Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.168 – 174

A NEW HYBRID APPROACH FOR NETWORK TRAFFIC CLASSIFICATION USING SVM AND NAÏVE BAYES ALGORITHM

Ruchika Aggarwal, Nanhay Singh

Aryabhatta College, University of Delhi, New Delhi, India

AIACTR, New Delhi

{Ruchika.aggarwal1989, Nsingh1973}@gmail.com

Abstract- Traffic classification is an automated process which categorizes computer network traffic according to various parameters into a number of traffic classes. Many supervised classification algorithms and unsupervised clustering algorithms have been applied to categorize Internet traffic. 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. In order to improve the classification accuracy, Support Vector Machine (SVM) estimator is proposed to categorize the traffic by application. In this, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow (BoF). This methodology uses flow statistical feature based traffic classification to enhance feature discretization. This approach for traffic classification improves the classification performance effectively by incorporating correlated information into the classification process. The experimental results show that the proposed scheme can achieve much better classification performance than existing state-of-the-art traffic classification methods..

Index Terms- Traffic Classification, Naïve Bayes, Support Vector Machine (SVM), Traffic flows

I. INTRODUCTION

1.1 Introduction

Internet traffic classification is the process of identifying network applications and classifying the corresponding traffic, which is considered to be the most fundamental functionality in modern network management and security systems. OR 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. 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 data set, 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.

In this paper we illustrate the high level of accuracy achievable with the Naive Bayes estimator. We further illustrate the improved accuracy of refined variants of this estimator. Our results indicate that with the simplest of Naive Bayes estimator we are able to achieve about 65% accuracy on per-flow classification and with two powerful refinements we can improve this value to better than 95%; this is a vast improvement over traditional techniques that achieve 50--70%. While our technique uses training data, with categories derived from packet-content, all of our training and testing was done using header-derived discriminators. We emphasize this as a powerful aspect of our approach: using samples of well-known traffic to allow the categorization of traffic using commonly available information alone. The Internet continually evolves in scope and complexity, much faster than our ability to characterize, understand, control, or predict it. The field of Internet traffic classification research includes many papers representing various attempts to classify whatever traffic samples a given researcher has access to, with no systematic integration of results. Here we provide a rough taxonomy of papers, and explain some issues and challenges in traffic classification. 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.

1.2 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. SVM is a new machine learning method based on SLT (Statistics Learning Theory) and SRM (structural risk minimization). Compared with other learning machine, SVM has some unique merits, such as small sample sets, high accuracy and strong generalization performance etc. Classifiers based on machine learning use a training dataset that consists of N tuples (xi , yi) and learn a mapping f (x) → y . In the traffic classification context, examples of attributes include flow statistics such as duration and total number of packets. The terms attributes and features are used interchangeably in the machine learning literature. In our supervised Internet traffic classification system, Let X={x1, x2. . . xn} be a set of flows. A flow instance xi is

characterized by a vector of attribute values, xi= { xij |1≤ j ≤ m }, where m is the number of attributes, and xij is the value of the j-th attribute of the i-th flow, and xi is referred to as a feature vector. Also, let Y= {y1, y2. . . yq} be the set of traffic classes, where q is the number of classes of interest. To build a robust classifier, three factors to be considered.

(i) A set of discriminating features such as protocols, ports, IP address.

(ii) An effective classification algorithm; the SVM is chosen, which consistently outperformed all others.

(iii) A correct and complete training set for building the classifier model. Support Vector Machine (SVM), based on statistical learning theory, is known as one of the best machine learning algorithms for classification purpose and has been successfully applied to many classification problems such as image recognition, text categorization, medical diagnosis, remote sensing, and motion classification. SVM method is selected as classification algorithm due to its ability for simultaneously minimizing the empirical classification error and maximizing the geometric margin classification space. These properties reduce the structural risk of over-learning with limited samples.

1.4 Naive Bayes

One of the recent approaches classifies the traffic by using the simple and effective probabilistic Naive Bayes (NB) classifier. It employs the Bayes' theorem with naive feature independence assumptions. The main reason for the underperformance of a number of traditional classifiers including NB is the lack of the feature discretization process. 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

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

1.5 Supervised Methods

The supervised traffic classification methods analyze the supervised training data and produce an inferred function which can predict the output class for any testing flow. In supervised traffic classification, sufficient supervised training data is a general assumption. To address the problems suffered by payload-based traffic classification, such as encrypted applications and user data privacy, Moore and Zuev applied the supervised naive Bayes techniques to classify network traffic based on

flow statistical features. Williams et al. evaluated the supervised algorithms including naive Bayes with discretization, naive Bayes with kernel density estimation, C4.5 decision tree, Bayesian network, and naive Bayes tree. Nguyen and Armitage proposed to conduct traffic classification based on the recent packets of a flow for real-time purpose. Auld et al. extended the work of with the application of Bayesian neural networks for accurate traffic classification.

