International Journal of Research in Advent Technology, Vol.3, No.3, March 2015
E-ISSN: 2321-9637

Internet Traffic Classification by Aggregating Correlated NB Predictions

Pandit Memane, Atul Karche, Vaibhav Dengane
Departmentof computer engineering,TAE pune
Email: sambhajiraje01@gmail.com , karcheatul333@gmail.com

Abstract- The classification and identification of network application from network traffic flow provides various advantages to a number of fields such as security monitoring, intrusion detection and to tackle a number of network security problems including lawful interception. In this paper traffic flow is described by using the discretized statistical NB features. The flow correlation information of the network traffic flow is modeled by Flow Container (FC). In this paper novel NB classifier is proposed. First, low density graph and high density graph is analyzed. For high density flow graph Naïve Bayesian classifier is used and finally aggregated result is provided. The aggregated result is compared with machine learning algorithm such as Single NB predictor. The proposed system enhances the accuracy as well as improves the performance of the network.

Index Terms-Internet traffic classification by aggregating correlated navie bayes prediction

1. INTRODUCTION

Traffic classification is an automatic procedure which classifies computer network traffic according to various constraints into a number of traffic. Application related traffic classification is basic technology for recent network security. The traffic classification can be used to find out the worm propagation, intrusions detection, and patterns indicative of denial of service attacks(DOS attacks), and spam spread. Internet traffic is mainly occurs due to File transfer, Streaming Media, Videoconferencing, etc,. The development of network technology and application leads to severe shortage of the network resource. Identifying and control of the internet traffic flows with high efficiency is the key problem to solve. previous method of traffic classification techniques may focused on Port Based application Deduction, Packet based analysis, Payload based application deduction and the modern techniques used to classify the internet traffic is statistics based classification, Flow-based Classification.

2. RELATED WORK

Many supervision classification algorithms and unsupervision clustering algorithms have been applied to categorize network traffic. In network traffic classification, the traffic classes are defined according to real system application and a set of label training samples dataset are also manually collected for classifier construction. Este,all 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. During the *evaluation>phase*, a classification mechanism examines the evaluation set and associates its members to the classes that are available. After analyzing a few packets of each TCP flow, the monitoring node's purpose is to assign the flow to one of the application protocol class . They deap reviewed to important areas - IP quality of service schemes, and lawful interception. A key criterion on which to differentiate between classification techniques is predictive accuracy.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 , the reconstruction of protocol signatures in its payload . They list a number of possible features, and classify them into five categories*: a) Packet Level b) Flow Level c) Connection Level d) Intra-flow / connection features e) Multi-flow*

3. PROPOSED SYSTEM

3.1. Traffic classification approach with flow correlation

This section presents a framework, named **Traffic Classification** using **Correlation** information. 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 five-tuple: source IP, source port, destination IP, destination 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 algo classifies traffic flows into application-based classes by taking all information of statistical features and flow correlation into dataset.

3.2. System Model

The novelty of system model is discover correlation information in the traffic flows and incorporate it into the classification process.
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 = {$\mathbf{x}_1, \ldots, \mathbf{x}_n$},
where $\mathbf{x}_i$ is a feature vector representing the *i* th flow in the BoF Q. The BoF Q explicitly denotes the correlation among *n* flows, {$\mathbf{x}_1, \ldots, \mathbf{x}_n$}. The power of modeling correlation information with a bag has been demonstrated in preliminary work for image ranking

4. CORRELATION ANALYSIS

Correlation analysis is conducted using a tuple, which can quickly discover BoF in the real traffic data.
Three-tuple heuristic: in a certain period of time, the flows sharing the same three tuple {*destination IP, destinationt port, protocol}* form a Bag of flow

5. AGGREGATION OF CORRELATED NB PREDICTION

According to the Bayesian decision theory, the maximum-a-post order classifier can minimize the average classification error. The key point is to estimate the post order testing belongs to network traffic class. Given a flow x={$x_1, x_2 \ldots x_n$}, the posterior probability corresponding to class ω is P (ω | x) = P (ω | $x_1, x_2, \ldots, x_n$) **(2)** Using Bayes' theorem, P (ω | $x_1, x_2, \ldots, x_n$) = P(ω)p($x_1, x_2, \ldots, x_n$ | ω) p($x_1, x_2, \ldots, x_n$) **(3)** Under the naive conditional independence assumptions that each feature x_i is conditionally independent of every other feature x_j, **(2)** becomes P (ω | x) = P (ω) **(4)** here C=p($x_1, x_2, \ldots, x_n$) is a scaling factor.

6. IMPLEMENTATION ISSUES

In the experiments, 20 unidirectional flow statistical features are extracted and used to represent traffic flows, which are listed in Table. 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 BFS algorithm is used to create candidate sets of features. Feature discretization can significantly improve the classification performance of many supervision classification algo. Feature discretization is also incorporated into the proposed scheme.

