Feature Selection Technique Impact for Internet Traffic Classification Using Naïve Bayesian

[1]Satish Chadokar, [2]Ashish Kumbhare

Department of Computer Science and Engineering,
RGPV, Bhopal, India

Abstract— **Feature selection technique has an important role for internet traffic classification. This technique will present more accurate data and more accurate internet traffic classification which will provide precise information for bandwidth optimization. One of the important considerations in the feature selection technique that should be looked into is how to choose the right features which can deliver better and more precise results for the classification process. This research will compare feature selection algorithms where the Internet traffic has the same correlation that could fit into the same class. Internet traffic dataset will be collected, formatted, classified and analyzed using Naïve Bayesian. Formerly, the Correlation Feature Selection (CFS) is used in the feature selection to find a collection of the best sub-sets data from the existing data but without the discriminate and principal of a body dataset. We plan to use Principal Component Analysis technique in order to find discriminate and principal feature for internet traffic classification. Moreover, this paper also studied the process to fit the features. The result also shows that the internet traffic classification using Naïve Bayesian and Correlation Feature Selection (CFS) have more than 90% accuracy while the classification accuracy reached 75% for feature selection using Principal Component Analysis (PCA).**

Index Terms—**ISP, DAG, CFS, FCBF, BOF, PCA**

Introduction: 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. It can be used for intrusion detection system by finding patterns of denial of service (Dos) and other attacks.

It can be used for intrusion detection system by finding patterns of denial of service (Dos) and other attacks. It can help to ISPs to monitor network traffic flow and troubleshoot the faults and other problems, it can also be used in "lawful inspection" of the payload of a packet by government to obtain users information.

2. Problem Definition

The traffic on internet network is growing exponentially and today traffic reaches up to TB/S. Now, there are several new applications like online gaming, social network sites, online chatting, live video streaming etc use dynamic port numbers or unregistered port number so port based techniques can cot classify traffic correctly this problem overcomes using payload based techniques overcomes this problem but it cannot deal with encrypted data they uses deep packet inspection and because of deep packet inspection the processing efficiency become very low.

There are three main goals of internet traffic classification are Internet traffic classification,

- **Accuracy**- correctly classified traffic i.e. It is the percentage of correctly classified instances over all classified samples this called classification accuracy
- **Processing Efficiency**- the speed of classification. How much it takes to classify internet traffic.
- **New Knowledge Adaption**-to deal with new application and data and knowledge i.e. nature of techniques will be adaptive.

It is play most important role in intrusion detection system because with the help of accurately classify traffic we find the pattern and information related to data packets. Machine learning techniques can help us to find all three goals and also overcome the problems of previous techniques. In this thesis we work on machine learning based internet traffic classification with datasets which having large number of samples.

3. Proposed Methodology

In our proposed methodology first we take the well known dataset of Cambridge University which have 249 features of each packet, the features are like source IP address, destination IP address, port number, time interval between send packet and received packet etc. We perform feature selection first because feature is a very important task and eliminate redundant and irrelevant features. It is also called feature optimization from dataset and then performs classification using four well known machine learning approaches.

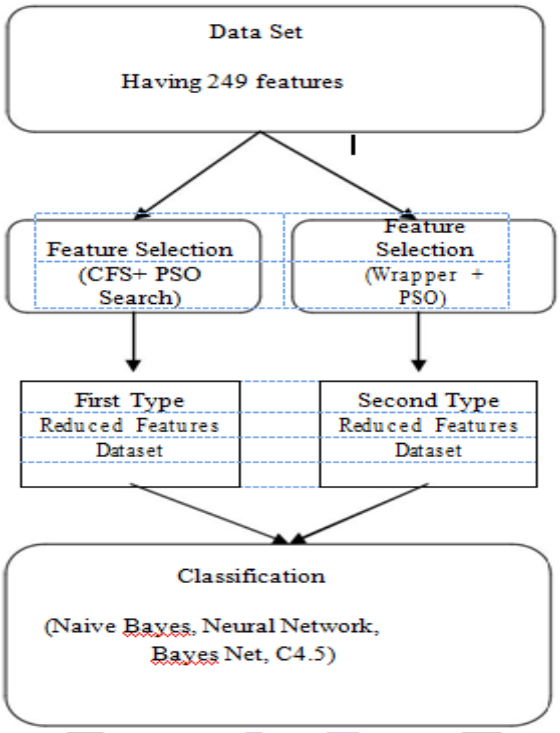

Figure 1 Proposed Methodology

4. Experimental Matrices

In this research work we use weka toolkit to perform various experiments of classification. Weka is basically a data mining tool and specially used for classification purpose. In this paper we consider confusion matrix to find overall performance and get model building time, in this we find the real class of a packet where it is belong to.

