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. 4, April 2017, pg.300 – 307

Survey: Feature Selection Technique Impact for Internet Traffic Classification Using Naïve Bayesian

Satish Chadokar, Ashish Kumbhare

Department of Computer Science and Engineering, RGPV, Bhopal, India
Department of Computer Science and Engineering, RGPV, Bhopal, India
chadokarsatish111@gmail.com; ashishkumbhare99@gmail.com

Abstract— Internet traffic defines as the density of data or information presented on the Internet or in another language we can say it's a flow of data on the internet. Internet traffic classification has power to solve many network difficulties and manage different type of network problems. There are some basic functions provided to government, Internet service providers (ISPs) and network administrator through Internet traffic classification. Machine learning approaches overcome many problems of traditional approaches of internet traffic classification. In supervised approaches, we discuss five well known supervised machine learning approaches these are Naïve Bayes, Feed Forward Neural Network, Bayes Net, RBF and C4.5 decision tree approach.

Keywords— "ISP, ML, DAG, RBF, FCBF, BOF"

I. INTRODUCTION

Machine learning based internet traffic classification is very popular now because machine learning approaches give better results. And also overcome the problems of previous traditional port based and payload based internet traffic classification. It is also deal with new P2P application, online gaming application and many other application which do not have any registered port number in IANA (Internet authority of number Assign) and they have dynamic port number so to classify it with traditional approach its difficult and payload based techniques do not deal with encrypted data so that's why using machine learning approach for classification of internet traffic. In this chapter we discuss previous work done in machine leaning based internet traffic classification. There are two types of ML techniques first is supervised learning (Classification) and another one is unsupervised Learning (Clustering).

II. LITERATURE SURVEY

i. Supervised Machine Learning Approach

Supervised machine learning approaches have two working phases first is Training (Learning) and the second one is Testing phase.

a. Bayes Net Approach

Bayes Net is structured like a combination of a DAG (directed acyclic graph) of nodes and links, and a set of conditional probability tables. It is also called belief network and here, Nodes represent attributes or classes, while links between nodes represent the relationship amongst them. Conditional probability tables determine the strength of the links. There is one probability table for every node (attribute) that defines the probability distribution for the node given its parent nodes. If a node has no parents the probability distribution is unconditional. If a node has one or more than one parents the probability distribution is a conditional distribution, where the probability of each feature value depends on the values of the parents.

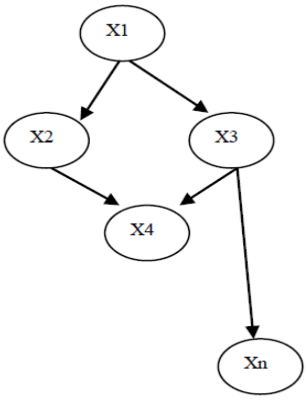

Fig. 1: Basic structure of Bayes Net

Conditional Probability Distribution function
P (U) = Πni=1 P(xi| Paxi) 1
Where,
P(U) = Set of variables (x1,x2,x3...........,xn)
Pa = parents of xi

In [2013], Kuldeep Singh et al. [6] uses five machine learning algorithms (MLP, RBF, C4.5, Naïve Bayes, Bayes Net) to classify real time IP traffic. In this they prepared dataset by using a packet capturing tool Wireshark and captured packets for duration of 2 second and prepared datasets and now they apply feature selection algorithms to eliminate irrelevant features for this they using correlation and consistency based feature selection algorithms for feature reduction. Correlation based FS (feature selection) algorithm is used for identifying and reducing number of features which are redundant and not defining a particular type of traffic of internet and consistency based FS algorithm first compute different number of subsets of features and after that it select the optimal subset of features which contain less number of features. Result reported in this paper show 91% of classification accuracy of Bayes net.

