SVM and Naïve Bayes Network Traffic Classification using Correlation Information

Dipti Tiwari
Dept. of computer science,
Galgotias College of
Engineering and Technology,
Greater Noida

Bhawna Mallick
(Head of department)
Dept. of Computer Science and
Information Technology,
Galgotias College of
Engineering and Technology,
Greater Noida

ABSTRACT

Traffic classification is an automatic method that categorizes network traffic in line with varied parameters into variety of traffic categories. Many supervised classification algorithms and unsupervised clustering algorithms have been applied to categorise web traffic. Traditional traffic classification strategies embrace the port-based prediction strategies and payload-based deep examination strategies. In current network environment, the traditional strategies suffer from variety of sensible issues, such as dynamic ports and encrypted applications. In order to boost the classification accuracy, Support Vector Machine (SVM) and Naïve Bayes estimator is planned to categorise the traffic by application. In this, traffic flows are represented exploitation the discretized statistical options and flow correlation data is sculptured by bag-of-flow (BoF). This methodology uses flow statistical feature primarily based 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.

1. INTRODUCTION

Internet traffic classification is the method of distinctive network applications and classifying the corresponding traffic, which is thought of to be the foremost basic practicality in trendy network management and security systems. OR Traffic classification is an automatic procedure that classifies computer network traffic consistent with varied constraints into variety of traffic. Application related traffic classification is basic technology for recent network security. The traffic classification can be accustomed to resolve the worm propagation, intrusions detection, and patterns indicative of denial of service attacks (DOS attacks), and spam spread methods. In current network environment, the traditional strategies suffer from variety of sensible issues, such as dynamic ports and encrypted applications. Recent research efforts have been centered on the applying of machine learning techniques to traffic classification supported flow applied math options. Machine learning can mechanically search for and describe helpful structural patterns in a very equipped 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 math options has not been solved .

In this paper we illustrate the high level of accuracy possible with the Naive bayes estimator. We more illustrate the improved accuracy of refined variants of this calculator. Our results indicate that with the simplest of Naive Bayes estimator we tend to are ready to achieve regarding 65th accuracy on per-flow classification and with two powerful refinements we will improve this price to raised than 95%; this is often an enormous improvement over ancient techniques that achieve 50--70%. While our technique uses training knowledge, with categories derived from packet-content, all of our training and testing was done victimisation header-derived discriminators. We emphasize this as a powerful facet of our approach: victimisation samples of well-known traffic to permit the categorization of traffic victimisation normally offered info alone. The Internet regularly evolves in scope and quality, much quicker than our ability to characterize, understand, control, or predict it. The field of Internet traffic classification analysis includes several papers representing varied tries 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 victimisation supervised classification algorithms or unsupervised classification (clustering) algorithms. In unsupervised traffic classification, it is very tough to construct an application orienting traffic classifier by victimisation the cluster results while not knowing the important traffic categories.

1.1 Support Vector Machine (SVM)

A Support Vector Machine (SVM) may be a discriminative classifier formally outlined by a separating hyperplane. In different words, given labeled training knowledge (supervised learning), the rule outputs an optimum hyperplane that categorizes new examples. SVM may be a new machine learning technique supported SLT (Statistics Learning Theory) and SRM (structural risk minimization). Compared with different learning machine, SVM has some distinctive deserves, like tiny sample sets, high accuracy and powerful generalization performance etc. Classifiers supported machine learning use a coaching dataset that consists of N tuples (xi , Yi) and learn a mapping f (x) → y . within the traffic classification context, samples of attributes embrace flow statistics like length and total variety of packets. The terms attributes and options square measure used interchangeably within the machine learning literature. In our supervised net traffic organization, Let X= be a group of flows. A flow instance xi is characterised by a vector of attribute values, xi= 1≤ j ≤ m , wherever m is that the variety of attributes, and xij is that the price of the j-th attribute of the i-th flow, and xi is named as a feature vector. Also, let Y= be the set of traffic categories, wherever alphabetic character is that the variety of categories of interest. to make a strong classifier, three factors to be thought-about.