In this experiment after perform operation of classification it generate a confusion matrix for 12 classes these are WWW, MAIL, FTP-Control, FTP-PASV,ATTACK, P2P, DATABASE, FTP-DATA, MULTIMEDIA, SERVICE, INTERACTIVE and GAMES on the basis of parameters which is given blow.

- False Negatives (FN): Percentage of samples of class P incorrectly classified as not belonging to class P.
- False Positives (FP): Percentage of samples of other classes incorrectly classified as belonging to class P.
- True Positives (TP): Percentage of samples of class P correctly classified as belonging to class P.
- True Negatives (TN): Percentage of samples of other classes correctly classified as not belonging to class P.

Overall accuracy: The percentage of samples that are correctly classified or packets belongs to its real class called classification accuracy. To find classification accuracy we use formula:

$$\text{Overall Accuracy} = \frac{\sum_{k=0}^{n} TP}{\sum_{k=0}^{n} (TP + FP)}$$

Overall accuracy we count in terms of percentage (%) i.e. how many percentage of instances correctly classified.

Model Building Time: time taken to build a model after training and then performed testing on the basis of this model. Building time counts in seconds.

4.1. Classification

Now we perform classification using four well known machine learning approaches on two types of reduced set dataset. First is prepared by correlation based feature selection algorithm with PSO search technique and second is prepared by wrapper subset evaluator with PSO search.

4.1.1. Naive Bayes

Naive Bayes is a probabilistic model and in which each feature is independent from others. In this experiment naive bayes perform well with second type reduced dataset compare than first type in terms of classification accuracy, it gives approx. 95% classification accuracy with second type dataset and its 4-5% more compare than first type dataset but the model building time of second type is little higher than first type.

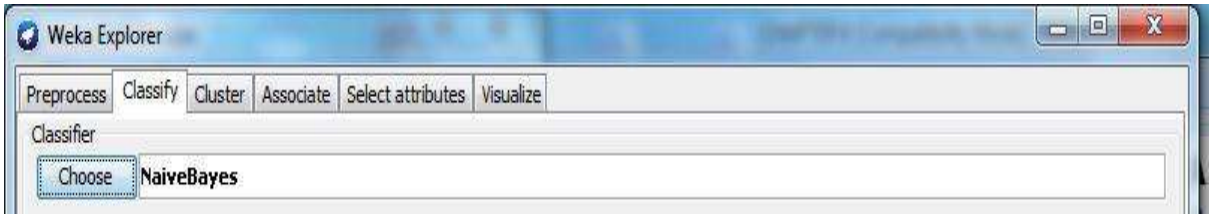

Figure 2: Naive Bayes Classifier

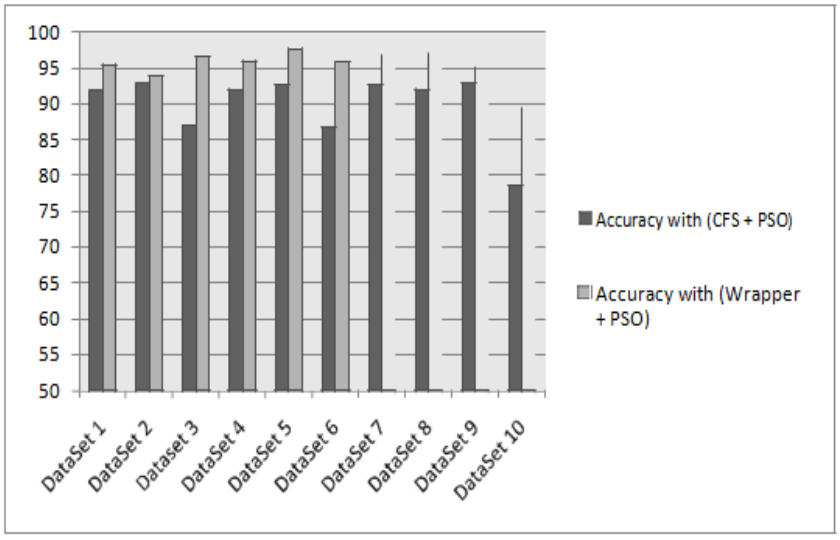

Figure 3: Classification Accuracy (%) of Naive Bayes Classifier on both types of reduced datasets

This table and graph shows the classification accuracy of each dataset in terms of percentage with both type reduced datasets. here highest classification accuracy 97.72 % gain by naïve bayes on dataset 5 in second type reduced dataset and smallest classification accuracy 78.52 % on dataset 10 in first type of reduced dataset.

