IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 6, June 2016

A Survey on SVM and Naives Bayes Network Traffic Classification Using Correlation Information

Dipti Tiwari[1], Bhawna Mallick [2]

Dept. of Computer Science, Galgotia College of Engineering and Technology, Greater Noida, India [1]

Professor, Dept. of Computer Science and Information Technology, Galgotia College of Engineering and Technology, Greater Noida, India [2]

Abstract: Traffic classification is an automatic method that categorizes electronic network traffic per varied parameters into variety of traffic categories. Many supervised classification algorithms and unsupervised clump algorithms have been applied to reason web traffic. Traditional traffic classification strategies embody the port-based prediction strategies and payload-based deep scrutiny strategies. In current network environment, the traditional strategies suffer from variety of sensible issues, such as dynamic ports and encrypted applications. In order to enhance the classification accuracy, Support Vector Machine (SVM) and Naïve Bayes estimator is projected to reason the traffic by application. In this, traffic flows are represented mistreatment the discretized applied math options and flow correlation data is modelled by bag-of-flow (BoF). This methodology uses flow statistical feature based mostly traffic classification to boost feature discretization. This approach for traffic classification improves the classification performance effectively by incorporating correlated data into the classification method. The experimental results show that the proposed theme will come through far better classification performance than existing progressive traffic classification strategies.

Keywords- Support Vector Machine (SVM), Traffic Classification, Supervised algorithm, Naïve Bayes

I. Introduction

Internet traffic classification is the method of distinctive network applications and classifying the corresponding traffic, which is thought-about to be the foremost basic practicality in fashionable network management and security systems. OR Traffic classification is an automatic procedure that classifies laptop network traffic in step with varied constraints into variety of traffic. Application related traffic classification is basic technology for recent network security. The traffic classification can be wont to determine the worm propagation, intrusions detection, and patterns indicative of denial of service attacks (DOS attacks), and spam spread. Traditional traffic classification ways embody the port-based prediction ways and payload-based deep scrutiny ways. In current network environment, the traditional ways suffer from variety of sensible issues, such as dynamic ports and encrypted applications. Recent research efforts have been centred on the appliance of machine learning techniques to traffic classification supported flow applied mathematics options. Machine learning can mechanically search for and describe helpful structural patterns in a very provided traffic knowledge set, which is useful to showing intelligence conduct traffic classification. However, the problem of correct classification of current network traffic supported flow applied mathematics options has not been resolved.

In this paper we illustrate the high level of accuracy possible with the Naive Bayes computer. We any illustrate the improved accuracy of refined variants of this computer. Our results indicate that with the simplest of Naive Bayes computer we have a tendency to able to come through regarding sixty fifth accuracy on per-flow classification and with 2 powerful refinements we will improve this price to raised than 95%; this can be a colossal improvement over ancient techniques that come through 50--70%. While our technique uses training knowledge, with categories derived from packet-content, all of our training and testing was done mistreatment header-derived discriminators. We emphasize this as a powerful side of our approach: mistreatment samples of well-known traffic to permit the categorization of traffic mistreatment ordinarily out there data alone. The Internet regularly evolves in scope and complexness, much quicker than our ability to characterize, understand, control, or predict it. The field of Internet traffic classification analysis includes several papers representing varied makes an attempt to classify no matter traffic samples a given investigator has access to, with no systematic integration of results. Here we give a rough taxonomy of papers, and explain some problems and challenges in traffic classification. The flow statistical feature-based traffic classification will be achieved by mistreatment supervised classification algorithms or unattended classification (clustering) algorithms. In unsupervised traffic classification, it is very tough to construct And application homeward traffic classifier by mistreatment the bunch results while not knowing the important traffic categories .In the last decade, considerable analysis works were

DOI 10.17148/IARJSET.2016.3647

rumoured on the application of machine learning techniques to traffic classification. These works can be classified as supervised ways or unattended ways.