1.6 Unsupervised Methods

The unsupervised methods (or clustering) try to find cluster structure in unlabeled traffic data and assign any testing flow to the application-based class of its nearest cluster. McGregor et al. proposed to group traffic flows into a small number of clusters using the expectation maximization (EM) algorithm and manually label each cluster to an application. Zander et al. used AutoClass to group traffic flows and proposed a metric called intraclass homogeneity for cluster evaluation. Bernaille et al. applied the k-means algorithm to traffic clustering and labeled the clusters to applications using a payload analysis tool. Erman et al. evaluated the k-means, DBSCAN and AutoClass algorithms for traffic clustering on two empirical data traces. The empirical research showed that traffic clustering can produce high-purity clusters when the number of clusters is set as much larger than the number of real applications. Generally, the clustering techniques can be used to discover traffic from previously unknown applications . Wang et al. proposed to integrate statistical feature-based flow clustering with payload signature matching method, so as to eliminate the requirement of supervised training data. Finamore et al. combined flow statistical feature-based clustering and payload statistical feature-based clustering for mining unidentified traffic. However, the clustering methods suffer from a problem of mapping from a large number of clusters to real applications.

II. RELATED WORK

R.S.Anu Gowsalya, Dr. S.Miruna Joe Amali [01], In this paper they explain, Traffic classification is an automated process which categorizes computer network traffic according to various parameters into a number of traffic classes. Many supervised classification algorithms and unsupervised clustering algorithms have been applied to categorize Internet traffic. 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. In order to improve the classification accuracy, Support Vector Machine (SVM) estimator is proposed to categorize the traffic by application. In this, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow (BoF). This methodology uses flow statistical feature based traffic classification to enhance feature discretization. This approach for traffic classification improves the classification performance effectively by incorporating correlated information into the classification process. The experimental results show that the proposed scheme can achieve much better classification performance than existing state-of-the-art traffic classification methods.

Kuldeep Singh, Manoj Kumar [02], In this paper they explain, traffic classification has wide applications in network management, from security monitoring to quality of service measurements. Recent research tends to apply machine learning techniques to flow statistical feature based classification methods. The nearest neighbor (NN)-based method has exhibited superior classification performance. It also has several important advantages, such as no requirements of training procedure,

no risk of overfitting of parameters, and naturally being able to handle a huge number of classes. However, the performance of NN classifier can be severely affected if the size of training data is small. In this paper, we propose a novel nonparametric approach for traffic classification, which can improve the classification performance effectively by incorporating correlated information into the classification process. We analyze the new classification approach and its performance benefit from both theoretical and empirical perspectives. A large number of experiments are carried out on two real-world traffic data sets to validate the proposed approach. The results show the traffic classification performance can be improved significantly even under the extreme difficult circumstance of very few training samples.

R.S. ANU GOWSALYA, S. MIRUNA JOE AMALI [03], In this paper they explain, 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.

Ms. Zeba Atique Shaikh, Prof. Dr. D.G. Harkut [04], In this paper they explain, Network traffic classification can be used to identify different applications and protocols that exist in a network. Actions such as monitoring, discovery, control and optimization can be performed by using classified network traffic. The overall goal of network traffic classification is improving the network performance. Once the packets are classified as belonging to a particular application, they are marked. These markings or flags help the router determine appropriate service policies to be applied for those flows. This paper gives an overview of available network classification methods and techniques. Researchers can utilize this paper for approaching real time network traffic classification. Traffic classification using payload, statistical analysis, deep packet inspection, naïve Bayesian estimator and Bayesian neural networks are reviewed in this paper.

Pallavi Singhal Rajeev Mathur [05], In this Paper they explain, Network traffic classification is extensively required mainly for many network management tasks such as flow prioritization, traffic shaping/policing, and diagnostic monitoring. Similar to network management tasks, many network engineering problems such as workload characterization and modeling, capacity planning, and route provisioning also benefit from accurate identification of network traffic .This paper presents review on all the work done related to Network Traffic Management since 1993 to 2013 in various fields like artificial intelligence, neural network, ATM and wireless networks.