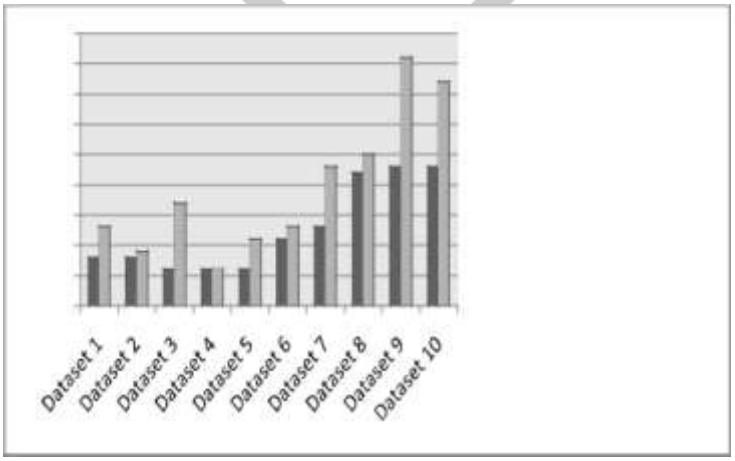

Figure 4: Model Building Time (Seconds) of Naive Bayes Classifier on both types of reduced datasets

This graph and table shows model building time of naive bayes classifier in terms of seconds and it shows the highest time 0.41 seconds taken on dataset 9 in second type reduced dataset and smallest time 0.06 seconds taken on datasets 3,4,5 in first type reduced dataset. Here dataset 9 is having highest number of instances and dataset 5 is having smallest number of instances.

4.1.2. Feed Forward Neural Network

Feed forward neural network is simplest type of neural network. In this experiment we use one hidden layer with three neurons. we use linear summation at output layer.

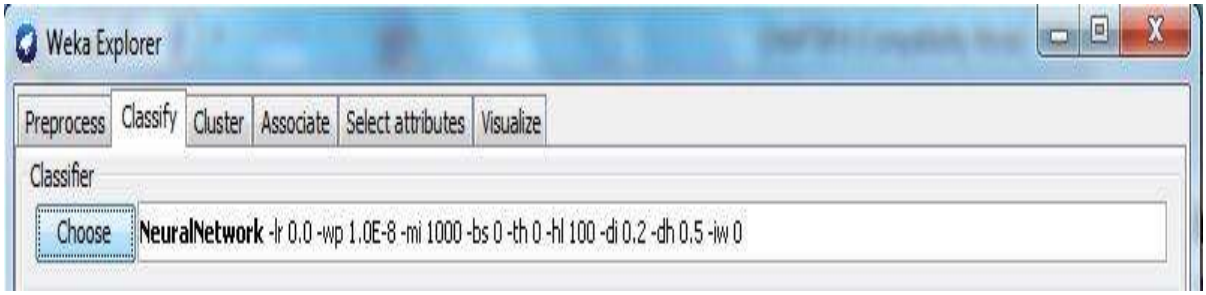

Figure 5 Feed Forward Neural Network

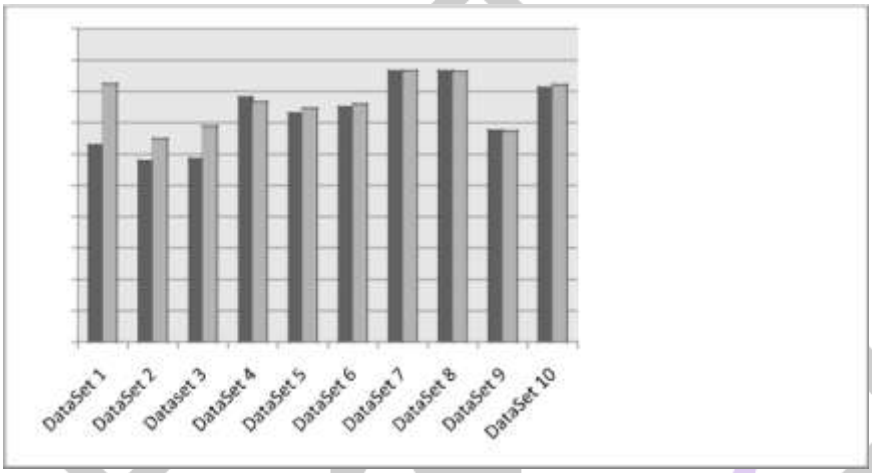

Figure 6: Classification Accuracy (%) of Feed Forward Neural Network Classifier on both types of reduced datasets

This table shows that feed forward neural network is not very effective with both type of reduced datasets it gives the highest classification accuracy 93 % classification accuracy and get blow 90% with maximum datasets. Performance of feed forward neural network is very poor compare than other classification approaches.

