional probabilities for each flow.



.Classification Framework

Single NB Predictor: NB algorithm to produce a set of posterior probabilities as predictions for each testing flow. It is different to the conventional NB classifier which directly assigns a testing flow to a class with the maximum posterior probability. Considering correlated flows, the predictions of multiple flows will be aggregated to make a final prediction

Aggregated Predictor: Under Kittler's theoretical framework, a number combination methods can be derivedfrom the Bayesian decision theory which can be used for aggregated predictor.

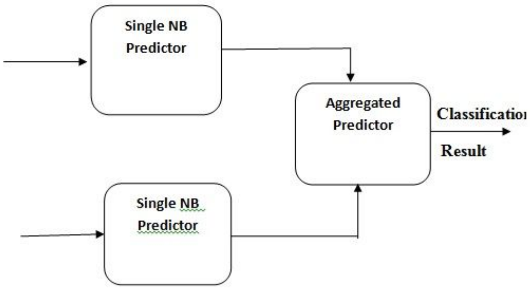


Figure Aggregate Predictor

3.Multi boosting

The effect of combining different classifiers can be explained with the theory of bias-variance decomposition. Bias refers to an error due to a learning algorithm while variance refers to an error due to the learned model. The total expected error of a classifier is the sum of the bias and the variance. In order to reduce bias and variation, some ensemble approaches have been introduced: Adaptive Boosting(AdaBoost) ,Bootstrap Aggregating (Bagging),Wagging and Multiboosting. This is why the idea emerged of combining both in order to profit from theadvantages of both algorithms and obtain overall error reduction

Algorithm Description Naive Bayes Predictions

Definition:

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independentfeature model". Naive Bayes belongs to a group of statistical techniques that are called 'supervised classification' as opposed to 'unsupervised classification.' In 'supervised classification' the algorithms are told about two or more classes to which texts have previously been assigned by some human(s) on whatever basis.

Explanation

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing inBayesian probability or using any Bayesian methods.

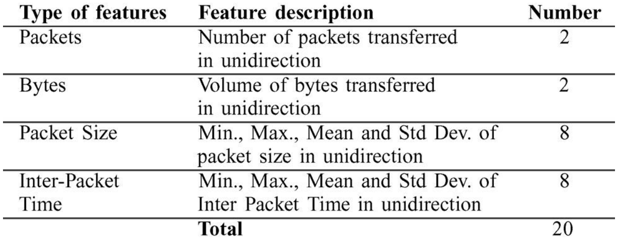
In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers.[1] Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests.[2]

An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

IV. Experimental Evaluation

In this section, we evaluate the proposed BoF-NB scheme on two real-world traffic datasets. The proposed BoF-NB scheme is compared to four state-of-the-art traffic classification method including C4.5, k-NN, NB [10] and Erman'ssemisupervisedmethod [27] in the situation of a small number of supervised training samples.

To establish the ground truth for the testing datasets,we Have developed a deep packet inspection (DPI) tool that matches regular expression signatures against flow payload content [29]. A number of application signatures are developed based on previous experience and some well-known tools such as l7-filter (http://l7filter.sourceforge.net) and Tstat (http://tstat.tlc.polito.it). Also, several encrypted and new applications are investigated by manual inspection of the unidentified traffic.



Type of features	Feature description	Number
Packets	Number of packets transferred in unidirection	2
Bytes	Volume of bytes transferred in unidirection	2
Packet Size	Min., Max., Mean and Std Dev. of packet size in unidirection	8
Inter-Packet Time	Min., Max., Mean and Std Dev. of Inter Packet Time in unidirection	8
	Total	20

Fig. 2. Impact of feature discretization (a) on isp dataset and (b) on

[34] and *isp* [29], respectively. The wide dataset consists of 182 k traffic flows which are randomly selected from the *wide* trace and carefully recognized by the DPI tool and manual inspection. All flows in the wide dataset are categorized into 6 application oriented classes. For the wide dataset, there are only a small number of classes and the HTTP flows dominate the whole dataset. The other is the isp dataset created from our *isp* trace.

