IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

An Efficient NB-IWD Based Network Traffic Classification over KDD Dataset

Amit Kumar[1*], Daya Shankar Pandey[2], Varsha Namdeo[3]

RKDF IST College, SRK University, India

Corresponding Author: amitkumar1989.niec@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.706710 | Available online at: www.ijcseonline.org

Accepted: 16/Apr/2019, Published: 30/Apr/2019

***Abstract*—** The Network traffic arrangement is a procedure by which the large chips get away at different parameters for instance port and convention based which are utilized to identify the classes of the traffic. Thus these types of classification methods are very helpful in providing security at two levels- network as well as system. The main focus of this paper is sorting out the problem which comes while handling network traffic whereas some of the traffic classification methods are unable to find out the special requirements of individual datasets because there are massive measures of network traffic datasets and restricted quantities of resources are accessible to deliver classification examination. The paper uncovers that traffic arrangement should be refreshed normally to keep up the precision and ought to have the capacity to adjust the dynamic conduct of network stream.

***Keywords*—** Network Traffic, Network Traffic Class, Network Features, Statistical features, Classification.

I. INTRODUCTION

The main concern of network monitoring is measuring the performance of the high speed network in a centralized way. The different types of networks carry data for many different kinds of applications which have their own requirements of performance. As soon as the system is connected to a network, the flow of packet starts right from there. Flows are generally considered to bidirectional i.e. one from the local machine or system to the server to which it is connected and other vice versa. A very simple approach to classify network is through realization of the group of packets having same IP address, same transport protocols and port number. This type of technique is very effective for protocols using fixed port number but in case of dynamic port number it doesn't work up to its expectations. On the other hand deep packet inspection (DPI) technique is quiet slow which requires a lot of processing power. Machine learning algorithms on the other hand require training data for their initial phase of learning. Our paper introduces Naïve Bayes algorithm in traffic classification of the network and leaves a very clear impression about how it is better than other previous techniques and methodologies as it is able to give an accuracy of above 99% in the classification.

1. Need for Network Traffic Classification: Arrangement of system traffic is the primary method to recognize different applications and conventions. When the parcels are delegated having a place with a specific application or convention, they are checked or hailed. Stamping is the procedure that hues the bundles dependent on certain grouping strategies, to give suitable treatment to those parcels.

2. Classification Attributes: Classification attributes generally fall under many criteria's. Generally using this type of algorithm for classification, attributes are divided into two types of disjoint sets. Actually flow of the packets is recognized on the basis of proportions of inbound to outbound payload bytes of the classified flow. Each flow is divided into X groups of Y packets where Y depends on the current iteration and X stands for the count of obtained groups. Each group from these two disjoint sets is used for the generation of one training case and another testing case for the classifier. Many features are taken into account like number of inbound/outbound/total payload bytes in the sample, ratio of all small inbound and outbound data packets, ratio of all large data packets, applications used etc. The other part exclusively contains protocol dependent attributes like which transport protocol is being used, number of ACK/PSH flags for inbound/outbound direction, local port and remote part etc.

The purpose of network classification system:

It is needed to have a proper understanding of the applications and protocols in the network.

It is also required in implementation of appropriate security policies. It helps in informing about the real attacks which is far below the false alarm rate.

Last but not the least the prime concern lies in boosting the quality of service and analysis of the traffic.

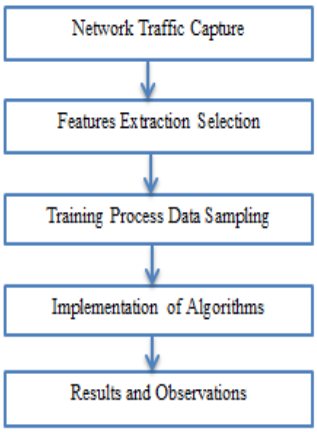

Figure 1: Traffic Classification.

The above figure 1 states the process involves in the classification of the traffic.