1. a group of discriminating options like protocols, ports, IP address.
2. an efficient classification algorithm; the SVM is chosen, that systematically outperformed all others.
3. an accurate and complete coaching set for building the classifier model. Support Vector Machine (SVM), supported statistical learning theory, is understood united 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, diagnosis, remote sensing, and motion classification. SVM technique is chosen as classification rule as a result of its ability for at the same time minimizing the empirical classification error and increasing the geometric margin classification house. These properties cut back the structural risk of over-learning with minimum samples.

1.2 Naive Bayes

One of the recent approaches classifies the traffic by mistreatment the easy and effective probabilistic Naive bayes (NB) classifier. It employs the Bayes'theorem with naive feature independence assumptions. main reason for the underperformance of variety of ancient classifiers as well as NB is that the lack of the feature discretization method. NB algorithmic program is employed to supply a collection of posterior chances as predictions for every testing flow. it's totally different to the traditional NB classifier that directly assigns a testing flow to a category with the most posterior chance. Considering correlative flows, the predictions of multiple flows are aggregated to create a final prediction.

Naive bayes has been studied extensively since the 1950s. it had been introduced below a special name into the text retrieval community within the early 1960s, and remains a preferred (baseline) technique for text categorization, the matter of judgment documents as happiness to at least one class or the opposite (such as spam or legitimate, sports or politics, etc.) with word frequencies because the options. With applicable pre-processing, it's competitive during this domain with a lot of advanced strategies as well as support vector machines. It additionally finds application in automatic diagnosis.

Naive bayes classifiers square measure extremely climbable, requiring range|variety} of parameters linear within the number of variables (features/predictors) during a learning downside. Maximum-likelihood coaching are often done by evaluating a closed-form expression, that takes linear time, instead of by expensive repetitious approximation as used for several different kinds of classifiers.

An advantage of naive bayes is that it solely needs alittle quantity of training knowledge to estimate the parameters necessary for classification.

1.3 Supervised Methods

The supervised traffic categoryification strategies analyze the supervised coaching knowledge and turn out an inferred operate which may predict the output class for any testing flow. In supervised traffic classification, adequate supervised training knowledge may be a general assumption. to deal with the issues suffered by payload-based traffic classification, like encrypted applications and user knowledge privacy, Moore and Zuev applied the supervised naive bayes techniques to classify network traffic supported flow applied mathematics options. Williams et al. evaluated the supervised algorithms as well as naive bayes with discretization, naive bayes with kernel density estimation, C4.5 decision tree, theorem network, and naive bayes tree. Nguyen and Armitage planned to conduct traffic classification supported the recent packets of a flow for time period purpose. Auld et al. extended the work of with the applying of bayesian neural networks for correct traffic classification.

1.4 Unsupervised Methods

The unsupervised methods attempt to realize cluster structure in unlabelled traffic data and assign any testing flow to the application-based category of its nearest cluster. McGregor et al. planned to cluster traffic flows into atiny low range of clusters mistreatment the expectation maximization (EM) formula and manually label every cluster to an application. Zander et al. used AutoClass to cluster traffic flows and planned a metric known as intraclass homogeneity for cluster analysis. Bernaille et al. applied the k-means formula to traffic agglomeration and labeled the clusters to applications employing a payload analysis tool. Erman et al. evaluated the k-means, DBSCAN and AutoClass algorithms for traffic agglomeration on two empirical knowledge traces. The enquiry showed that traffic clustering will turn out high-purity clusters once the amount of clusters is about the maximum amount larger than the amount of real applications. Generally, the cluster techniques will be accustomed discover traffic from antecedently unknown applications . Wang et al. planned to integrate applied mathematics feature-based flow clustering with payload signature matching methodology, thus on eliminate the need of supervised coaching knowledge. Finamore et al. combined flow applied mathematics feature-based clustering and payload applied mathematics feature-based clustering for mining unidentified traffic. However, the clustering strategies suffer from a problem of mapping from an outsized range of clusters to real applications.

2. 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 overfitting 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.

