Naive Bayes Based Network Traffic Classification Using Correlation Information

R.S. ANU GOWSALYA, S. MIRUNA JOE AMALI
Computer Science & Engineering, Anna University, India

Abstract — Traffic classification is of fundamental importance to numerous other network activities, from security monitoring to accounting, and from Quality of Service to providing operators with useful forecasts for long-term provisioning. Naive Bayes estimator is applied to categorize the traffic by application. Uniquely, this work capitalizes on hand-classified network data, using it as input to a supervised Naive Bayes estimator. A novel traffic classification scheme is used to improve classification performance when few training data are available. In the proposed scheme, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow (BoF). A novel parametric approach for traffic classification, which can improve the classification performance effectively by incorporating correlated information into the classification process. Then analyze the new classification approach and its performance benefit from both theoretical and empirical perspectives. Finally, a large number of experiments are carried out on large-scale real-world traffic datasets to evaluate the proposed scheme. The experimental results show that the proposed scheme can achieve much better classification performance than existing state-of-the-art traffic classification methods.

Keywords — Traffic classification, Traffic flows, Naïve Bayes, Bag-of-Flow (BoF), Correlation information, Parametric approach

I. INTRODUCTION

Network traffic classification has drawn significant attention over the past few years. Classifying traffic flows by their generation applications plays very important role in network security and management, such as quality of service (QoS) control, lawful interception and intrusion detection. Traditional traffic classification methods include the port-based prediction methods and payload-based deep inspection methods. In current network environment, the traditional methods suffer from a number of practical problems, such as dynamic ports and encrypted applications. Recent research efforts have been focused on the application of machine learning techniques to traffic classification based on flow statistical features. Machine learning can automatically search for and describe useful structural patterns in a supplied traffic dataset, which is helpful to intelligently conduct traffic classification. However, the problem of accurate classification of current network traffic based on flow statistical features has not been solved.

The flow statistical feature based traffic classification can be achieved by using supervised classification algorithms or unsupervised classification (clustering) algorithms. In unsupervised traffic classification, it is very difficult to construct an application oriented traffic classifier by using the clustering results without knowing the real traffic classes. Given a set of pre-labeled training data, the supervised traffic classification can be divided into two categories: parametric classifiers and non-parametric classifiers. Parametric classifiers, such as C4.5 decision tree, Bayesian network, SVM and neural networks, require an intensive training procedure for the classifier parameters. Non-parametric classifiers, e.g., k-Nearest Neighbor (k-NN), usually require no training phase and make classification decision based on closest training samples in the feature space.

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, which is a simple and effective probabilistic classifier employing the Bayes’ theorem with naive feature independence assumptions. 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. The main reason for the underperformance of a number of traditional classifiers including NB is the lack of the feature discretization process.

In this paper, 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 into the NB-based classification process with feature discretization.
In this paper, propose a new framework, Traffic Classification using Correlation information (TCC), to address the problem of very few training samples. The correlation information in network traffic can be used to effectively improve the classification accuracy. The major contributions of this work are summarized as follows:

- Propose a novel parametric approach which incorporates correlation of traffic flows to improve the classification performance.
- Provide a detailed analysis on the novel classification approach and its performance benefit from both theoretical and empirical aspects.
- The performance evaluation shows that the traffic classification using very few training samples can be significantly improved.

The remainder of the paper is organized as follows. Section 2 reviews related work in traffic classification. A novel classification approach is proposed in Section 3. Section 4 presents a large number of experiments and results for performance evaluation. Finally, the paper is concluded in Section 6.

II. RELATED WORK

In the area of network traffic classification, the state-of-the-art methods employ flow statistical features and machine learning techniques. Many supervised classification algorithms and unsupervised clustering algorithms have been applied to categorize internet traffic. In supervised traffic classification, the traffic classes are predefined according to real applications and a set of labelled training samples are also manually collected for classifier construction.

