ISSN:0975-9646

G.Vivek et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1814-1818

Aggregating Correlated Naive Predictions to Detect Network Traffic Intrusion

G.Vivek[#1], B.Logesshwar[#2], Civashritt.A.B[#3], D.Ashok[#4]

UG Student, Department of Computer Science and Engineering, SRM University, Chennai, Tamil Nadu, India[#1,#2,#3]
Assistant Professor, Department Of Computer Science and Engineering, SRM University, Chennai, Tamil Nadu, India[#4]

***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. In our proposed system, we present a novel traffic classification scheme 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). We solve the BoF-based traffic classification in a classifier combination framework and theoretically analyze the performance benefit. Furthermore, a new BoF-based traffic classification method is proposed to aggregate the naive Bayes (NB) predictions of the correlated flows. We also present an analysis on prediction error sensitivity of the aggregation strategies. Finally, a large number of experiments are carried out on two 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, Intrusion Detection System, Bag of Flows, Network Traffic Characterization, Naive Bayes.**

I. INTRODUCTION

The detection of DoS attacks is essential to the protection of online services. DoS attack detection mainly concentrates on the development of network-based detection mechanisms. In proposed scheme we implemented IDNB (intrusion detection by Naive Bayes) for Big Data scheme. The main aim of this implementation is to detect intrusion packets to increase the performance of Big Data Processing model. It detects intrusion packets data in client side when large complex data is arrived.

To minimize the effort of handling large complex data we are using a specialized tool. The specialized tool used for protecting network and monitoring available service. It provides security against hackers, malicious behaviours, Denial of service attacks. NB with feature discretization significantly has higher accuracy and also improves classification speed.

NB-based traffic classifier improves classification process and decreases the size of the training samples. A Naive Bayes classifier is a probabilistic classifier based on applying Bayes' theorem with an independence assumptions. A descriptive term for the probability model will be "independent assumption model" Naive Bayes classifiers can be trained efficiently in a supervised learning setting based on the nature of the probability model. NB classifier is one which requires a small amount of training data to estimate the parameters of a classification model.

A. *Intrusion Detection System:*

An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. There are several ways to categorize an IDS:

1) Misuse detection vs. Anomaly detection:

1A. *Misuse detection:* In misuse detection, the IDS analyzes the information it gathers and compares it to large databases of attack signatures. Essentially, the IDS looks for a specific attack that has already been documented. Like a virus detection system, misuse detection software is only as good as the database of attack signatures that it uses to compare packets against.

1B. *Anomaly detection*: In anomaly detection, the system administrator defines the baseline, or normal, state of the network's traffic load, breakdown, protocol, and typical packet size. The anomaly detector monitors network segments to compare their state to the normal baseline and look for anomalies.

2) Network-based vs. Host-based systems:

2A. *Network-based:* In a network-based system, or NIDS, the individual packets flowing through a network are analyzed. The NIDS can detect malicious packets that are designed to be overlooked by a firewall's simplistic filtering rules.

2B. *Host-based:* In a host-based system, the IDS examines at the activity on each individual computer or host.

3) Passive system vs. Reactive system:

3A. *Passive System:* In a passive system, the IDS detects a potential security breach, logs the information and signals an alert.

3B. *Reactive system*: In a reactive system, the IDS responds to the suspicious activity by logging off a user or by reprogramming the firewall to block network traffic from the suspected malicious source.

II. RELATED WORKS

In the area of network traffic classification, the state-of-the-art methods employ flow statistical features and machine learning techniques [1]. Many supervised classification