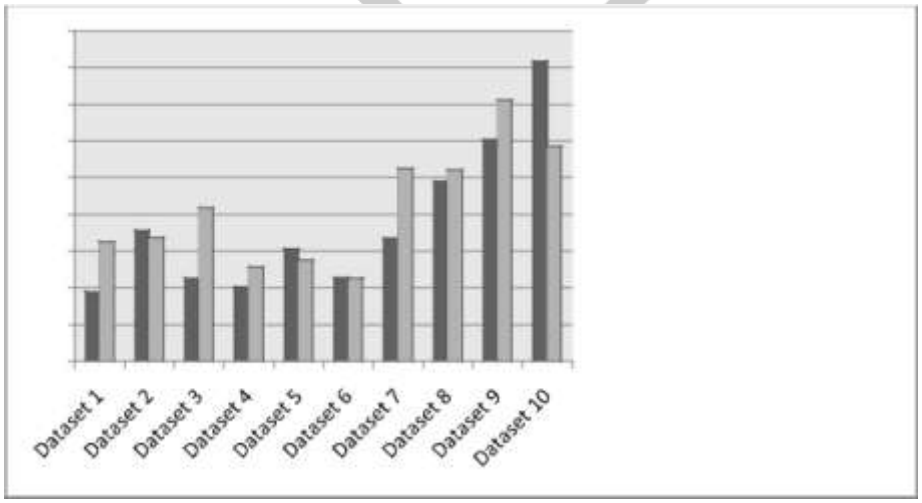

Figure 7: Model Building Time (Seconds) of Feed Forward Neural Network Classifier on both types of reduced datasets

Model building time taken by feed forward neural network is very high in this experiment it take 16.25 seconds highest on dataset 10 and in maximum cases it takes more than 5 seconds so its processing efficiency is also very poor.

4.1.3. Bayes Net

Bayes Net is belief network and uses a DAG (Directed Acyclic Graph) between features. In this paper we use K2 search algorithm and Simple estimator to estimate table.