3. PROPSED METHODLOLGY

The problems suffered by payload-based traffic classification, like encrypted applications and user data privacy, Moore and applied the supervised naive techniques to classify network traffic supported flow applied math choices.Evaluated the supervised algorithms along side naive Thomas Bayes with discretization, naive Thomas Bayes with kernel density estimation, C4.5 decision tree, Bayesian network, and naive Thomas Bayes tree. Nguyen and Armitage planned to conduct traffic classification supported the recent packets of a flow for amount of your time purpose. Extended the work of with the appliance of Bayesian neural networks for proper traffic classification. Used unidirectional statistical choices for traffic classification inside the network core associate degreed projected an formula with the potential of estimating the missing options. planned to use only the size of the first packets of associate degree SSL affiliation to acknowledge the encrypted applications projected to analyze the message content randomness introduced by the secret writing method victimisation Pearson's chi-Square test-based technique. The likelihood density perform (PDF)-based protocol fingerprints to specific 3 traffic statistical properties throughout a compact approach. Their work is extended with a continuing improvement procedure.

Advantages

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

- Evaluated three supervised ways for an ADSL provider managing many points of presence, the results of that are corresponding to deep review solutions.
- Applied one class SVMs to traffic classification and given a straightforward improvement formula for each set of SVM in operation parameters projected to classify P2P-TV traffic using the count of packets modified with completely different peers throughout the insufficient time windows.

4. 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 I, it's additionally clear that MLP algorithmic rule provides very poor performance in terms of classification accuracy and coaching time. moreover, classification accuracy is of RBF Neural internetwork Classifier is additionally lesser than that of different ml classifiers and its coaching time is incredibly massive 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	RBF	C4.5	Bayes Net	Naive Bayes
Classification Accuracy (%)	27.75	81.25	83.75	88.125	88.875
Training Time (Seconds)	17.79	6.14	1.34	0.7	0.16

5. IMPLEMENTATION DETAILS:

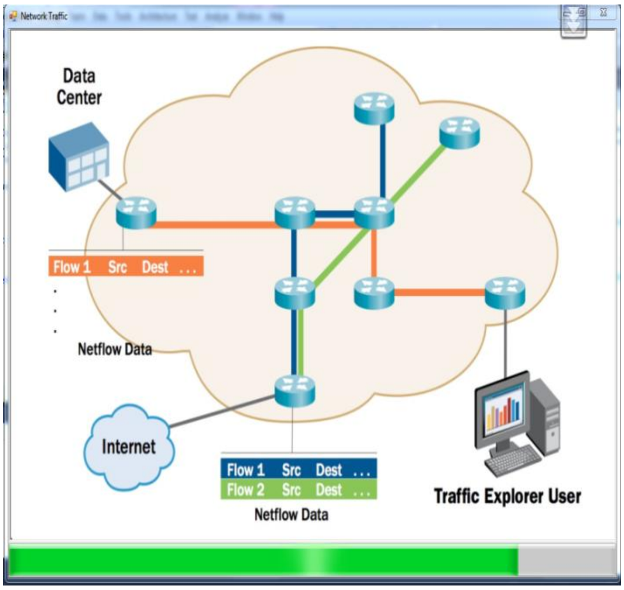
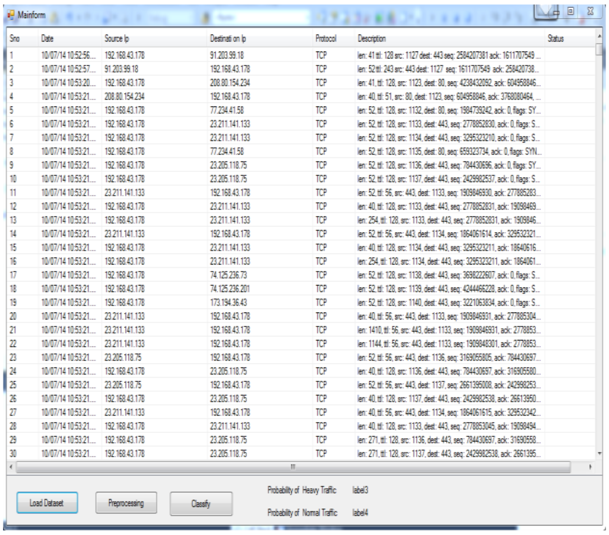