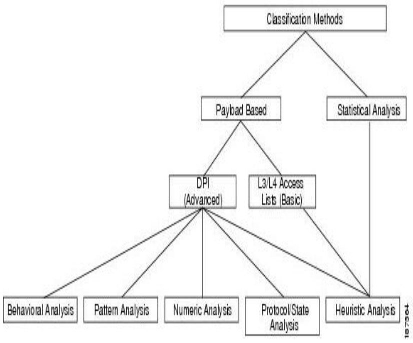

Fig1: Network traffic classification methods and techniques.

Supervised Methods

The supervised traffic categorisation strategies Analyze the supervised training information and turn out an inferred operate that will predict the output class for any testing flow. In supervised traffic classification, sufficient supervised training information is a general assumption. To address the issues suffered by payload-based traffic classification, such as encrypted applications and user data privacy. Although traffic classification by looking application signatures in payload content is a lot of correct, deriving the signatures manually is terribly time intense. To address this problem, researchers proposed to apply the supervised learning algorithms to mechanically establish signatures for a variety of applications. Additionally they planned application signatures exploitation applied mathematics characterization of payload and applied supervised algorithms, such as SVM, to conduct traffic classification. Similar to the supervised strategies supported flow applied mathematics options, these payload-based methods need decent supervised training information.

Unsupervised Methods

The unsupervised strategies (or clustering) strive to realize cluster structure in unlabelled traffic information and assign any testing flow to the application-based category of its nearest cluster. McGregor et al. proposed to cluster traffic flows into a tiny variety of clusters exploitation the expectation maximization (EM) algorithmic program and manually label every cluster to an application typically, the clustering techniques will be wont to discover traffic from antecedently unknown applications but, the mapping method can turn out a giant proportion of “unknown” clusters, especially once the supervised training information is terribly tiny. In this paper, we study the drawback of supervised traffic classification exploitation only a few training samples. From the supervised learning purpose of read, several supervised samples area unit out there for every category. Without the method of unattended clump, the mapping between clusters and applications can be avoided. Our work focuses on nonparametric classification strategies and address the troublesome drawback of traffic classification exploitation terribly few training samples. The motivations are twofold. First, as mentioned in Section one, nonparametric NN methodology has three vital blessings that area unit appropriate for traffic classification in current complicated network scenario. Second, labelling training information is time intense and the capability of classification exploitation only a few training sample is extremely helpful.

II. RELATED WORKS

SVM Based Network Traffic Classification Using Correlation Information, 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.

Network Traffic Classification Using Correlation Information, During this paper they explain, traffic classification has wide applications in network management, from security observance to quality of service measurements. Recent analysis tends to use machine learning techniques to flow statistical feature primarily based classification ways. The closest neighbor (NN)-based methodology has exhibited superior classification performance. It conjointly has many vital benefits, like no needs of training procedure, no risk of over fitting of parameters, and naturally having the ability to handle a large range of categories. However, the performances of NN classifier are often severely affected if the scale of training data is small. During this paper, we tend to propose a unique nonparametric approach for traffic classification, which might improve the classification performance effectively by incorporating related info into the classification method. We tend to analyze the new classification approach and its performance has the benefit of each theoretical and empirical perspectives. An oversized range of experiments are distributed on two real-world traffic data sets to validate the proposed approach. The results show the traffic classification performance are often improved considerably even beneath the extreme tough circumstance of only a few training samples.

Naive Bayes Based Network Traffic Classification Using Correlation Information during this paper they justify, Traffic classification is of basic importance to various alternative network activities, from security monitoring to accounting, and from Quality of Service to providing operators with helpful forecasts for long-run provisioning. Naive Bayes estimator is applied to categorise the traffic by application. Uniquely, this work capitalizes on hand-classified network information, victimization it as input to a supervised Naive Bayes estimator. a unique traffic classification theme is employed to boost classification performance once few coaching information are accessible. Within the planned theme, traffic flows are described using the discretized applied math options and flow correlation data is modeled by bag-of-flow (BoF). a unique parametric approach for traffic classification, which might improve the classification performance effectively by incorporating related to data into the classification method. Then analyze the new classification approach and its performance enjoys each theoretical and empirical views. Finally, an oversized variety of experiments are applied on large-scale real-world traffic datasets to judge the projected theme. The experimental results show that the planned theme are able to do far better classification performance than existing state-of-the-art traffic classification ways.