II. RELATED WORK

In machine learning based traffic classification there are three main challenges to consider. They are identifying the feature, identifying the best algorithm for classification and obtaining the training datasets.

In [3] an unsupervised learning approach Expectation Maximization for classifying the internet traffic which are represented by a set of flow attribute such as size of the packet, duration of connection, count of bytes and inter-arrival time.

K-means online approach for classifying the traffic by using just initial five parcels of the stream is introduced in [4]. Better outcomes were acquired with the initial five parcels that are over 80%. In any case, this outcome can't be accomplished if any of the bundles from the initial five parcels are missed.

Some classifiers like Bayesian network, Naïve Bayes, C4.5, Naïve Bayes tree are compared in [5]. From their analysis they have found that C4.5 algorithm performs well in terms of accuracy than other algorithms. An approach for classifying the traffic into different type of services is presented in [6]. Here along with the kernel estimation, a Naïve Bayes classifier is combined for classifying offline TCP traces. It achieves 96% accuracy.

Here features are selected using correlation based filter algorithm. Accuracy of decision tree algorithms like C5.0 and random forest are always stable than other algorithms. They also stated that build time of the real time features are low. So decision tree algorithms are also well suited for online traffic classification [7].

III. MODULES DESCRIPTION

Supervised Methods

Supervised techniques, otherwise called arrangement or classification strategies, separate information structures to arrange new cases in pre-characterized classes. It is essential to note that is called managed in light of the fact that the yield classes are pre-characterized. The procedure of a regulated ML techniques begin with a preparation dataset TS characterized as,

T S = < x1, y1 >, < x2, y1 >, ..., < xN , yM >, where xi is the vector of estimations of the highlights relating to the I th example, and yi is its yield class esteem. It finds the distinctive relations between the occurrences and yields a structure, normally a choice tree or order governs, that will characterize the cases in a discrete set y1, y2, yM. There is a great deal of related work that utilization administered procedures [8] with promising outcomes. The supervised traffic grouping strategies dissect the managed preparing information and produce a deduced capacity which can anticipate the yield class for any testing flow. In regulated rush hour gridlock characterization, adequate managed preparing information is a general supposition. To address the issues endured by payload-based traffic characterization, for example, scrambled applications and client information.

Unsupervised Methods

The unsupervised techniques (or grouping) endeavour to discover bunch structure in unlabelled rush hour gridlock information and allocate any testing stream to the application-based class of its closest group. The proposed to bunch traffic streams into few groups utilizing the desire expansion (EM) calculation and physically mark each group to an application [9-10].

A Traffic Classification Approach with Flow Correlation

It exhibits another structure which we call Traffic Classification utilizing Correlation data or TCC for short. An epic nonparametric methodology is additionally proposed to viably consolidate stream connection data into the order procedure.

Correlation Analysis

The associated flows having a similar three-tuple are produced by a similar application. For instance, a few flows started by various hosts are for the most part interfacing with an comparable host at TCP port 80 of each a concise period. These streams are in all regards likely created by a comparative application, for instance, a web program. The three-tuple heuristic about stream relationship has been considered in a few reasonable rush hour gridlock order plans proposed a payload based bunching strategy for convention surmising, in which they gathered flows into equality groups utilizing the heuristic. Tried the rightness of the three-tuple heuristic with certifiable follows [13].

Network Flexibility

The proposed system show is available to include extraction and relationship investigation. Initially, any sorts of flow measurable highlights can be connected in our system display. In this work, we extricate unidirectional factual highlights from full flows. The measurable highlights separated from parts of flows can likewise be utilized to speak to traffic flows in our system display. Second, any new relationship examination technique can be implanted into our system demonstrate. We acquaint flow relationship investigation with find connection data in rush hour gridlock flows to improve the strength of grouping.

IV. RESULTS AND DISCUSSION

In proposed system, a novel parametric approach is used to deal with the correlated flows in an effective way, which can significantly improve the classification performance.