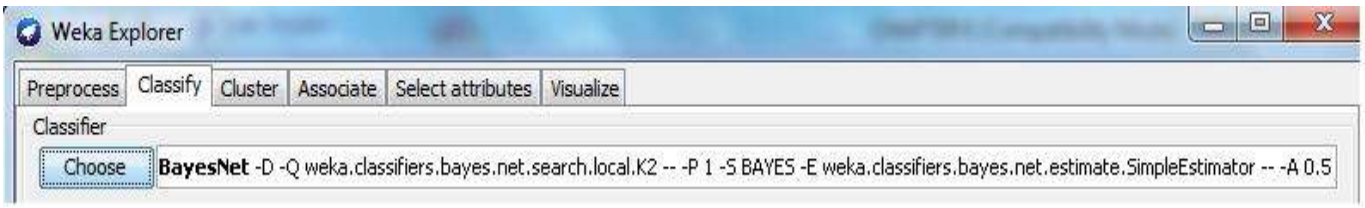

Figure 8: bayes Net Classifier

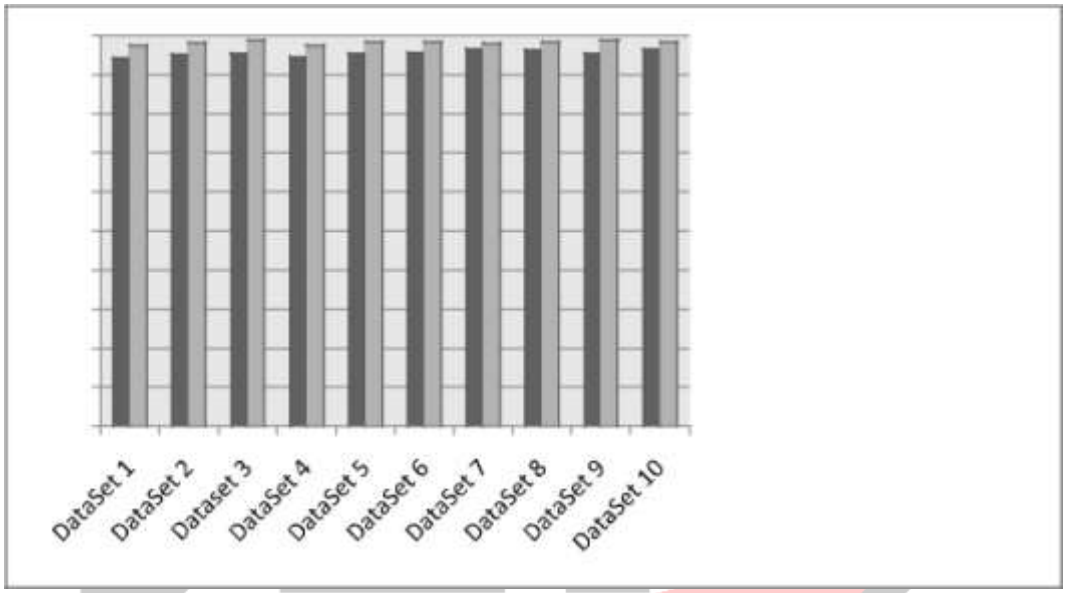

Figure 9: Classification Accuracy (%) of Bayes Net Classifier on both types of reduced datasets

The table and graph shows that it is good classifier. It is gives a high percentage of classification accuracy. We get above 95% correctly classified instances in all datasets of both types. Its results with second type reduced dataset are little good compare than results of first type. The highest percentage of classification accuracy 98% we get on in all 10 datasets of second type and get 95% accuracy with first type datasets.

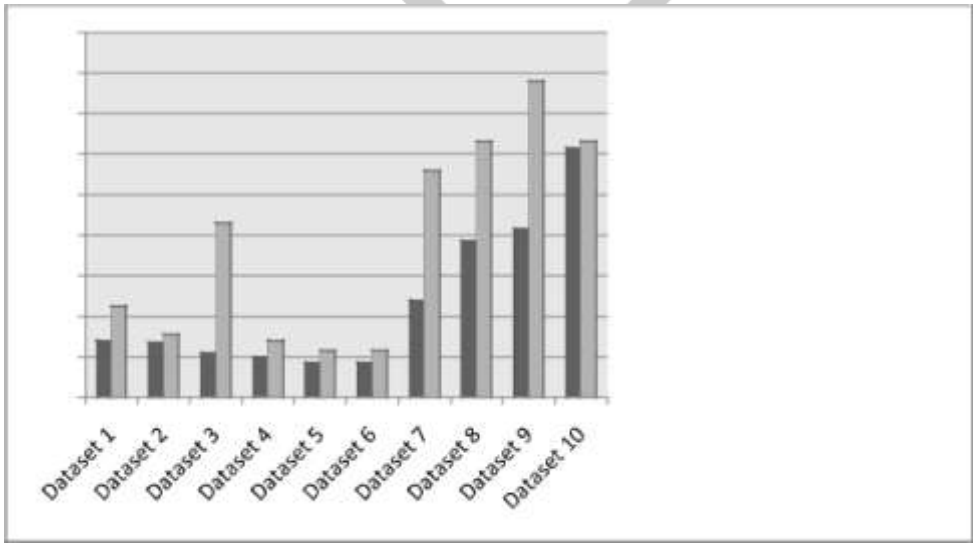

Figure 10: Model Building Time (Seconds) of Bayes Net Classifier on both types of reduced datasets

Its model building time is also small i.e. it has good processing efficiency. it does not have model building time more than 1.5 seconds. The time of second type reduced dataset is a little more compare with first type.

4.1.4. C4.5 (J48) Decision Tree

C4.5 is a decision tree based classification approach. C4.5 gives better results in all four classification approaches. We get highest percentage of classification accuracy. for this experiment the parameters used are Confidence factor = 0.25, Minimum number of instances per leaf = 2 , Number of folds to reduce error pruning = 3