Figure 1: Welcome page

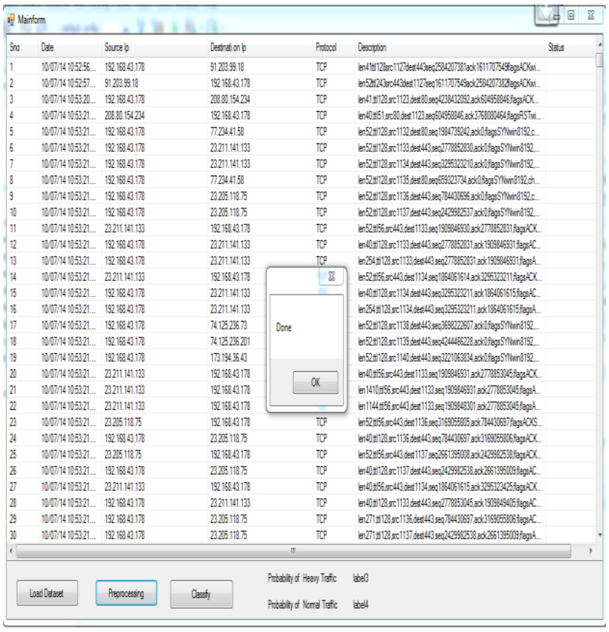


Sno	Date	Source Ip	Destination Ip	Protocol	Description	Status
1	10/07/14 10:52:56...	192.168.43.178	91.203.99.18	TCP	len: 41 ttl: 128 src: 1127 dest: 443 seq: 2584207381 ack: 1611707549 ...	
2	10/07/14 10:52:57...	91.203.99.18	192.168.43.178	TCP	len: 52 ttl: 243 src: 443 dest: 1127 seq: 1611707549 ack: 258420738...	
3	10/07/14 10:53:20...	192.168.43.178	208.80.154.234	TCP	len: 41, ttl: 128, src: 1123, dest: 80, seq: 4238432092, ack: 604958846...	
4	10/07/14 10:53:21...	208.80.154.234	192.168.43.178	TCP	len: 40, ttl: 51, src: 80, dest: 1123, seq: 604958846, ack: 3768080464, ...	
5	10/07/14 10:53:21...	192.168.43.178	77.234.41.58	TCP	len: 52, ttl: 128, src: 1132, dest: 80, seq: 1984739242, ack: 0, flags: SY...	
6	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 52, ttl: 128, src: 1133, dest: 443, seq: 2778852830, ack: 0, flags: S...	
7	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 52, ttl: 128, src: 1134, dest: 443, seq: 3295323210, ack: 0, flags: S...	
8	10/07/14 10:53:21...	192.168.43.178	77.234.41.58	TCP	len: 52, ttl: 128, src: 1135, dest: 80, seq: 659323734, ack: 0, flags: SYN...	
9	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len: 52, ttl: 128, src: 1136, dest: 443, seq: 784430696, ack: 0, flags: SY...	
10	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len: 52, ttl: 128, src: 1137, dest: 443, seq: 2429982537, ack: 0, flags: S...	
11	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len: 52, ttl: 56, src: 443, dest: 1133, seq: 1909846930, ack: 277885283...	
12	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 40, ttl: 128, src: 1133, dest: 443, seq: 2778852831, ack: 19098469...	
13	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 254, ttl: 128, src: 1133, dest: 443, seq: 2778852831, ack: 1909846...	
14	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len: 52, ttl: 56, src: 443, dest: 1134, seq: 1864061614, ack: 329532321...	
15	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 40, ttl: 128, src: 1134, dest: 443, seq: 3295323211, ack: 18640616...	
16	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 254, ttl: 128, src: 1134, dest: 443, seq: 3295323211, ack: 1864061...	
17	10/07/14 10:53:21...	192.168.43.178	74.125.236.73	TCP	len: 52, ttl: 128, src: 1138, dest: 443, seq: 3698222607, ack: 0, flags: S...	
18	10/07/14 10:53:21...	192.168.43.178	74.125.236.201	TCP	len: 52, ttl: 128, src: 1139, dest: 443, seq: 4244466228, ack: 0, flags: S...	
19	10/07/14 10:53:21...	192.168.43.178	173.194.36.43	TCP	len: 52, ttl: 128, src: 1140, dest: 443, seq: 3221063834, ack: 0, flags: S...	
20	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len: 40, ttl: 56, src: 443, dest: 1133, seq: 1909846931, ack: 277885304...	
21	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len: 1410, ttl: 56, src: 443, dest: 1133, seq: 1909846931, ack: 2778853...	
22	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len: 1144, ttl: 56, src: 443, dest: 1133, seq: 1909848301, ack: 2778853...	
23	10/07/14 10:53:21...	23.205.118.75	192.168.43.178	TCP	len: 52, ttl: 56, src: 443, dest: 1136, seq: 3169055805, ack: 784430697...	
24	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len: 40, ttl: 128, src: 1136, dest: 443, seq: 784430697, ack: 316905580...	
25	10/07/14 10:53:21...	23.205.118.75	192.168.43.178	TCP	len: 52, ttl: 56, src: 443, dest: 1137, seq: 2661395008, ack: 242998253...	
26	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len: 40, ttl: 128, src: 1137, dest: 443, seq: 2429982538, ack: 26613950...	
27	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len: 40, ttl: 56, src: 443, dest: 1134, seq: 1864061615, ack: 329532342...	
28	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len: 40, ttl: 128, src: 1133, dest: 443, seq: 2778853045, ack: 19098494...	
29	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len: 271, ttl: 128, src: 1136, dest: 443, seq: 784430697, ack: 31690558...	
30	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len: 271, ttl: 128, src: 1137, dest: 443, seq: 2429982538, ack: 2661395...	