In 2012 S. Agrawal et al. [13] use three machine learning algorithm (C4.5, Bayes Net and RBF) to classify internet traffic classification for academic perspective. They classify the website of an educational institution into two category first is an educational website which includes website like www.ieeexplore.ieee.org, www.sciencedirect.com etc., and the second one is non educational website which include like www.yahoomassenger.com, www.movies.com etc. they prepared dataset by using network capturing tools Wireshark and captured traffic of an educational institution for the duration of 1 minute in middle session of a day and prepare samples for testing and training purposes. They measure the performance on the basis of classification accuracy and training time, and they got that Bayes Net gives the better performance as compared to other two methods C4.5 and RBF. Bayes Net gives 76.67% classification accuracy with training time of 2 seconds.

b. Feed Forward Neural Network

Neural Network machine learning technique is inspired by animal brain. In this A collection of neurons connected together in a network is called neural Network and nodes are representing the neurons, and arrows are representing the links between nodes. Every node has its number, and a link connecting two nodes will have a pair of numbers. There are several types of neural network like feed forward neural network, back propagation neural network, radial basis network etc but the feed forward neural network is the simplest type in all above Networks without feedback to input or does not has a cycles (loops) are called a feed-forward networks (or perceptron). In feed forward neural network there are three basic terms these are one Input layer, one output layer and Hidden layers. Hidden layers may have more than one and at each neuron of hidden layers are contain some weight and we get the linear sum of all weight. The structure of feed forward neural network is shown in figure.

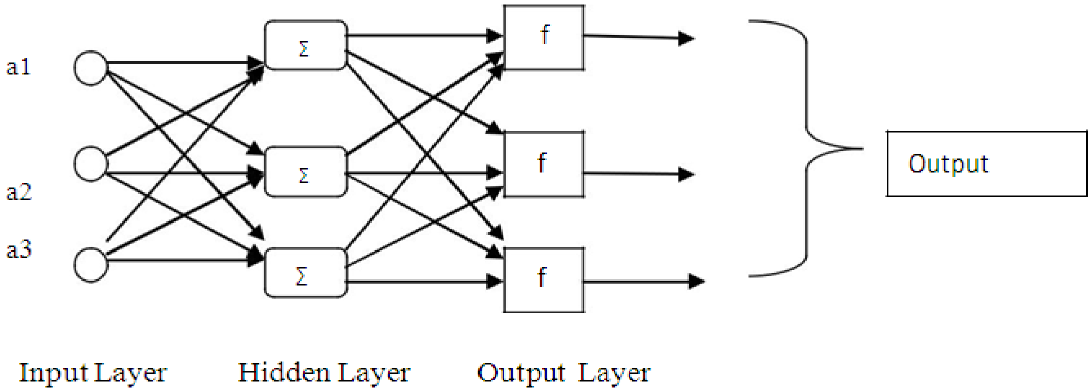

Fig. 2: Feed forward neural network

Where,
Transfer function f is determined by the user
Input = Aj (1 ≤ j ≤ k)
Weight = Wij (1 ≤ i ≤ u, 1 ≤ j ≤ k)
F = transfer function define by user
Output=ji (1 ≤ i ≤ u)
B bias added = bi (1 ≤ i ≤ u)
Here transfer function and weight are adjustable according to the output gain. In this structure there is a one hidden layer with three neurons.

In 2011 Wengang Zhou et al. [21] proposed an approach based on a feed forward neural network for accurate traffic classification and combined it with FCBF (Fast Correlation Based Feature) feature selection algorithm. FCBF is used for eliminating the redundant features and chosen the valuable features and feed forward neural network work as classifier. In this Bayesian regularization technique is used for training and this technique reduces a linear combination of squared errors and squared network parameters to keep safe the model from over-fitting for the datasets. In this paper, proposed method is compared with naïve bayes method and experimented result verifies that the proposed method is more robust and better.

c. Naive Byes Approach

The Naive Bayes is very simple Classifier technique is based on the Bayesian theorem (Bayes Rule's) with Strong and Weak independence (naive) assumptions. In the other way a Naive Bayes classifier assumes that the presence or absence of a particular feature of a class has not any relation with the presence or absence of any other features given in the same class variable and it is particularly suited when the dimensionality of the inputs is high.

Bayes Rule

$$P(E \mid H) = \frac{P(E \mid H).P(H)}{P(E)}$$

The concept of Bayes's rule is that the result outcome of an event or a hypothesis (H) can be predicted based on the evidences (E) that can be observed. From Bayes's rule, we have A priori probability of H, it is P(H) probability of a hypothesis before the evidence is observed. And a posterior probability of H, it is P(H | E) probability of a hypothesis after the evidence is observed.