The isp dataset consists of 200 k flows randomly sampled from 11 major classes. To avoid the dominating classes, we randomly select up to 30 k flows from every class. The wide and ispdatasets can well represent the different natures of two real-world network traffic traces. A large number of experimental results obtained on the two datasets with different characteristics are statistically significant. The experimental results can effectively demonstrate the classification capability of various traffic classification methods.

In the experiments, 20 unidirectional flow statistical Features are extracted and used to represent traffic flows, which are listed in Table I. We apply feature selection to remove irrelevant and redundant features from the feature set [35], [5]. The correlation-based feature subset selection is used in the experiments, which searches for a subset of features with high class-specific correlation and low intercorrelation. A Best First search [36] is used to create candidate sets of features. The process of feature selection [36] yields 6 features for the isp dataset and 6 features for the wide dataset, respectively. Feature discretization can significantly improve the classification performance of many supervised classification algorithms [10]. We also incorporate feature discretization [37] into our proposed scheme.

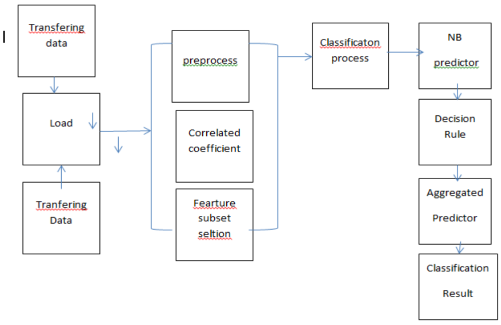
Two common metrics are used to measure the classificationperformance [5], overall accuracy and F-Measure. Overall accuracy is the ratio of the sum of all correctly classified flows to the sum of all testing flows. This metric is used to measure the

Classification Process

Below figure illustrates the classification process ofour proposed scheme, which is focused on flow-level traffic classification.In the preprocessing, the system captures IP packets crossinga target network and constructs traffic flows by checking theheaders of IP packets. A flow consists of successive IP packets with the same 5-tuple: source IP, source port, destination IP,destination port, and transport layer protocol.

We apply aheuristic way to determine the correlated flows and model them

using “bag-of-flows (BoF)”. If the flows observed in a certain period of time share the same destination IP, destination port,and transport layer protocol, they are determined as correlatedflows and form a BoF. For the classification purpose, a setnamely aggregation of correlated NB predictions,which consists of two steps. In the first step, the single NB predictor produces the posteriori class-conditional probabilities for each flow. In the second step, the aggregated predictor aggregates the flow predictions (posteriori probabilities) to determine the final class for BoFs



B. A BoF-Based Classification Framework

In the proposed scheme, a set of correlated flows are generated the ated by the same application, which is modelled using a bag of belong to the same application-based class, such correlation in formation can be utilized to improve the classification results.Therefore, we aim to aggregate the individual predictions of the correlatedflows so as to conduct more accurate classification.Our research shows that the goal can be achieved by following the approach of classifier combination. The BoF-based classification can be fitted into Kittler’s theoretical framework [30] for classifier combination.Consider a traffic classification problem where pattern (BoFis to be assigned to one of the possible traffic classes. Let us assume that we have a classifier, but thegiven pattern can be represented by using distinct measurement vectors (flows in this BoF),This is atypical classifier combination architecture of repeated measure-ments [31]. In the measurement space, each classis modelled by the probability density function and its *priori*probability of occurrence is denoted by According to theBayesian decision theory, given measurements the pattern (BoF) should be assigned to class provided thea posteriori probability of that interpretation is maximum