Probability of Heavy Traffic label3

Probability of Normal Traffic label4

Figure 2: Loaded dataset



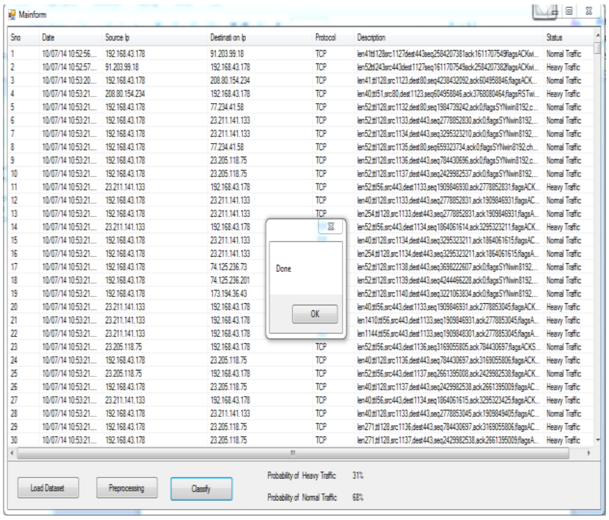
Sno	Date	Source Ip	Destination Ip	Protocol	Description	Status
1	10/07/14 10:52:56...	192.168.43.178	91.203.99.18	TCP	len41ttl128src1127dest443seq2584207381ack1611707549flagsACKwi...	
2	10/07/14 10:52:57...	91.203.99.18	192.168.43.178	TCP	len52ttl243src443dest1127seq1611707549ack2584207382flagsACKwi...	
3	10/07/14 10:53:20...	192.168.43.178	208.80.154.234	TCP	len41,ttl128,src1123,dest80,seq4238432092,ack604958846,flagsACK...	
4	10/07/14 10:53:21...	208.80.154.234	192.168.43.178	TCP	len40,ttl51,src80,dest1123,seq604958846,ack3768080464,flagsRSTwi...	
5	10/07/14 10:53:21...	192.168.43.178	77.234.41.58	TCP	len52,ttl128,src1132,dest80,seq1984739242,ack0,flagsSYNwin8192,c...	
6	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len52,ttl128,src1133,dest443,seq2778852830,ack0,flagsSYNwin8192,...	
7	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len52,ttl128,src1134,dest443,seq3295323210,ack0,flagsSYNwin8192,...	
8	10/07/14 10:53:21...	192.168.43.178	77.234.41.58	TCP	len52,ttl128,src1135,dest80,seq659323734,ack0,flagsSYNwin8192,ch...	
9	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len52,ttl128,src1136,dest443,seq784430696,ack0,flagsSYNwin8192,c...	
10	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len52,ttl128,src1137,dest443,seq2429982537,ack0,flagsSYNwin8192,...	
11	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len52,ttl56,src443,dest1133,seq1909846930,ack2778852831,flagsACK...	
12	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len40,ttl128,src1133,dest443,seq2778852831,ack1909846931,flagsAC...	
13	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len254,ttl128,src1133,dest443,seq2778852831,ack1909846931,flagsA...	
14	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len52,ttl56,src443,dest1134,seq1864061614,ack3295323211,flagsACK...	
15	10/07/14 10:53:21...	192.168.43.178	23.211.141.133		len40,ttl128,src1134,dest443,seq3295323211,ack1864061615,flagsAC...	
16	10/07/14 10:53:21...	192.168.43.178	23.211.141.133		len254,ttl128,src1134,dest443,seq3295323211,ack1864061615,flagsA...	
17	10/07/14 10:53:21...	192.168.43.178	74.125.236.73		len52,ttl128,src1138,dest443,seq3698222607,ack0,flagsSYNwin8192,...	
18	10/07/14 10:53:21...	192.168.43.178	74.125.236.201		len52,ttl128,src1139,dest443,seq4244466228,ack0,flagsSYNwin8192,...	
19	10/07/14 10:53:21...	192.168.43.178	173.194.36.43		len52,ttl128,src1140,dest443,seq3221063834,ack0,flagsSYNwin8192,...	
20	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len40,ttl56,src443,dest1133,seq1909846931,ack2778853045,flagsACK...	
21	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len1410,ttl56,src443,dest1133,seq1909846931,ack2778853045,flagsA...	
22	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len1144,ttl56,src443,dest1133,seq1909848301,ack2778853045,flagsA...	
23	10/07/14 10:53:21...	23.205.118.75	192.168.43.178	TCP	len52,ttl56,src443,dest1136,seq3169055805,ack784430697,flagsACKS...	
24	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len40,ttl128,src1136,dest443,seq784430697,ack3169055806,flagsACK...	
25	10/07/14 10:53:21...	23.205.118.75	192.168.43.178	TCP	len52,ttl56,src443,dest1137,seq2661395008,ack2429982538,flagsACK...	
26	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len40,ttl128,src1137,dest443,seq2429982538,ack2661395009,flagsAC...	
27	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len40,ttl56,src443,dest1134,seq1864061615,ack3295323425,flagsACK...	
28	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len40,ttl128,src1133,dest443,seq2778853045,ack1909849405,flagsAC...	
29	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len271,ttl128,src1136,dest443,seq784430697,ack3169055806,flagsAC...	
30	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len271,ttl128,src1137,dest443,seq2429982538,ack2661395009,flagsA...	