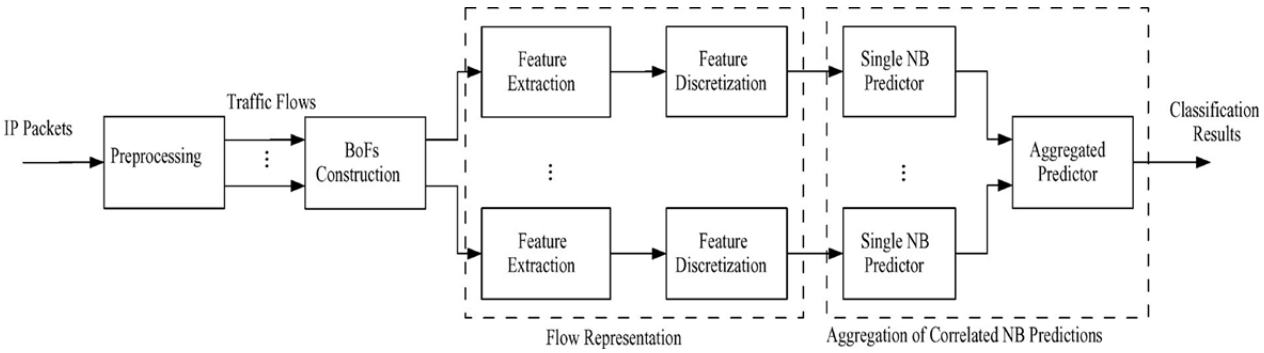
algorithms and unsupervised clustering algorithms have been applied to categorize Internet traffic. In supervised traffic classification, the traffic classes are predefined according to real applications and a set of labelled training samples are also manually collected for classifier construction. In contrast, the clustering-based methods can automatically 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 number of real applications [7]–[11]. 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 [3] applied the naive Bayes techniques to classify network traffic based on the flow statistical features. Later, several well-known algorithms were also applied to traffic classification, such as Bayesian neural networks [12] and support vector machines [13]. Erman et al. [14] proposed to use unidirectional statistical features to facilitate traffic classification in the network core. Taking into account the real-time purpose, several supervised classification methods [15], [16] were proposed, which only used the first few packets. Other existing works include the Pearson's chi-Square test based technique [17], probability density function (PDF) based protocol fingerprints [18], and small time-windows based packet count [19]. Different methods may have their own advantages in different network situations. Some empirical study [20], [4], [2], [21] evaluated the traffic classification performance of different methods for practical usage. Roughan et al. [20] have tested NN and LDA methods for traffic classification using five categories of statistical features. Williams et al. [4] 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. [2] 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 [5]. 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 [5]. Since the performance of supervised methods is sensitive to the size of training data, some proposals tried to address this problem. Erman et al. [22] 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. [6] 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. [23] tested the correctness of the 3-tuple heuristic with real-world traces. In our previous work [24], 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 SYSTEM

A: *Traffic Classification:*

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.

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 can be described by Q = {x1, . . . , xn}, where xi is a feature vector representing the i^{th} flow in the BoF Q. The BoF Q explicitly denotes the correlation among *n* flows, {x1, . . .xn}. 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 Predictors:*
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={x1, x2… xn}, the posterior probability corresponding to class ω is P (ω | x) = P (ω | x1, x2, … , xn). Using Bayes' theorem, P (ω | x1, x2, … , xn) = P(ω)p(x1, x2, … , xn | ω) p(x1, x2, … , xn) .
Under the naive conditional independence assumptions that each feature xi is conditionally independent of every other feature xj, P (ω | x) = P (ω | x1, x2, … , xn) becomes

$$\prod_{i=1}^{n} p(x_i|\omega)$$

P (ω | x) = (1/C) P (ω)
where C=p(x1, x2,…, xn) 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.

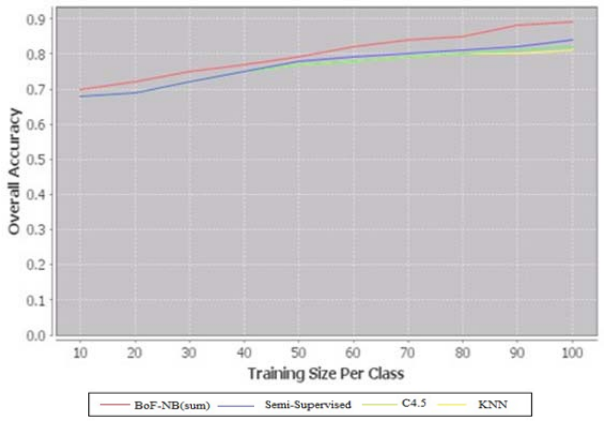
2) *Aggregated Predictor:* Under Kittler's theoretical framework, a number of combination methods can be derived from the Bayesian decision theory which can be used for aggregated predictor. The aggregated classifier can be expressed as φA (X,L) = ΘxϵX (φ(x,L)), where Θ is the combination method. In 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.