In naive bayes classifier we predict outcomes of a hypothesis on basis of observing evidences and conceptually it is better to us that more than one evidences in the support of a hypothesis for prediction of result. A Naïve-Bays ML algorithm has a simple structure show in fig.3.

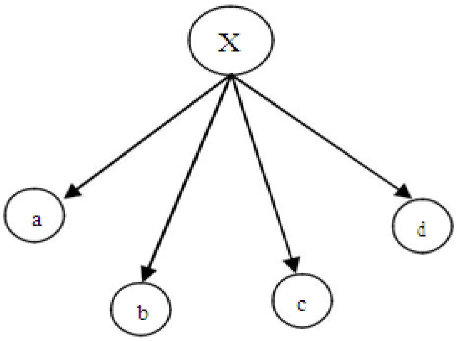

Fig. 3: Naive bayes

Where,
X = Class node
a, b, c, d = attribute nodes
This figure shows a parent node X and four children node a, b, c, d. In Naive byes all attribute are independent from each other and there is no learning structure required in naive bayes.

In 2013 Jun Zhang et al. [7] uses classify internet traffic by Aggregating Correlated Naive Bayes Prediction and get high accuracy with this approach. They proposed new (bag-of-flow) BoF-based traffic classification technique is to aggregate the Naive Bayes (NB) predictions of the correlated flow. They proposed a new approach of classification to utilize the information among the correlated traffic flows produced by the traffic. In the approach of classification there are two steps, in a first step the single naïve Bayes predictor generates the posteriori class-conditional probabilities or each flow and in a second step the aggregated predictor aggregates the flow predictions to determine the final class for BoFs.

d. C4.5 (Decision Tree) Approach

C4.5 is a popular decision tree Machine Learning algorithm used to develop Univariate decision tree. C4.5 is an enhancement of Iterative Dichotomiser 3 (ID3) algorithm which is used to find simple decision trees. C4.5 is also called a Statistical Classifier because of its good ability of classification. C4.5 makes decision trees from a set of training data samples, with the help of information entropy concept. The training data set contains of a greater number of training samples which are characterized by different attributes and it also consists of the target class. C4.5 selects a particular attribute of the data at each node of the tree which is used to split its set of data samples into subsets in one or another class. It is based on the criterion of normalized information gain that is obtained by selecting an attribute for splitting the data. The attribute with the highest normalized information gain is chosen and made a decision. After that, the C4.5 algorithm repeats the same action on the smaller subsets.

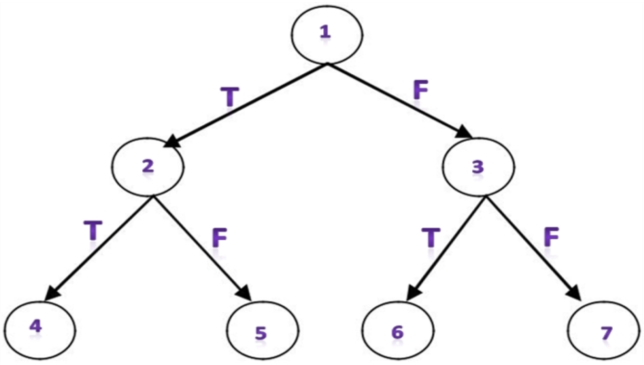

Fig. 4: Basic structure of C4.5

This figure shows simple structure of C4.5 decision tree where T and F represents True and False respectably. C4.5 has made various improvements to ID3 like it can handle both continuous attributes and discrete attributes, it can handle training data with missing attribute values, it can also handle attributes with differing costs etc.

In 2015 Kailas Elekar and et al.[2] They evaluated classification performance of five rule based classification algorithm these are Decision Table, OneR, JRip, ZeroR and PART. In this paper they uses KDD-CUP dataset and this dataset contains four type of attacks these are DoS, U2R, R2L and other attacks and they apply these algorithms to classify traffic and they got that PART algorithm is better than any of four algorithms in the terms of overall higher correctly classify instances and lower false attacks.