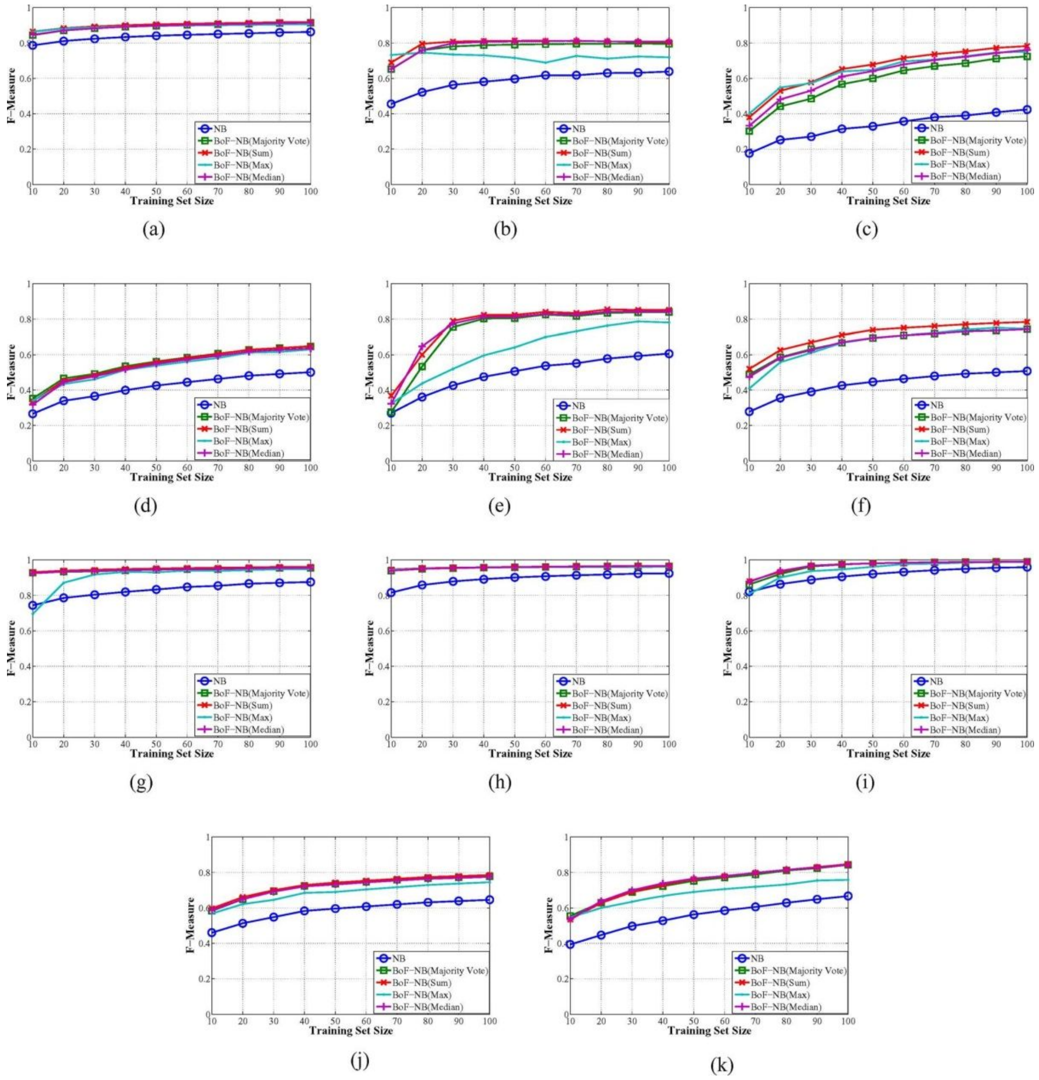
posterior×Likelihood
prior= Evidence

Impact of Feature Discretization

Firstly, a set of experiments are carried out to evaluate the effect of feature discretization. Fig. 2 reports the classificationaccuracy of NB with and without feature discretization on theisp and wide datasets.As shown in Fig. 2(a), on the ispdataset,feature discretization can improve the classification accuracy by approximately 5 percent when only 10 training samples areavailable for each class. The improvement increases with the rise of the training samples and it can achieve up to 20 per-cent. The results on the wide dataset (see Fig. 2(b)) is similar to that on the isp dataset, while the maximum improvement can be 30 percent. The experimental results demonstrate the benefitof feature discretization, i.e., feature discretization can signifi-cantly improve the classification accuracy of the NB classifier.Therefore, similar to [10], we apply feature discretization in our proposed scheme.