Este et al. applied one of the approaches to solve multi-class problems to the task of statistical traffic classification, and described a simple optimization algorithm that allows the classifier to perform correctly with as little training as a few hundred samples. Being a supervised method, it relies on two phases: during the training phase, the algorithm “acquires knowledge” about the classes by examining the training set that describes them. During the evaluation phase, a classification mechanism examines the evaluation set and associates its members to the classes that are available. They pursued the objective of recognizing the application protocol responsible for sending packets through a monitoring node. After analyzing a few packets of each TCP flow, the monitoring node’s purpose is to assign the flow to one of the application classes it has been trained with, or to the unknown protocol class.

Nguyen and Armitage proposed the importance of IP traffic classification. They briefly reviewed to important areas - IP quality of service (QoS) schemes, and lawful interception (LI). A key criterion on which to differentiate between classification techniques is predictive accuracy (i.e., how accurately the technique or model makes decisions when presented with previously unseen data). A number of metrics exist with which to express predictive accuracy. Traditional IP traffic classification relies on the inspection of a packet’s TCP or UDP port numbers (port based classification), or the reconstruction of protocol signatures in its payload (payload based classification). They list a number of possible features, and classify them into five categories: 1) Packet Level 2) Flow Level 3) Connection Level 4) Intra-flow / connection features 5) Multi-flow

Finamore et al. explained a fully unsupervised algorithm to identify classes of traffic inside an aggregate. The algorithm leverages on the K-means clustering algorithm, augmented with a mechanism to automatically determine the number of traffic clusters. The signatures used for clustering are statistical representations of the application layer protocols. They focus attention on the inspection of the unclassified traffic. They proposed an unsupervised technique that, having no knowledge of the applications that generate the traffic, partitions a traffic aggregate into “clusters” that are distinguished based on common features, i.e., they exhibit a common treat. A simple clustering methodology based on the K-means algorithm is augmented with the capability to effectively determine the number of traffic clusters.

Nguyen and Armitage proposed training the classifier on a combination of short sub-flows (extracted from full flow examples of the target application’s traffic). They demonstrated this optimization using the Naïve Bayes ML algorithm, and show that approach results in excellent performance even when classification is initiated mid-way through a flow with windows as small as 25 packets long. They suggest future use of unsupervised ML algorithms to identify optimal sub flows for training. Packet inspection uses two assumptions: third parties unaffiliated with either source or recipient can easily parse each IP packet’s payload, and the classifier knows the precise syntax of each application’s packet payloads.

Xiang et al. presented a novel and practical IP traceback system called Flexible Deterministic Packet Marking (FDPM) which provides a defense system with the ability to find out the real sources of attacking packets that traverse through the network [10]. Kim et al. critically revisited traffic classification by conducting a thorough evaluation of three classification approaches, based on transport layer ports, host behavior, and flow features [5]. Ermam et al. demonstrated how cluster analysis can be used to effectively identify groups of traffic that are similar using only transport layer statistics using two unsupervised clustering algorithms, namely K-Means and DBSCAN[1]. Karagiannis et al. analyzed the patterns at three levels of increasing detail (i) the social, (ii) the functional and (iii) the application level [4]. Lim et al. explained the three sources of the discriminative power in classifying the Internet application traffic: (i) ports, (ii) the sizes of the first one-two (for UDP flows) or four-five (for TCP flows) packets, and (iii) discretization of those features by using C4.5 algorithm [6], Williams et al. explained the performance impact of feature set reduction, using Consistency based and Correlation-based feature selection is demonstrated on Naïve Bayes, C4.5, Bayesian Network and Naïve Bayes Tree algorithms [9].
III. PROPOSED SYSTEM

A. Traffic classification approach with flow correlation

This section presents a new framework, named Traffic Classification using Correlation information or TCC for short. A novel parametric approach is also proposed to effectively incorporate flow correlation information into the classification process.

Fig. 1 shows the proposed system model. In the preprocessing, the system captures IP packets crossing a computer network and constructs traffic flows by IP header inspection. A flow consists of successive IP packets having the same 5-tuple: \{src ip, src port, dst ip, dst port, protocol\}. After that, a set of statistical features are extracted to represent each flow. Feature selection aims to select a subset of relevant features for building robust classification models. Flow correlation analysis is proposed to correlate information in the traffic flows. Finally, the robust traffic classification engine classifies traffic flows into application-based classes by taking all information of statistical features and flow correlation into account.