An Overview of Network Traffic Classification Methods, In this paper they explain, Network traffic classification may be accustomed identify totally different applications and protocols that exist in a very network. Actions like obseving, discovery, control and optimization may be performed by using classified network traffic. the general goal of network traffic classification is rising up the network performance. Once the packets are classified as belonging to a selected application, they're marked. These markings or flags facilitate the router verify acceptable service policies to be applied for those flows. This paper provides an outline of obtainable network classification strategies and techniques. Researchers will utilize this paper for approaching real time network traffic classification. Traffic classification using payload, statistical analysis, deep packet review, naïve theorem estimator and bayesian neural networks are reviewed during this paper.

State of the Art Review of Network Traffic Classification based on Machine Learning Approach, In this Paper they explain, Network traffic classification is extensively required primarly 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. PROBLEM DEFINITION

Statistics based classification: Packet level trace generates n number of zero payload flows wherever peer attempt to connect one another. In this case some statistical feature of the packet-level-trace is grabbed and accustomed classify the network traffic. This approach is feasible to work out the appliance kind, but specific application/client cannot be determined generally. These flow characteristics can be extremely coded manually or in a different way is to mechanically extract the options of a specific quite traffic. This technique is achieved by combining applied math method with AI. There is various data processing approaches combination to use applied math based mostly classification. Applying statistical based mostly classification can offer high accuracy for traffic classification, but the result cannot be actual and settle for minor classification errors.

Flow-based Classification: Traffic application based on flow-level information with a similar and high level of accuracy is incredibly tough, because it consist of less elaborate input. For application behaviour, analyzing the application constraints makes the classification more possible. The connection patterns is the novel approach to classify traffic supported the applying teams, It is represented by graphs, where nodes provides scientific discipline address and port pairs data and edge represents flows between supply and destination nodes. Connection patterns square measure analyzed at 3 levels of details, the social, the functional and the application level. This method operates among the data having no access to payload data, no knowledge concerning port range and no data behind what current flow collectors offer. On the other hand, connection patterns need a high quality of flow data and finished flow amount to perform the analyses.
The proposed work of this paper uses the SVM and NB based traffic .

IV PROPOSED METHODOLOGY

The problems suffered by payload-based traffic classification, like encrypted applications and user information privacy, Moore and applied the supervised naive techniques to classify network traffic supported flow statistical options. Evaluated the supervised algorithms together with naive Bayes with discretization, naive Bayes with kernel density estimation, C4.5 decision tree, Bayesian network, and naive Bayes tree. Nguyen and Armitage planned to conduct traffic classification supported the recent packets of a flow for period of time purpose. Extended the work of with the appliance of Bayesian neural networks for correct traffic classification. Used one-way statistical options for traffic classification within the network core and projected an formula with the

DOI 10.17148/IARJSET.2016.3647

potential of estimating the missing options. Proposed to use solely the dimensions of the primary packets of an SSL affiliation to acknowledge the encrypted applications projected to investigate the message content randomness introduced by the encoding process using Pearson's chi-Square test-based technique. The probability density perform (PDF)-based protocol fingerprints to specific three traffic statistical properties during a compact approach. Their work is extended with a constant optimization procedure.

Advantages

These works use constant machine learning algorithms, that need an intensive coaching procedure for the classifier parameters and want the training for brand new discovered applications.

➢ Evaluated three supervised strategies for an ADSL provider managing several points of presence, the results of that are corresponding to deep review solutions.

➢ Applied oneclass SVMs to traffic classification and conferred an easy optimisation formula for every set of SVM operating parameters projected to classify P2P-TV traffic using the count of packets changed with different peers throughout the little time windows.

Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training information (supervised learning), the algorithm outputs associate best hyperplane that categorizes new examples. SVM is a new machine learning method supported SLT (Statistics Learning Theory) and SRM (structural risk minimization). Compared with other learning machine, SVM has some unique deserves, 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 like period and total variety of packets. The terms attributes and features are used interchangeably in the machine learning literature. In our supervised web traffic classification system, Let X= be a set of flows. A flow instance xi is characterized by a vector of attribute values, xi= 1≤ j ≤ m , where m is the variety 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= be the set of traffic classes, where letter of the alphabet is the variety of categories of interest. To build a strong classifier, three factors to be thought of. (i) A set of discriminating features like protocols, ports, IP address. (ii) An effective classification algorithm; the SVM is chosen, which systematically outperformed all others. (iii) A correct and complete training set for building the classifier model. Support Vector Machine (SVM), based on applied mathematics learning theory, is known jointly of the most effective machine learning algorithms for classification purpose and has been with success applied to several classification issues like image recognition, text categorization, medical diagnosis, remote sensing, and motion classification. SVM method is elect as classification formula owing to its ability for at the same time minimizing the empirical classification error and maximising the geometric margin classification area. These properties reduce the structural risk of over-learning with restricted samples.

V. NAVIE BAYES

One of the recent approaches classifies the traffic by using the easy and effective probabilistic Naive Thomas Bayes (NB) classifier. It employs the Bayes theorem with naive feature independence assumptions. The main reason for the underperformance of variety of traditional classifiers together with NB is that the lack of the feature discretization method. NB algorithm is used to provide a group of posterior possibilities as predictions for every testing flow. It is different to the standard NB classifier that directly assigns a testing flow to a category with the utmost posterior chance. Considering correlated flows, the predictions of multiple flows will be collective to create a final prediction

VI. RESULT ANALYSIS:

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 1, 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**	**Naive Bayes**
Classification Accuracy (%)	27.75	88.125	88.875
Training Time (Seconds)	17.79	0.7	0.16

VII. 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.

DOI 10.17148/IARJSET.2016.3647

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.

REFRENCES

[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

[6]. G.Suganya.M.sc.,B.Ed, **“**An Efficient Network Traffic Classification Based on Unknown and Anomaly Flow Detection Mechanism**”,** International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 4 – Apr 2014, ISSN: 2231-2803, http://www.ijcttjournal.org

[7]. Andrew W. Moore, Denis Zuev, “Internet Traffic Classification Using Bayesian Analysis Techniques”, SIGMETRICS 05, June 6– 10, 2005, Banff, Alberta, Canada. Copyright 2005 ACM 1-59593-022-1/05/0006.

[8]. M. Tamilkili, “A Survey on Recent Traffic Classification Techniques Using Machine Learning Methods”, International Journal of Advanced Research in Computer Science and Software Engineering, Research Paper Available online at: www.ijarcsse.com, Volume 3, Issue 12, December 2013, ISSN: 2277 128X.

[9]. G. Rubadevi , R. Amsaveni, “A Novel Hybrid Aggregated Classifier for Internet Traffic Classification**”,** International Journal of Computer Engineering and Applications, Volume VII, Issue II, Part II, August 14.

[10]. Bin Hu, Yi Shen, **“**Machine Learning Based Network Traffic Classification : A Survey**”**, Journal of Information & Computational Science 9: 11 (2012) 3161–3170 Available at http://www.joics.com.

[11]. G. Vivek, B. Logesshwar, Civashritt. A. B, D. Ashok, **“**Aggregating Correlated Naive Predictions to Detect Network Traffic Intrusion**”**, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015, 1814-1818, ISSN:0975-9646.

[12]. Pandit Memane, Atul Karche, Vaibhav Dengane, “Internet Traffic Classification by Aggregating Correlated NB Predictions”, International Journal of Research in Advent Technology, Vol.3, No.3, March 2015 E-ISSN: 2321-9637.