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.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.
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 semi supervised method. C4.5 and k-NN demonstrate superior traffic classification performance in recent research. Erman's semi supervised 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 semi supervised method when a small size of supervised training set is available. Erman's semi supervised 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.This 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 the second best one, the semi supervised method on the isp dataset. C4.5 and K-NN have the similar performance, which are slightly worse than the semi supervised method. The results show that BoF-NB can effectively improve the classification accuracy by aggregating correlated NB predictions.

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



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

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

[2] 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.

[3] 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.

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

[5] 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. 6th Int. Conf., Ser. Co-NEXT'10, New York, 2010, pp. 9:1–9:12, ACM.

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

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

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

[9] 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, pp. 23–26.

[10] Y.Wang, Y. Xiang, and S.-Z. Yu, "An automatic application signature construction system for unknown traffic," Concurrency Computat.: Pract. Exper., vol. 22, pp. 1927–1944, 2010.

[11] A. Finamore, M. Mellia, and M. Meo, "Mining unclassified traffic using automatic clustering techniques," in Proc. TMA Int. Workshop on Traffic Monitoring and Analysis, Vienna, Austria, Apr. 2011, pp. 150–163.

[12] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," IEEE Trans. Neural Netw., vol. 18, no. 1, pp. 223–239, Jan. 2007.

[13] A. Este, F. Gringoli, and L. Salgarelli, "Support vectormachines for tcp traffic classification," Comput. Netw., vol. 53, no. 14, pp. 2476–2490, Sep. 2009.

[14] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and discriminating between web and peer-to-peer traffic in the network core," in Proc. 16th Int. Conf. World Wide Web, New York, 2007, pp. 883–892.

[15] T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimize the use of machine learning classifiers in real-world ip networks," in Proc. Ann. IEEE Conf. Local Computer Networks, Los Alamitos, CA, 2006, pp. 369–376.

[16] L. Bernaille and R. Teixeira, "Early recognition of encrypted applications," in Proc. 8th Int. Conf. Passive and Active Network Measurement, Berlin, Heidelberg, Germany, 2007, pp. 165–175.

[17] D. Bonfiglio, M. Mellia, M. Meo, D. Rossi, and P. Tofanelli, "Revealing skype traffic: When randomness playswith you," in Proc.Conf. Applications, Technologies, Architectures, and Protocols for Computer Communications, New York, 2007, pp. 37–48.

[18] 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.

[19] S. Valenti, D. Rossi, M. Meo, M. Mellia, and P. Bermolen, "Accurate, fine-grained classification of P2P-TV applications by simply counting packets," in Proc. Int. Workshop on Traffic Monitoring and Analysis, Berlin, Heidelberg, Germany, 2009, pp. 84–92.

[20] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," in Proc. 4th ACM SIGCOMM Conf. Internet Measurement, New York, 2004, pp. 135–148.

[21] M. Pietrzyk, J.-L. Costeux, G. Urvoy-Keller, and T. En-Najjary, "Challenging statistical classification for operational usage: The

ADSL case," in Proc. 9th ACM SIGCOMM Conf. Internet Measurement Conf., New York, 2009, pp. 122–135.

[22] J. Erman, A.Mahanti, M. Arlitt, I. Cohen, and C.Williamson, "Offline/ realtime traffic classification using semi-supervised learning," Performance Evaluation, vol. 64, no. 9-12, pp. 1194–1213, Oct. 2007.

[23] M. Canini, W. Li, M. Zadnik, and A. W. Moore, "Experience with high-speed automated application-identification for network-management," in Proc. 5th ACM/IEEE Symp. Architectures for Networking and Communications Systems, New York, 2009, pp. 209–218.

[24] Y. Wang, Y. Xiang, J. Zhang, and S.-Z. Yu, "A novel semi-supervised approach for network traffic clustering," in Proc. Int. Conf. Network and System Security, Milan, Italy, Sep. 2011, pp. 169–175.