B. System Model

The novelty of system model is to discover correlation information in the traffic flows and incorporate it into the classification process. Conventional supervised classification methods treat the traffic flows as the individual and independent instances. They do not take the correlation among traffic flows into account. The correlation information can significantly improve the classification performance, especially when the size of training data is very small. In the proposed system model, flow correlation analysis is a new component for traffic classification which takes the role of correlation discovery. Robust classification methods can use the correlation information as input.

In this paper, “bag of flows” (BoF) is used to model the correlation information in traffic flows.

- A BoF consists of some correlated traffic flows which are generated by the same application.

A BoF can be described by

\[ Q = \{x_1, \ldots, x_n\}, \]  

where \(x_i\) is a feature vector representing the \(i\) th flow in the BoF \(Q\). The BoF \(Q\) explicitly denotes the correlation among \(n\) flows, \(\{x_1, \ldots, x_n\}\). The power of modeling correlation information with a bag has been demonstrated in preliminary work for image ranking. In this paper, the proposed flow correlation analysis will produce and analyze a large number of BoFs. A robust classification method should be able to deal with BoFs instead of individual flows.

C. Correlation Analysis

Correlation analysis is conducted using a 3-tuple heuristic, which can quickly discover BoFs in the real traffic data.

- 3-tuple heuristic: in a certain period of time, the flows sharing the same 3-tuple \{dst ip, dst port, protocol\} form a BoF.

The correlated flows sharing the same 3-tuple are generated by the same application. For example, several flows initiated by different hosts are all connecting to a same host at TCP port 80 in a short period. These flows are very likely generated by the same application such as a web browser. The 3-tuple heuristic about flow correlation has been considered in several practical traffic classification schemes.

D. Aggregation of Correlated NB Predictions

A new approach, BoF-based NB (BoF-NB) is used to aggregate correlated NB predictions in this work, which results in a more accurate aggregated predictor for traffic classification.
1) Single NB Predictor: Naive Bayes classifier is chosen due to two reasons. Firstly, it has demonstrated high classification speed and good performance using the discretized statistical features in traffic classification. Secondly, it is easy for naive Bayes classifier to produce the posterior probability that a testing flow belongs to a traffic class.

According to the Bayesian decision theory, the maximum-a-posterior classifier can minimize the average classification error. The key point is to estimate the posterior probability that a testing flow belongs to a traffic class. Given a flow \( x = \{x_1, x_2, \ldots, x_n\} \), the posterior probability corresponding to class \( \omega \) is

\[
P(\omega \mid x) = \frac{P(x_1, x_2, \ldots, x_n \mid \omega) P(\omega)}{P(x_1, x_2, \ldots, x_n)}
\]

(2)

Using Bayes’ theorem,

\[
P(\omega \mid x_1, x_2, \ldots, x_n) = \frac{P(\omega) P(x_1, x_2, \ldots, x_n \mid \omega)}{P(x_1, x_2, \ldots, x_n)}
\]

(3)

Under the naive conditional independence assumption that each feature \( x_i \) is conditionally independent of every other feature \( x_j \), (2) becomes

\[
P(\omega \mid x) = \frac{1}{C} P(\omega) \prod_{i=1}^{n} P(x_i \mid \omega)
\]

(4)

Where \( C = P(x_1, x_2, \ldots, x_n) \) is a scaling factor.

In the proposed scheme, the NB algorithm is used 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.

2) Aggregated Predictor: Under Kittler’s theoretical framework, a number of combination methods can be derived from the Bayesian decision theory which can be used for aggregated predictor. The aggregated classifier can be expressed as

\[
\phi_A (X, L) = \Theta_{\omega \in \mathcal{X}} (\phi(x, L))
\]

(5)

where \( \Theta \) is the combination method. In this paper, the equal prior assumption for all combination rules is used. Based on the previous research, the product rule and the min rule are pretty sensitive to noisy samples and weak classifiers. Therefore, the sum rule, the max rule, the median rule and the majority vote rule are used for flow aggregation and evaluate these rules in the experiments.