III. PROPOSED METHODOLOGY

. The problems suffered by payload-based traffic classification, such as encrypted applications and user data privacy, Moore and applied the supervised naive techniques to classify network traffic based on flow statistical features. Evaluated the supervised algorithms including naive Bayes with discretization, naive Bayes with kernel density estimation, C4.5 decision tree, Bayesian network, and naive Bayes tree. Nguyen and Armitage proposed to conduct traffic classification based on the recent packets of a flow for real-time purpose. Extended the work of with the application of Bayesian neural networks for accurate traffic classification. used unidirectional statistical features for traffic classification in the network core and proposed an algorithm with the capability of estimating the missing features. Proposed to use only the size of the first packets of an SSL connection to recognize the encrypted applications proposed to analyze the message content randomness introduced by the encryption processing using Pearson's chi-Square test-based technique. The probability density function (PDF)-based protocol fingerprints to express three traffic statistical properties in a compact way. Their work is extended with a parametric optimization procedure.

Advantages

- These works use parametric machine learning algorithms, which require an intensive training procedure for the classifier parameters and need the retraining for new discovered applications.
- Evaluated three supervised methods for an ADSL provider managing many points of presence, the results of which are comparable to deep inspection solutions.
- Applied oneclass SVMs to traffic classification and presented a simple optimization algorithm for each set of SVM working parameters proposed to classify P2P-TV traffic using the count of packets exchanged with other peers during the small time windows.

IV. RESULT ANALYSIS / IMPLEMENTATION

Table I shows classification accuracy and training time of five ML classifiers namely MLP, RBF, C4.5, Bayes Net and Naïve Bayes for Dataset 1 which has been developed by considering packet capture duration of 2 seconds only. It is clear from this table and figure 5 that maximum classification accuracy is provided by Bayes Net classifier for Dataset 1 which is 88.125 % with training time or model building time of 0.7 seconds only.

From table I, it is also clear that MLP algorithm gives very poor performance in terms of classification accuracy and training time. Furthermore, classification accuracy is of RBF Neural Network Classifier is also lesser than that of other ML classifiers and its training time is very large as compared to Bayes Net, C4.5 and Naïve Bayes which make it inappropriate for efficient IP traffic classification. Therefore MLP and RBF algorithms are not taken into consideration for further discussion.

ML Classifiers	**MLP**	**Bayes Net**	**Naïve Bayes + SVM**
Classification Accuracy (%)	27.75	88.125	88.875
Training Time (Seconds)	17.79	0.7	0.16

Table 1 – Accuracy % & Time taken to train system of Proposed and Existing methodology

V. CONCLUSION

In this paper, firstly real time internet traffic has been captured using Wireshark software for packet capture durations of 2 seconds. After that, Internet traffic from this dataset is classified using five ML classifiers. Results show that Bayes Net Classifier gives better performance with classification accuracy of 88.125%. But the problem with this technique is large training time which makes it ineffective of real time and online IP traffic classification. Solution of this problem is reduction in number of features characterizing each internet application sample. For this Correlation based FS algorithm is better choice with which a reduced feature dataset has been developed. Using this new dataset, performance of five ML classifiers has been analyzed. Results show that Bayes Net classifier gives better performance among all other classifiers in terms classification accuracy of 91.875 %, training time of ML algorithms and recall and precision values of individual internet applications. Thus it is evident that Bayes Net is an effective ML techniques for near real time and online IP traffic classification with reduction in packet capturing time and reduction in number of features characterizing application samples with Correlation based FS algorithm.

In this research work, the packet capturing duration is reduced to 2 seconds to make this approach suitable for implementing real time IP traffic classification. For this purpose, the packet capturing duration should be as less as possible. This can be further reduced to fraction of seconds which will make this classification technique more real time compatible. Secondly, this internet traffic dataset can be extended for many other internet applications which internet users use in their day to day life and it can also be captured from various different real time environments such as university or college campus, offices, home environments and other work stations etc.

REFERENCES

[1] R.S.Anu Gowsalya, Dr. S.Miruna Joe Amali, **"SVM Based Network Traffic Classification Using Correlation Information",** International Journal of Research in Electronics and Communication Technology (IJRECT 2014), ISSN : 2348 - 9065 (Online) ISSN : 2349 – 3143.

[2] Kuldeep Singh, Manoj Kumar, **"Review on Network Traffic Classification",** International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064.

[3] R.S. ANU GOWSALYA, S. MIRUNA JOE AMALI, **"Naive Bayes Based Network Traffic Classification Using Correlation Information",** International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 3, March 2014 ISSN: 2277 128X.

[4] Ms. Zeba Atique Shaikh, Prof. Dr. D.G. Harkut, **"An Overview of Network Traffic Classification Methods",** International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169 Volume: 3 Issue: 2.

[5] Pallavi Singhal Rajeev Mathur, Ph.D. Himani Vyas, **"State of the Art Review of Network Traffic Classification based on Machine Learning Approach",** International Journal of Computer Applications (0975 – 8887) International Conference on Recent Trends in engineering & Technology 2013.