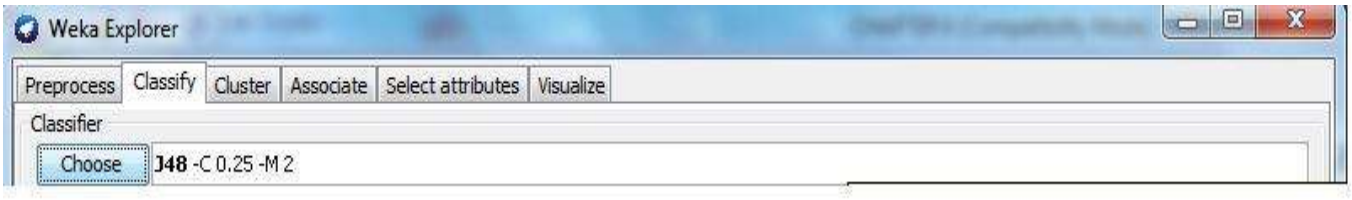

Figure 11: C4.5 Classifier

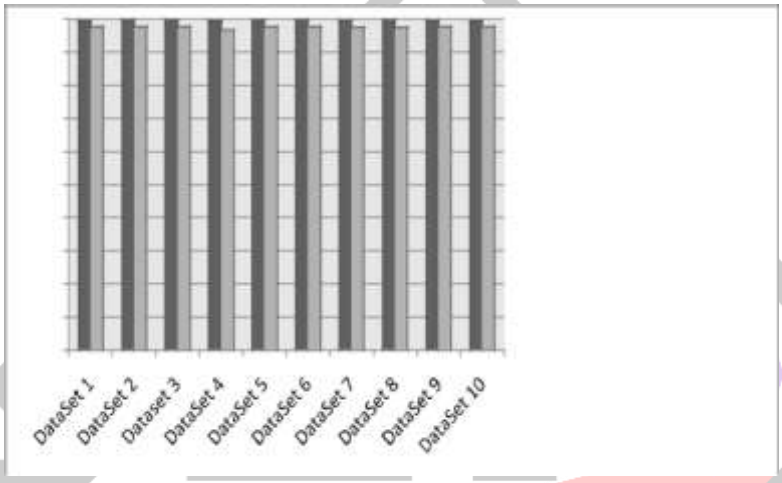

Figure 12 : Classification Accuracy (%) of Bayes Net Classifier on both types of reduced datasets

The results show that C4.5 decision tree is a very good approach for classification. C4.5 gives highest percentage no classification accuracy compare than other three approaches. We get above 99.5 % classification accuracy in all datasets of first type and also get above 97% classification accuracy in all datasets of second type.

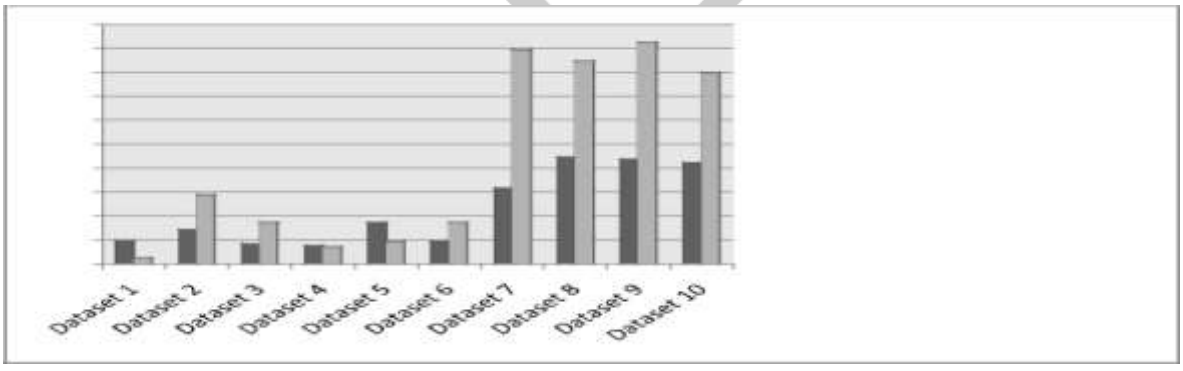

Figure 13: Model Building Time (Seconds) of C4.5 (J48) Classifier on both types of reduced data sets

The model building time of C4.5 decision tree approach is good. It takes less time and give good results. Highest time taken by C4.5 is 4.6 seconds in dataset 9 which is largest dataset in all datasets it is having highest number of packets. So the overall performance of C4.5 is very well in classification accuracy as well as processing efficiency.