Impact of Aggregation Methods

We perform a set of experiments to evaluate the proposed BoF-NB scheme with different aggregation methods. The orig-inal NB classifier with feature discretization is used in the experiments as a baseline. Fig. 3 shows the classification accu-racy with different training date sizes. One can find that the pro-posed BoF-NB scheme outperforms NB whichever aggregation method is used. On the isp dataset, the classification accuracy of BoF-NB is higher than that of NB by about 10 percent.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 4. F-Measures of BoF-NB with different aggregation rules on isp dataset. (a) bt, (b) dns, (c) ftp, (d) http, (e) imap, (f) msn, (g) pop3, (h) smtp, (i) ssh, (j) ssl, and (k) xmpp.

The reason is that BoF-NB can effectively utilize the flowcorrela-tion information. Regarding the aggregation methods, the sumrule is slightly better than the majority vote rule and the median rule. The max rule is the worst one among the four competing aggregation methods, whose accuracy is lower than the sum rule by approximately 4 percentsum rule by approximately 4

percent. The similar results can be obtained on the wide dataset as shown in Fig. 3(b). BoF-NB exhibits better classification capability than NB anthe

Figs. 4 and 5 report the F-Measures of BoF-NB and NB for each class on the two datasets. In general, our BoF-NB scheme, especially with the sum rule,

The degree of improvement varies in different classes. For example on the isp dataset as shown in Fig. 4, the F-Measure of BoF-NB with the sum rule is morethan 15 percent greater than that of NB for dns class. In the class pop3, the improvement is about 10 percent. Among the four aggregation methods, the max rule does not work as well as otheraggregation methods for many traffic classes. For instance, the

F-Measure of using the max rule is lower than that of other rules by up to 15 percent for imap on the isp dataset. On the wide dataset as shown in Fig. 5, BoF-NB with the max rule has similar accuracy to NB. However, the sum rule consistentlydemonstratesgoodclassification performance for all traffic classes on the two datasets.

ANALYSIS ON ERROR SENSITIVITY

In order to explain why the sum rule works better than the max rule, we investigate the error sensitivity. An empirical finding reported in Section IV is that the sum rule (13) appearsto produce more reliable decisions than the max rule (15). I shall show that the sum rule is much less affected by prediction errors. This theoretical analysis result is consistent with the experimental finding.

In Section III, we assumed that the a posteriori class probabilities forXi,,P(Wj,|Xi a flow are computed correctly. In fact, each flow will produce only an estimate of the posteriori class probability for a BoF, which is denoted as the estimatedeviates from the probability which is denoted a The estimate deviates from the probability P(wj|x) by error eji,

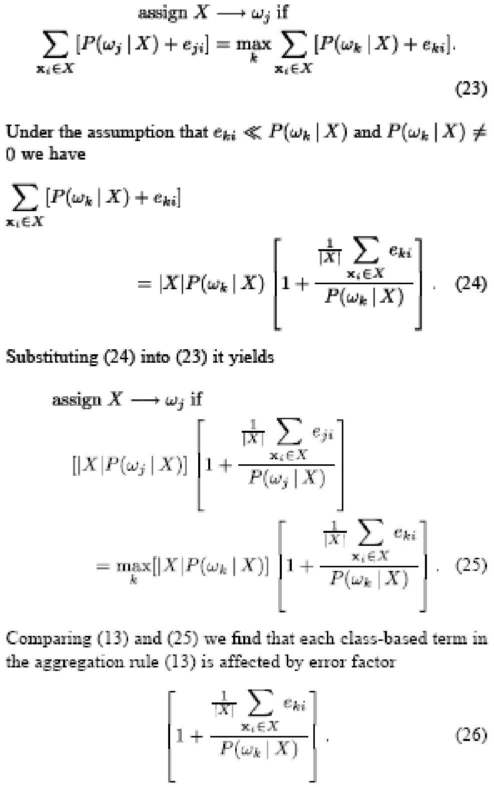
p(Wj|Xj)=P(Wj|X)+eij

These estimated probabilities, rather than the true probabilities, are used in the aggregated predictor rules.