Done

Probability of Heavy Traffic label3

Probability of Normal Traffic label4

Figure 3: Result after Preprocessing



Sno	Date	Source Ip	Destination Ip	Protocol	Description	Status
1	10/07/14 10:52:56...	192.168.43.178	91.203.99.18	TCP	len41ttl128src1127dest443seq2584207381ack1611707549flagsACKwi...	Normal Traffic
2	10/07/14 10:52:57...	91.203.99.18	192.168.43.178	TCP	len52ttl243src443dest1127seq1611707549ack2584207382flagsACKwi...	Heavy Traffic
3	10/07/14 10:53:20...	192.168.43.178	208.80.154.234	TCP	len41,ttl128,src1123,dest80,seq4238432092,ack604958846,flagsACK...	Normal Traffic
4	10/07/14 10:53:21...	208.80.154.234	192.168.43.178	TCP	len40,ttl51,src80,dest1123,seq604958846,ack3768080464,flagsRSTwi...	Heavy Traffic
5	10/07/14 10:53:21...	192.168.43.178	77.234.41.58	TCP	len52,ttl128,src1132,dest80,seq1984739242,ack0,flagsSYNwin8192,c...	Normal Traffic
6	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len52,ttl128,src1133,dest443,seq2778852830,ack0,flagsSYNwin8192,...	Normal Traffic
7	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len52,ttl128,src1134,dest443,seq3295323210,ack0,flagsSYNwin8192,...	Normal Traffic
8	10/07/14 10:53:21...	192.168.43.178	77.234.41.58	TCP	len52,ttl128,src1135,dest80,seq659323734,ack0,flagsSYNwin8192,ch...	Normal Traffic
9	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len52,ttl128,src1136,dest443,seq784430696,ack0,flagsSYNwin8192,c...	Normal Traffic
10	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len52,ttl128,src1137,dest443,seq2429982537,ack0,flagsSYNwin8192,...	Normal Traffic
11	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len52,ttl56,src443,dest1133,seq1909846930,ack2778852831,flagsACK...	Heavy Traffic
12	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len40,ttl128,src1133,dest443,seq2778852831,ack1909846931,flagsAC...	Normal Traffic
13	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len254,ttl128,src1133,dest443,seq2778852831,ack1909846931,flagsA...	Normal Traffic
14	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len52,ttl56,src443,dest1134,seq1864061614,ack3295323211,flagsACK...	Heavy Traffic
15	10/07/14 10:53:21...	192.168.43.178	23.211.141.133		len40,ttl128,src1134,dest443,seq3295323211,ack1864061615,flagsAC...	Normal Traffic
16	10/07/14 10:53:21...	192.168.43.178	23.211.141.133		len254,ttl128,src1134,dest443,seq3295323211,ack1864061615,flagsA...	Normal Traffic
17	10/07/14 10:53:21...	192.168.43.178	74.125.236.73		len52,ttl128,src1138,dest443,seq3698222607,ack0,flagsSYNwin8192,...	Normal Traffic