A. Pre-processing

Here the IP packets crossing across a network is collected and used for constructing the flows by examining the header of packets.

B. Correlation Based Feature Selection

Here measurable highlights are extricated and are utilized to speak to traffic streams that are finished by pre-preparing to apply include choice to expel superfluous and repetitive highlights from the list of capabilities.

C. Feature Discretization

Discretization is a process of converting numeric values into intervals and associating them to a nominal symbol. These symbols are then used as new values instead of the original numeric values.

D. Naïve Bayes Classification

A Naïve-Bayes (NB) ML algorithm is a simple structure consisting of a class node as the parent node of all other nodes. The basic structures of Naïve Bayes Classifier is shown in Figure 2 in which C represents main class and a, b, c and d represents other feature or attribute nodes of a particular sample. No other connections are allowed in a Naïve-Bayes structure. Naïve-Bayes has been used as an effective classifier. . It is easy to construct Naïve Bayes classifier as compared to other classifiers because the structure is given a priori and hence no structure learning procedure is required. Naive Bayes assumes that all the features are independent of each other. Naïve-Bayes works very well over a large number of datasets, especially where the features used to characterize each sample are not properly correlated.

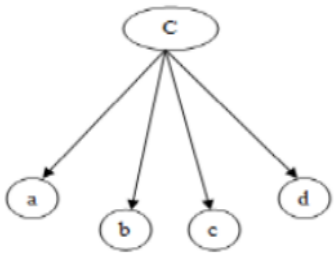

Figure 2: Naïve Bayes Classifier.

In the above figure 2 the Naïve Bayes Classifier is shown.

Table 1: Statistical Features.

Types of Features	Feature Description	Number
Packets	Number of packets transferred	1
Bytes	Volume of bytes transferred	1
Packet Size	Min, Max, Median and Standard. Deviation of packet size	6
Inter Packet Time	Min, Max, Median and Standard. Deviation of inter packet time	6
	Total	14

In the above table 1 the statistical features have been shown that contains types of features-packets, bytes, packet size, inter packet time comparison.

A. Performance Evaluation

Performance evaluation of the algorithm is done by using the following metrics: overall accuracy, precision, recall, F-measure, and classification speed.

	Negative (predicted)	Positive (predicted)
Negative (actual)	true negative	false positive
Positive (actual)	false negative	true positive

F-Measure: It can be defined as an information retrieval (IR) system has recall R and precision P on a test document collection and an information need.

$$F1 = 2 * \{(precision * recall)/(precision * recall)\}$$

Accuracy: It can be defined as the sum of all the values that is divided by the given set of numbers.

Accurcay= truepositives + truenegatives/total examples

Recall: The ratio of True Positives over the sum of True Positives and False Negatives.

Table 2: Comparative Analysis.

Algorithms	Naïve Bayes	BayesNet
Correctly classified	270	331
Incorrectly Classified	86	92
Overall Accuracy (%)	73.232	90.07
Error (%)	3.76	2.92
Time (sec)	0.04	0.5

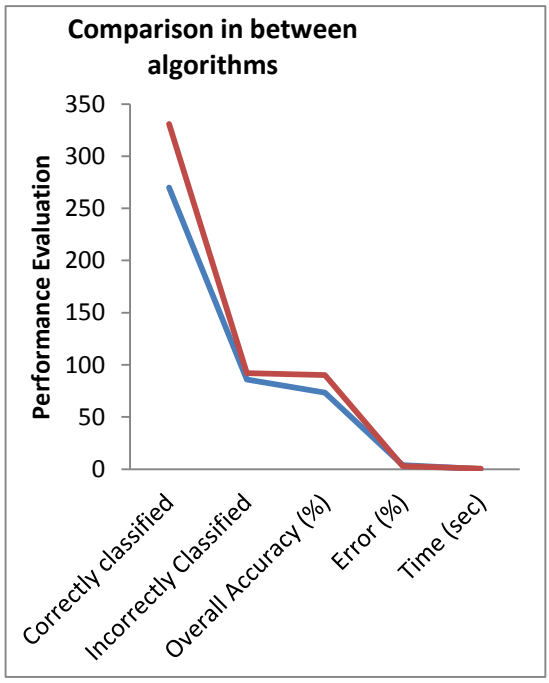

Figure 3: Graphical Analysis.