The aggregated classifier using the sum rule is

\[
\phi_n (X, L) = \Theta_{\omega \in \mathcal{X}} (\phi(x, L)) = \sum_{x \in \mathcal{X}} P(\omega \mid x)
\]

(6)

The aggregated classifier using the max rule is

\[
\phi_n (X, L) = \Theta_{\omega \in \mathcal{X}} (\phi(x, L)) = \max_{x \in \mathcal{X}} P(\omega \mid x)
\]

(7)

The aggregated classifier using the median rule is

\[
\phi_n (X, L) = \Theta_{\omega \in \mathcal{X}} (\phi(x, L)) = \text{med}_{x \in \mathcal{X}} P(\omega \mid x)
\]

(8)

The aggregated classifier using the majority vote rule is

\[
\phi_n (X, L) = \Theta_{\omega \in \mathcal{X}} (\phi(x, L)) = \sum_{i} \Delta_i
\]

(9)
Where $\Delta_i$ is binary valued function as

$$
\Delta_i = \begin{cases} 
1 & \text{if} P(\omega_i | x) = \max_j P(\omega_j | x_i) \\
0 & \text{otherwise}
\end{cases}
$$

(10)

IV. EXPERIMENTAL EVALUATION

In the experiments, 20 unidirectional flow statistical features are extracted and used to represent traffic flows, which are listed in Table I. Feature selection is applied to remove irrelevant and redundant features from the feature set. 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 is used to create candidate sets of features. Feature discretization can significantly improve the classification performance of many supervised classification algorithms. Feature discretization is also incorporated into the proposed scheme.

TABLE I

UNIDIRECTIONAL STATISTICAL FEATURES

<table>
<thead>
<tr>
<th>Type of features</th>
<th>Feature description</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packets</td>
<td>Number of packets transferred in unidirection</td>
<td>2</td>
</tr>
<tr>
<td>Bytes</td>
<td>Volume of bytes transferred in unidirection</td>
<td>2</td>
</tr>
<tr>
<td>Packet Size</td>
<td>Min., Max., Mean and Std Dev. of packet size in unidirection</td>
<td>8</td>
</tr>
<tr>
<td>Inter-Packet Time</td>
<td>Min., Max., Mean and Std Dev. of Inter Packet Time in unidirection</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>20</td>
</tr>
</tbody>
</table>

Two common metrics are used to measure the classification performance, 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 accuracy of a classifier on the whole testing data. F-measure is calculated by

$$
F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
$$

(11)

Where precision is the ratio of correctly classified flows over all predicted flows in a class and recall is the ratio of correctly classified flows over all ground truth flows in a class. F-Measure is used to evaluate the per-class performance.

A. Comparison With State-of-the-Art Methods

A number of experiments conducted to compare the classification performance of the proposed BoF-NB scheme with three state-of-the-art methods: C4.5, $k$-NN, and Erman’s semisupervised method. C4.5 and $k$-NN demonstrate superior traffic classification performance in recent research. Erman’s semisupervised method employs the K-means clustering algorithm and a supervised cluster-application mapping strategy. A large proportion of testing flows will be labelled as unknown by the semisupervised method when a small size of supervised training set is available. Erman’s semisupervised method is implemented with ignoring the unknown class in the training stage for fair comparison. In the experiments, the sum rule is selected for BoF-NB scheme based on the experimental results.
Fig. 2: classification accuracy of four methods (a) on isp dataset

Fig. 2 shows the classification accuracy of the four competing classification methods versus training data size. One can see that BoF-NB outperforms the other three state-of-the-art methods. For example, the classification accuracy of BoF-NB is higher than that of the second best one, the semisupervised method on the isp dataset. C4.5 and K-NN have the similar performance, which are slightly worse than the semisupervised method. The results show that BoF-NB can effectively improve the classification accuracy by aggregating correlated NB predictions.

V. CONCLUSION

In this paper, a new traffic classification scheme is proposed which can effectively improve the classification performance in the situation that only few training data are available. The proposed scheme is able to incorporate flow correlation information into the classification process. A new BoF-NB method was also proposed to effectively aggregate the correlation naive Bayes (NB) predictions. The experiments performed on real-world network traffic datasets demonstrated the effectiveness of the proposed scheme. The experimental results showed that BoF-NB with the sum rule outperforms existing state-of-the-art methods by large margins. This study provides a solution to achieve high-performance traffic classification without time-consuming training samples labelling.

REFERENCES