I consider the effect of the estimation errors on the aggregationrules. Substituting (22) into (13) I have to evaluate the proposed BoF-NB scheme with different aggregation methods. The original NB classifier with feature discretization is used in the experiments as a baseline. Fig. 3 shows th classification accuracy with different training date sizes. One can find that the proposed BoF-NB scheme outperforms NB whichever aggregation method is used. On the isp dataset, the classification accuracy of BoF-NB is higher than that of NB by about 10 percentTo establish the ground truth for the testing datasets, we have developed a deep packet inspection (DPI) tool that matches regular expression signatures against flow payload content [29]. A number of application signatures are developed based on previous experience and some well-known tools such as l7-filter (http://l7-filter.sourceforge.net) and Tstat (http://tstat.tlc.polito. it). Also, several encrypted and new applications are investigated by manual inspection of the unidentified traffic.Thewide dataset consists of 182 k traffic flows which are randomly selected from the *wide* trace and carefullyrecognized by the DPI tool and flows in the wide dataset are categorized into 6application oriented classesFor the wide dataset, there are only a small dataset.

number of classes and the HTTP flows dominate the whole

dataset.



$$\text{assign } X \longrightarrow \omega_j \text{ if}$$

$$\sum_{\mathbf{x}_i \in X} [P(\omega_j \mid X) + e_{ji}] = \max_k \sum_{\mathbf{x}_i \in X} [P(\omega_k \mid X) + e_{ki}]. \tag{23}$$

Under the assumption that $e_{ki} \ll P(\omega_k \mid X)$ and $P(\omega_k \mid X) \neq 0$ we have

$$\sum_{\mathbf{x}_i \in X} [P(\omega_k \mid X) + e_{ki}] = |X| P(\omega_k \mid X) \left[1 + \frac{\frac{1}{|X|} \sum_{\mathbf{x}_i \in X} e_{ki}}{P(\omega_k \mid X)} \right]. \tag{24}$$

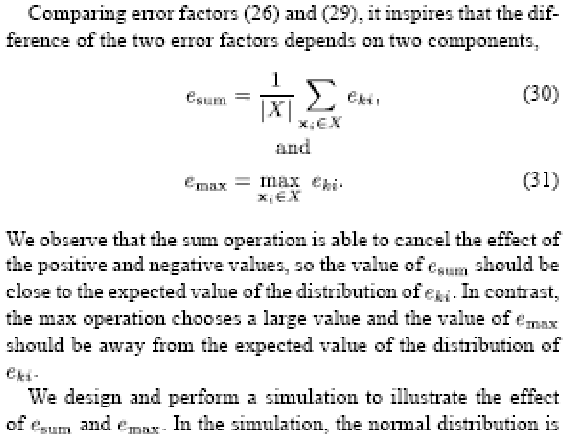
Substituting (24) into (23) it yields

$$\text{assign } X \longrightarrow \omega_j \text{ if}$$

$$[|X| P(\omega_j \mid X)] \left[1 + \frac{\frac{1}{|X|} \sum_{\mathbf{x}_i \in X} e_{ji}}{P(\omega_j \mid X)} \right] = \max_k [|X| P(\omega_k \mid X)] \left[1 + \frac{\frac{1}{|X|} \sum_{\mathbf{x}_i \in X} e_{ki}}{P(\omega_k \mid X)} \right]. \tag{25}$$

Comparing (13) and (25) we find that each class-based term in the aggregation rule (13) is affected by error factor

$$\left[1 + \frac{\frac{1}{|X|} \sum_{\mathbf{x}_i \in X} e_{ki}}{P(\omega_k \mid X)} \right]. \tag{26}$$



Comparing error factors (26) and (29), it inspires that the difference of the two error factors depends on two components,

$$e_{\text{sum}} = \frac{1}{|X|} \sum_{\mathbf{x}_i \in X} e_{ki}, \tag{30}$$

and

$$e_{\max} = \max_{\mathbf{x}_i \in X} e_{ki}. \tag{31}$$

We observe that the sum operation is able to cancel the effect of the positive and negative values, so the value of e_{sum} should be close to the expected value of the distribution of e_{ki}. In contrast, the max operation chooses a large value and the value of $e_{\max}$ should be away from the expected value of the distribution of e_{ki}.