18	10/07/14 10:53:21...	192.168.43.178	74.125.236.201		len52,ttl128,src1139,dest443,seq4244466228,ack0,flagsSYNwin8192,...	Normal Traffic
19	10/07/14 10:53:21...	192.168.43.178	173.194.36.43		len52,ttl128,src1140,dest443,seq3221063834,ack0,flagsSYNwin8192,...	Normal Traffic
20	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len40,ttl56,src443,dest1133,seq1909846931,ack2778853045,flagsACK...	Heavy Traffic
21	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len1410,ttl56,src443,dest1133,seq1909846931,ack2778853045,flagsA...	Heavy Traffic
22	10/07/14 10:53:21...	23.211.141.133	192.168.43.178		len1144,ttl56,src443,dest1133,seq1909848301,ack2778853045,flagsA...	Heavy Traffic
23	10/07/14 10:53:21...	23.205.118.75	192.168.43.178	TCP	len52,ttl56,src443,dest1136,seq3169055805,ack784430697,flagsACKS...	Normal Traffic
24	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len40,ttl128,src1136,dest443,seq784430697,ack3169055806,flagsACK...	Heavy Traffic
25	10/07/14 10:53:21...	23.205.118.75	192.168.43.178	TCP	len52,ttl56,src443,dest1137,seq2661395008,ack2429982538,flagsACK...	Normal Traffic
26	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len40,ttl128,src1137,dest443,seq2429982538,ack2661395009,flagsAC...	Heavy Traffic
27	10/07/14 10:53:21...	23.211.141.133	192.168.43.178	TCP	len40,ttl56,src443,dest1134,seq1864061615,ack3295323425,flagsACK...	Heavy Traffic
28	10/07/14 10:53:21...	192.168.43.178	23.211.141.133	TCP	len40,ttl128,src1133,dest443,seq2778853045,ack1909849405,flagsAC...	Normal Traffic
29	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len271,ttl128,src1136,dest443,seq784430697,ack3169055806,flagsAC...	Heavy Traffic
30	10/07/14 10:53:21...	192.168.43.178	23.205.118.75	TCP	len271,ttl128,src1137,dest443,seq2429982538,ack2661395009,flagsA...	Heavy Traffic

Done

Probability of Heavy Traffic 31%

Probability of Normal Traffic 68%

Figure 4: Classify Preprocessed dataset

6. 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 Naïve 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.

7. 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

[5] 2015 E-ISSN: 2321-9637.

[6] Ms. G. Rubadevi, Mrs. R. Amsaveni, **"**Internet traffic classification using Hybrid Aggregated classifier and Neural Network**",** International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 3, Issue 10 October, 2014 Page No. 8964-8971.

[7] Jamuna .A, Vinodh Ewards S .E, "Efficient Flow based Network Traffic Classification using Machine Learning"**,** International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 2, March -April 2013, pp.1324-1328.