In the above figure 1 the graphical comparison has been shown.

V. Conclusion and Future Scope

Traffic classification plays an important role in the network security as the applications and their behaviour are changing day to day. As a result there increased the need for accurate classification of the network flows. Here we have proposed a Naïve Bayes model with feature selection for the accurate classification of internet traffic. We have compared the method with three other Bayesian models. Our experiment shows that it provides an accuracy of 96.5% which is better than that of the other state-of-the art methods. NBD is easy to build and is applicable to various real world applications.

Acknowledgment

I would like to thank my guide (Daya Shankar Pandey) & Head of Department (Dr. Varsha Namdeo) from RKDF IST College, SRK University, India for supporting this work.

References

[1]. R. Kwitt and U. Hofmann, "Unsupervised oddity recognition in system traffic by methods for strong PCA", in Proc. Int. Conf. Processing in the Global Information Technology (ICCGI), **2007, pp. 3737.**

[2]. Anshul Vishwakarma1* , Amit Khare2, "Vehicle Detection and Tracking for Traffic Surveillance Applications: A Review Paper", **Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693.**

[3]. McGregor, M. Lobby, P. Lorier, and J. Brunskill. "Stream grouping utilizing AI methods". LectureNotes in Computer Science, 2004, **pp.205– 214.**
[4]. L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. "Salamatian. Traffic order on the flySIGCOMM Comput. Commun". 2006.
[5]. Williams N., Zander S., Armitage G., "A Prelimenary Performanc Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Comparison", ACM SIGCOMM Computer Communication Review, **Vol. 36, No. 5, 2006, pp. 5-16.**
[6]. Andrew W. Moore and Denis Zuev. "Web traffic order utilizing bayesian examination procedures". In Proceedings of ACM Sigmetrics,**2005, pp.50-60.**
[7]. ZHAO Jing-jing, HUANG Xiao-hong, SUN Qiong, MA Yan, "Real-time highlight choice in rush hour gridlock grouping", ELSEVIER, **2008**.
[8]. W. Yurcik and Y. Li, "Internet security representation contextual analysis: Instru-menting a system for netflow security perception devices", in Proc. Yearly Computer Security Applications Conf. (ACSAC 05), Tucson, AZ, **Dec. 59, 2005.**
[9]. K. Thyagarajan 1* , N. Vaishnavi 2, "Performance Study on Malicious Program Prediction Using Classification Techniques", **Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693.**
[10]. Y. Lim, H. Kim, J. Jeong, C. Kim, T. Kwon, and Y. Choi, "Internet Traffic Classification Demystified: On the Sources of the Discriminative Power", Proc. ACM CoNEXT, **2010, p. 9.**

Authors Profile

Mr. Amit kumar had completed B.Tech in CSE Branch from GGSIPU New Delhi in 2011, and currently pursuing M.Tech from RKDF IST College, SRK University, India.

Dayashankar pandey is Assistant professor in the Department of Computer Science & Applications in Sarvepalli Radhakrishnan University, Bhopal, India. He is a teacher in the field of computer science and information technology.

Dr. Varsha Namdeo is Professor in the Department of Computer Science & Applications in Sarvepalli Radhakrishnan University, Bhopal, India. She is a teacher and researcher in the field of computer science and information technology. She earned her Master degree in Computer Application from Barkatullah University Bhopal (M.P.) in 2000 and in Computer Science and Engineering from Barkatullah University Bhopal (M.P.) in 2009, and PhD degree from Maulana Azad National Institute of Technology, Bhopal (M.P.). Currently, she is guiding several PhD research scholars.