We design and perform a simulation to illustrate the effect of e_{sum} and $e_{\max}$. In the simulation, the normal distribution is used for analyzing errors eki.This system reduce the Bof size compare before process.

CONCLUSION

Iproposed a new traffic classification Scheme whichcan effectively improve the classification performance in the situation that onlyfew training data are available. I proposed a new traffic classification schemewhichcan effectively improvetheclassification performance in the situation that only few training data are available. The proposedscheme is able to incorporate flow correlationinformationinto the classification process. I presented a theoretical analysison why and how the proposed scheme does work. A new BoF-NB and C5.0 method was also proposed to effectively aggregate the correlation naive Bayes (NB) predictions. The experiments performedon two real-world network traffic datasets demonstrated the effectiveness of the proposed scheme. The experimental results showed that BoF-NB with the sum rule outperforms existingstate-of-the-art methods by large margins. This study provides a solution to achieve high-performance traffic classification and also detect the unrelavent attribute with time-consuming training samples labelling.

REFERENCES

[1] T. T. Nguyen and G.Armitage, “A survey of technique es for internet traffic classification using machine learning,” Commun. Surveys Tuts.,vol. 10, no 4, pp. 56–76, 4th Quarter 2008.

[2] Y. Xiang, W. Zhou, and M. Guo, “Flexible deterministic packet marking: An iptraceback system to find the real source of attacks,”IEEE Trans. Parallel Distrib. Syst., vol. 20, no .567–580, Apr.2009.

[3] Snort 2011 [Online]. Available: http://www.snort.org/

[4] Bro 2011 [Online]. Available: http://bro-ids.org/index.html

[5] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K.Lee, “Internet traffic classification demystified:Myths, caveats, and the best practices,” in Proc. ACM CoNEXT Conf., New York, 2008, pp.1–12.

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC: Multilevel traffic classification in the dark,” in Proc. SIGCOMM Comput.Commun. Rev., Aug. 2005, vol. 3 pp. 229–240.

[7] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” in SIGMETRICS Perform. Eval.Rev.,Jun. 2005, vol. 33, pp. 50–60.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. NewYork: Wiley, 2001.

[9] N.Williams, S. Zander, and G. Armitage, “A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification,” in Proc.SIGCOMMComput. Commun. Rev., Oct. 2006, vol. 36, pp. 5–16.

[10] Y.-S. Lim, H.-C.Kim, J. Jeong, C.-K. Kim, T. T. Kwon, and Y. Choi,“Internet traffic classification demystified: On the sources of the discriminative power,” in Proc. 6^{t}Int New York, 2010, pp. 9:1–9:12, ACM.

[11] J. Ma, K. Levchenko, C. Kreibich, S. Savage, and G. M. Voelker, “Unexpected means of protocol inference,” in Proc. 6th ACM SIGCOMMConf. Internet Measurement, New York, 2006, pp. 313–326.

[12] S. Zander, T. Nguyen, and G. Armitage, “Automated traffic classification and application identification using machine learning,” in Proc.Ann. IEEE Conf. Local Networks, Los Alamitos, CA, 2005,pp. 250–257.

[13] J. Erman, M. Arlitt, and A. Mahanti, “Traffic classification using clustering algorithms,”inProc. SIGCOMM Workshop on Mining NetworkData, New York, 2006,

[14] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian,Traffic classification on the fly,” in Proc. SIGCOMM Comput.Commun. Rev., Apr. 2006, vol. 36

[15] Y.Wang, Y. Xiang, and S.-Z. Yu, “An automatic application signature construction system for unknown traffic,” Concurrency Computat.:Pract. Exper., vol. 22,

[16] A. Finamore, M. Mellia, and M. Meo, “Mining unclassified trafficusing automatic clustering techniques,” in Proc. TMA Int.Workshopon