In 2012 Dong Shi et al. [15] they used to classify and identify the network with both supervised and unsupervised learning techniques. They use two types of dataset full features based and optimized features based. Here experiment result shows that the supervised ML algorithms give better result with feature reduction algorithms as compare to unsupervised ML algorithms. Simulation result concludes 99% classification accuracy with C4.5 algorithm.

e. Radial Basis Function (RBF) Approach

Radial basis function (RBF) networks have three layers architecture: an input layer, a hidden layer with a non-linear RBF function it an activation function and a linear output layer. Radial Basis Function (RBF) is a three layer feed forward artificial neural network which uses radial basis functions at each hidden layer neuron. The output gain of this RBF neural network is a weighted linear superposition of all these basis functions. The basic model of RBF neural network is shown in Fig. 5.

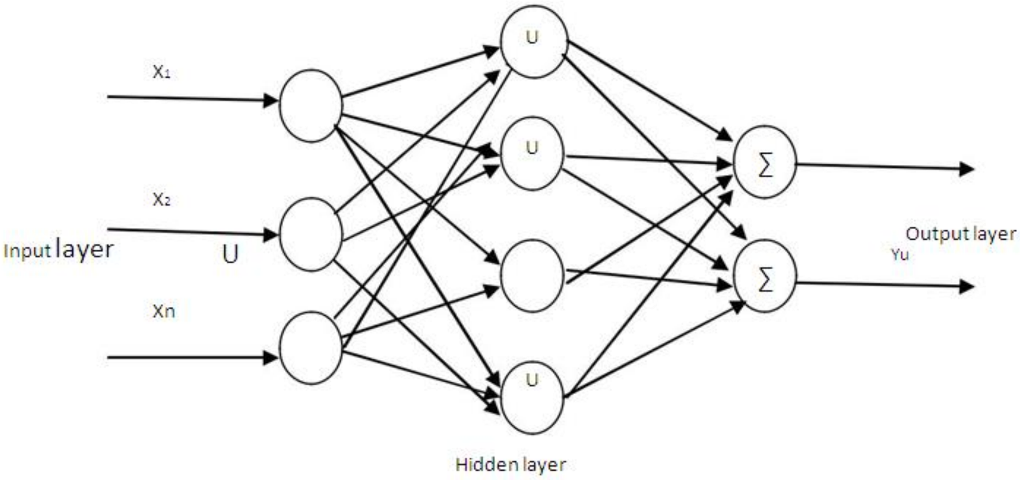

Fig. 5: RBF network

In this network, weights for input-hidden layer interconnections are fixed, while the weights for hidden-output layer interconnections are trainable. Following input - output mapping function 1 as:

$Y(X) = \sum Wi\ U(\|X\text{-}Xi\|)$
Where,
U () = basis function of hidden layer which is applied at each neuron of hidden layer,
M = basis functions consisting of the Euclidean distance between applied inputs X,
Y(x) = output mapping function.

In 2013 Mussab M. Hassan et al. [9] uses hybrid statistical traffic classifier to classify the P2P (peer to peer) traffic. Here also the works in two steps, firstly offline heuristics learning corpus generation and second is online statistical classification, In this first part, Heuristic classify the traffic flow and second part machine learning algorithm are used to classify network traffic. They apply 64 ML algorithms to classify traffic and find that RBF ML algorithms give good result.

III. COMPARISION OF CLASSIFICATION METHOD

Table 1: Performance of various methods

Reference no. research paper	Year	Classification Method used	Feature selection algorithm	Classification Accuracy	Dataset
7	2013	C4.5 Decision Tree Algorithm	N/A	90%	UTM campus network
9	2013	RBF, Bayes Net and C4.5	N/A	76.67%	Proprietary Hand Classified Traces
6	2013	MLP,RBF,C4.5,Bayes Net and Naïve Bayes	Correlation and consistency based feature selection algorithm	91.87%	Proprietary Hand Classified Traces
1	2015	K-means and EM	CSF	88%	Proprietary Hand Classified Traces
4	2014	Heuristic based co-clustering algorithm	N/A	86%	large scale Wi-Fi ISP N/W
2	2015	Decision Table, oneR, JRip, ZeroR and PART	N/A	Above 90%	KDD-CUP
3	2014	Variation edited nearest Neighbour (VNNN)	N/A	86.90%	UNBIS and Cambridge university

IV. ADVANTAGES AND DISADVANTAGES

Table 2: Advantages and Disadvantages of Different types of Approaches

Classification Method	Advantages	Disadvantages
Supervised ML Techniques		
Naïve bayes classifier	• Easy to implement. • We are getting good results in most of the cases	• Assumption of class conditional independence • Dependencies among classes cannot be modeled by Naive Bayesian Classifier