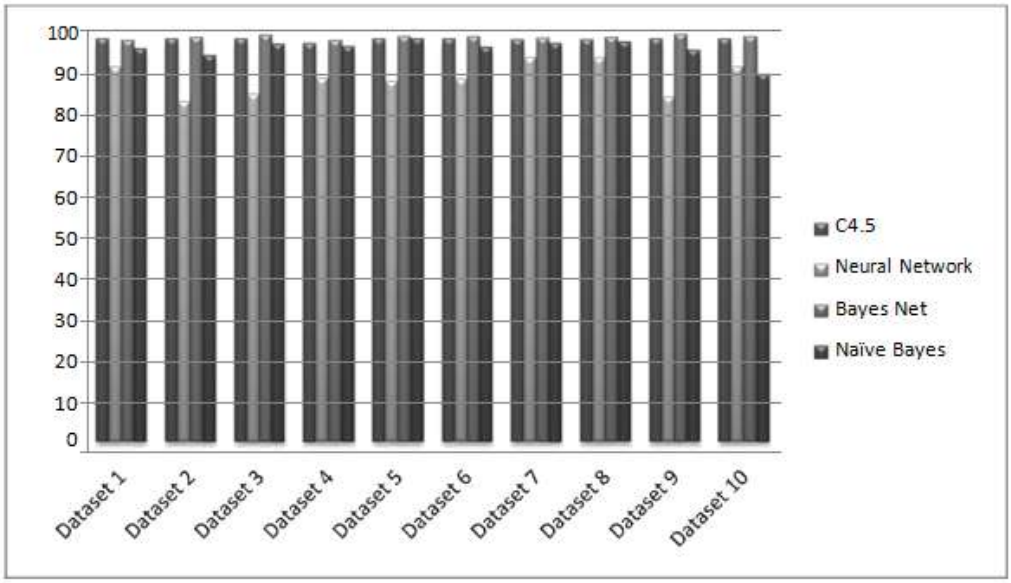

Figure 14: Comparatively study of four classifiers in terms of classification accuracy on first type reduced dataset

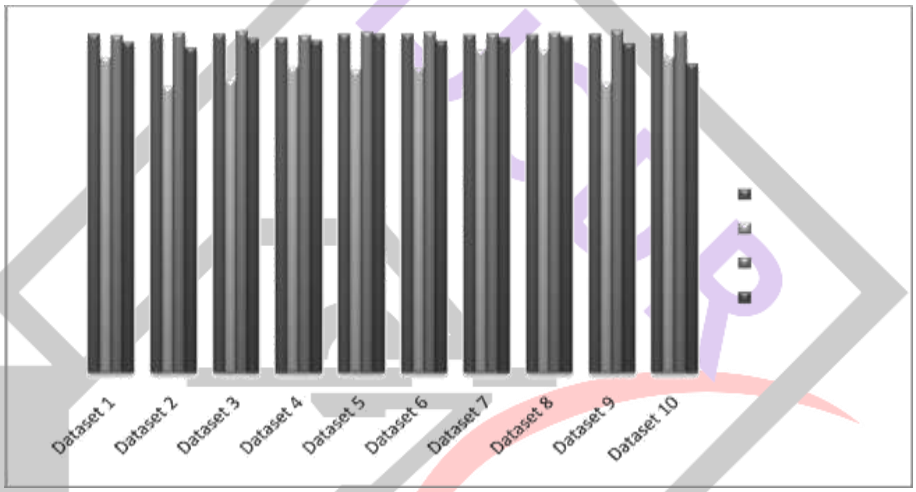

Figure 15: Comparatively study of four classifiers in terms of classification accuracy on second type reduced dataset

Table Number 6.7.9 and 6.7.10 give the overall classification accuracy of all four classifiers with both types of datasets and we find that in first type C4.5 gives highest classification accuracy and Feed forward neural network gives lowest classification accuracy. Bayes Net and Naive Byaes also give good classification accuracy but not so effective. In second type datasets Bayes Net gives the highest percentage of classification accuracy and feed forward neural network gives lowest classification accuracy. C4.5 and Naive Bayes also give good classification accuracy.

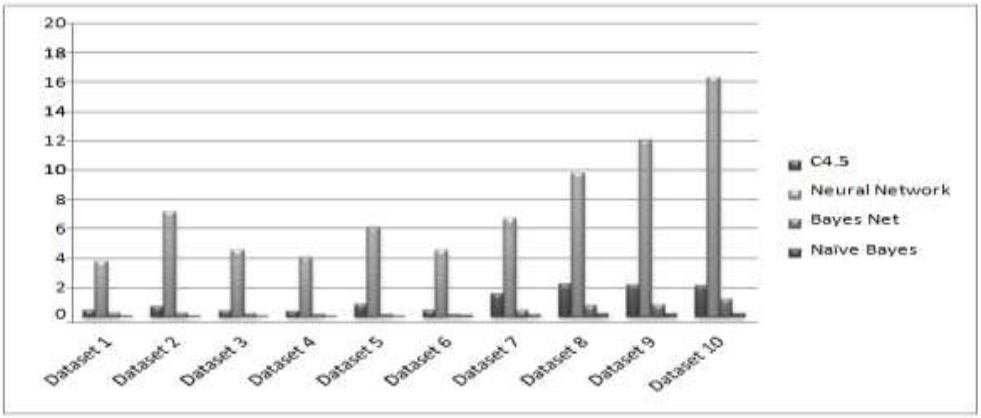

Figure 16: Comparatively study of four classifiers in terms of Model building time (Seconds) on first type reduced dataset