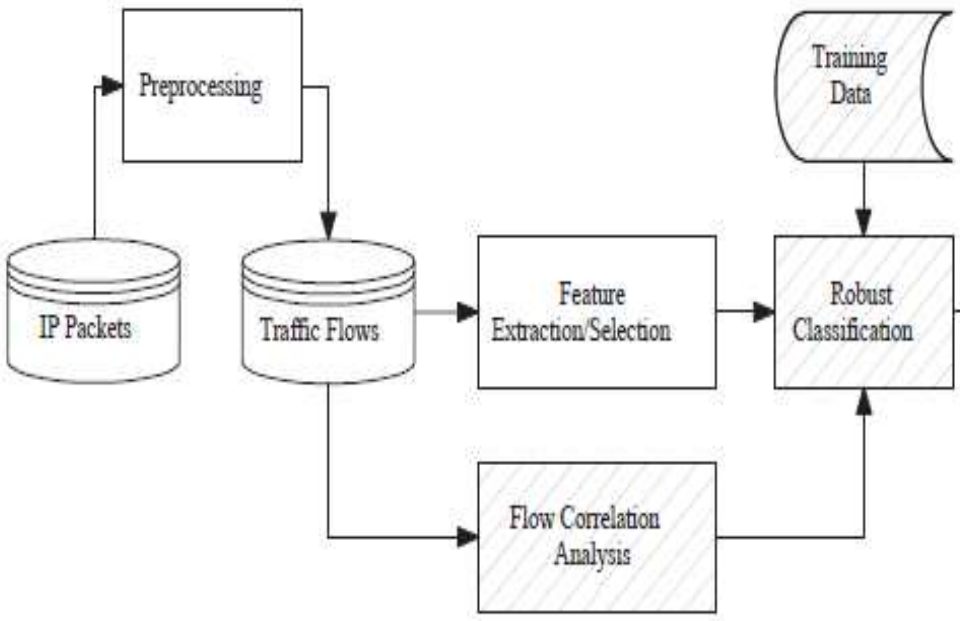

Fig. 1 shows the proposed system model

7. ***Results in images***

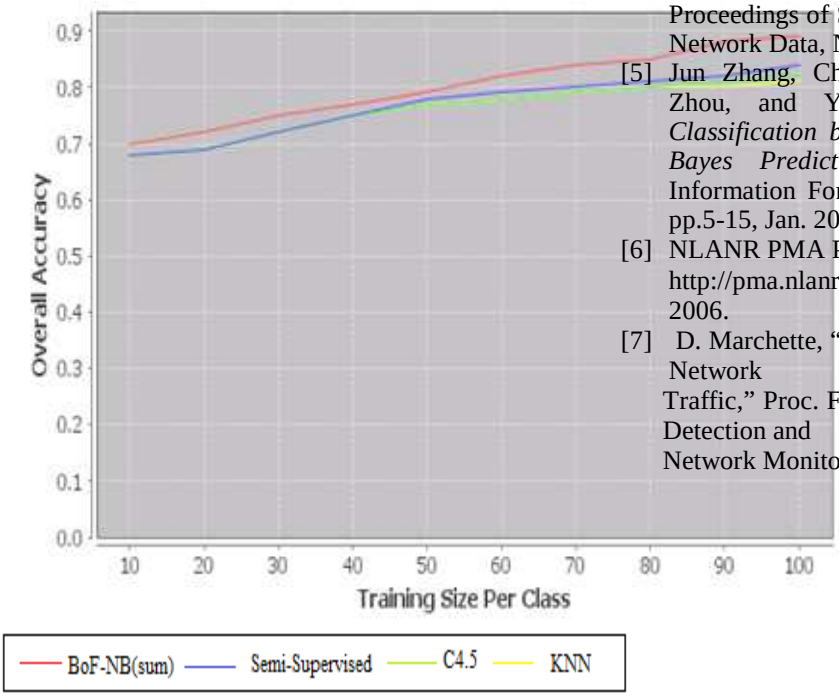

Fig.2 classification accuracy of four methods

8. 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 analysis process. A new bag of flow NB method was also proposed to effectively aggregate the correlation naive Bayes (NB) predictions.The experimental results showed that BoF-NB with the sum rule outperforms existing state-of-the-art methods by large dataset. This study provides a solution to achieve high-performance traffic classification without time-consuming training samples labels

REFERENCES

[1] N. Williams, S. Zander, and G. Armitage, "Evaluating machine learning methods for online game traffic identification," Centre for Advanced InternetArchitectures, http://caia.swin.edu.au/reports/060410C/CAIA-TR-060410C.pdf, Tech. Rep. 060410C, as of August 14, 2007.

[2] M. Crotti,M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," in *Proc. SIGCOMM Comput.Commun. Rev.*, Jan. 2007, vol. 37, pp. 5–16.

[3] MAWI Working Group Traffic Archive [Online]. Available:http://mawi.wide.ad.jp/mawi/

[4] J. Erman, M. Arlitt, and A. Mahanti, "*Traffic classification using clustering algorithms*", in Proceedings of SIGCOMM Workshop on Mining Network Data, New York,2006,pp. 281–286.

[5] Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhou, and Yong Xiang, "*Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions*", IEEE Transactions on Information Forensics and Security, vol.8, no.1, pp.5-15, Jan. 2013.

[6] NLANR PMA Packet Trace Data, http://pma.nlanr.net/Traces, 2006.

[7] D. Marchette, "A Statistical Method for Profiling Network Traffic," Proc. First USENIX Workshop Intrusion Detection and Network Monitoring, Apr. 1999