C4.5 and C5.0	• Easy to implement • We Can use it with both values categorical and continuous • It can Deal with noise	• Small variation in data can lead to different decision trees • Does not work very well on a small training set.
RBF	• We use enough number of nodes to find high accuracy. • Simple layer structure.	• Training time is very long and it increases when we increase the numbers of node •
Bayesian Net Classifier	• Implementation is very complicated	• Processing efficiency is high

V. CONCLUSION

This paper surveys significant works in the field of traffic classification during the peak period of 2012 to early 2015. Motivated by a desire to move away from port-based or payload-based traffic classification, it is clear that Machine Learning can be applied well in the task. The use of a number of different Machine Learning algorithms for offline analysis, such as Auto Class, Expectation Maximisation, Decision Tree, Naive Bayes etc. has demonstrated high accuracy (up to 99%) for a various range of Internet applications traffic. Early Machine Learning techniques relied on static, offline analysis of previously captured traffic. More recent work is beginning to address the requirements for practical, Machine Learning-based real-time IP traffic have outlined a number of critical operational requirements for real-time classifiers and qualitatively critiqued the reviewed works against these requirements.

There is still a lot of room for further research in the field. While most of the approaches build their classification models based on sample data collected at certain points of the Internet, those models' usability needs to be carefully evaluated. The accuracy evaluated on the test dataset collected at the same point of measurement might not be true when being applied in different point of measurement.

REFERENCES

1. Hardeep Singh. Performance Analysis of Unsupervised Machine Learning Techniques for Network Traffic Classification ; fifth International Conference on Advanced Computing and Communication Technology , IEEE; 2015 ; p. 401-404.
2. Kailas Elekar, M. M. Waghmare and Amrit Priyadarshi. Use of rule base data mining algorithm for Intrusion Detection; International Conference on Pervasive Computing (ICPC), IEEE; 2015; p. 1-5.
3. Wang Rue-yu, LIU Zhen and Zhang Ling. Method of data cleaning for network traffic classification; the journal of Chine Universities of post and Telecommunication, Elsevier; 2014; p. 35-45.
4. Wei Lu and Ling Xue. A Heuristic-Based Co-clustering Algorithm for the Internet Traffic Classification; 28th International conference on Advanced Information Networking and Application Workshop, IEEE; 2014; p. 49-54.
5. Yibo Xue, Dawei Wang and Luoshi Zhang. Traffic Classification: Issues and Challenges; International conference on computing, networking and communication (ICNC); IEEE, 2013; p. 545-549.
6. Kuldeep Singh, S. Agrawal and B. S. Sohi. A Near Real-time IP Traffic Classification Using Machine Learning; International Journal of Intelligent Systems and Applications(IJISA); 2013;vol. 5; p 83-93.
7. Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhao and Yong Xiang. Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions; IEEE transactions on information forensics and security; vol. 8,;2013;p. 5-15.
8. Shezad Shaikh, ashphak P. Khan and Vinod S. Mahajan. Implementation of DBSCAN Algorithm for Internet Traffic Classification; International Journal of Computer Science and Information Technology Research (IJCSITR); 2013; p. 25-32.
9. Mussab M. Hassan and Muhammad N. Marsono. A Hybrid Heuristics-Statistical Peer-to-peer Traffic Classifier; International conference on computer system and industrial information (ICCSII); 2013; IEEE; p. 1-6.

10. Megha Aggarwal. Performance analysis of different feature selection Method in Intrusion Detection; international Journal of Scientific and Technology Research(IJSTR);2013.
11. Anugerah Galang Persada, Noor Akhmad Setiawan and Hanung Adi Nugroho. Comparative study of Attribute Reduction on Arrhythmia Classification Dataset; international conference on Information Technology and Electrical Engineering(ICITEE);2013;IEEE;p.68-72.
12. Harvinder Chauhan, 2Anu Chauhan. Implementation of decision tree algorithm C4.5, International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013.
13. Jaspreet Kaur, Sunil Agrawal and B. S. Sohi. Internet Traffic Classification for Education Institutions Using Machine Learning; International Journal of Intelligent Systems and Applications (IJISA); 2012; MECS; vol. 4; p. 37-45.