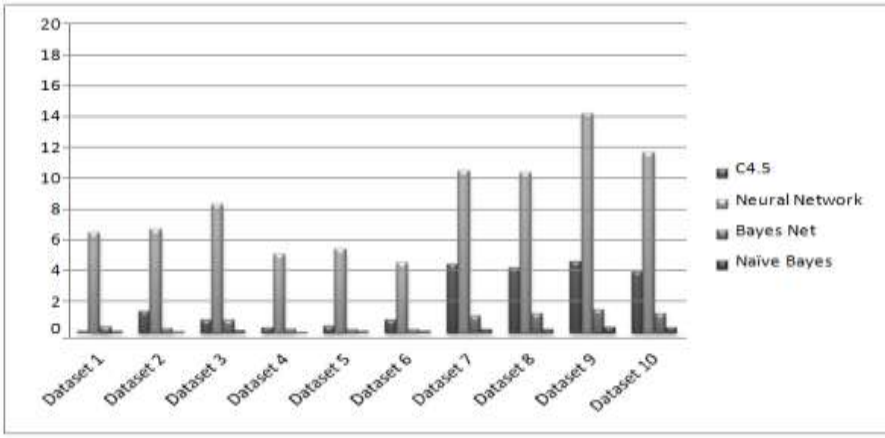

Figure 17: Comparatively study of four classifiers in terms of Model building time (Seconds) on Second type reduced datasets

Table Numbers 6.7.11 and 6.7.12 give model building time of all four classifiers with both two types of datasets and we find that with both datasets that Naive bayes gives highest processing efficiency and Feed forward neural network has lowest processing efficiency. We also observe that model building time increases with size of datasets in all classifiers.

4.2 Summary

Machine learning approaches give good classification results. In this chapter we do experiments and analyze the performance of different classifiers in terms of classification accuracy and model building time. for experiments we uses 10 different datasets having large number of features so firstly we apply two feature selection methods these are Correlation based and Wrapper subset evaluator with PSO search technique and prepared two datasets with selected features. and then applied four machine learning classifiers are Naive Bayes, Feed Forward Neural Network and Bayes Net to classify on these two types of datasets. All operation were performed on weka tool and use 10 fold cross validation. Here we find that C4.5 decision tree approach gives highest percentage of classification accuracy and Naive Bayes approach gives highest processing efficiency but it get low percentage of classification accuracy.

Conclusion

Internet traffic classification is a very hot research area with wide range of applications. By using internet traffic classification we can solve many problems related to network and also examine activities of it. The traditional techniques of internet traffic classification are Port based and Payload techniques but port based techniques suffers to deal with new application and unknown knowledge because they use random or dynamic port number address which is not registered with IANA. And payload based techniques able to solve this but it suffers from encrypted data i.e. it cannot deal with encrypted information because it uses deep packet inspection. In this research work we classify internet traffic using machine learning, machine learning techniques overcome these problem of both. Machine learning techniques give superiority over the traditional techniques.

In this research work we are applied four well known supervised machine learning techniques to classify internet traffic and these techniques are Naive Bayes, Feed Forward Neural Network, Bayes Net and C4.5 decision tree. There are three phases of whole process which are first phase is to prepare datasets, second is feature selection (to reduced irrelevant and redundant features) and then last classification. For this work we take more datasets and use two different techniques for feature selection, first correlation based feature subset selection (CFS) with PSO search and second wrapper feature subset selection with PSO search. In our research we uses 10 fold cross validation in all experiments and uses weka tool kit to perform operations. We Find that CFS gives good set of features in comparison of wrapper subset feature selection and C4.5 decision tree approach gives higher percentage of classification accuracy compare than other three, it gives above 99.5% classification accuracy with first type reduces features datasets and its performance also very good with second type reduced feature datasets in it we get 97% accuracy. Naive bayes gives higher processing efficiency in all approach with both type reduced feature datasets but it classification accuracy is poor. Bayes Net is a good classifier it gives high percentage of accuracy as well as processing efficiency and it also give higher classification accuracy with second type reduced datasets it give 98% classification accuracy. The performance of feed forward neural network is poor in both parameters (accuracy and model building time).

ACKNOWLEDGMENT

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.