INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET) E-ISSN: 2395-0056
VOLUME: 07, SPECIAL ISSUE | JUNE 2020 WWW.IRJET.NET P-ISSN: 2395-0072
International Conference on Recent Trends in Science & Technology-2020 (ICRTST - 2020)
Organised by: ATME College of Engineering, Mysuru, INDIA

Classification of Internet Traffic Data using Feedforward Network: An Approach using Reduced Number of Features

N Manju[1], B S Harish[2]

[1]*Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, India*
[2]*Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, India*

Abstract— Analysis of internet traffic is a recent research trend in the area of networking and communication. The main aim of the data analysis is to perform network management effectively. Internet traffic classification poses a major role in most of the management tasks. Traffic classification has become a challenging issue due to the drawback of both port based and payload based classification approach. Hence, in this paper, we are using flow based statistical approach to classify the traffic which overcomes the drawback of port and payload based approach. In this paper a) we used a feedforward neural network (FNN) to handle the issue of multiclass imbalance problem and b) evaluated the model using less (08 out of 248 features) number of features which are selected from ensemble tree based feature selection method. The proposed model is evaluated using Cambridge dataset. Experimentation is carried out on two variants of datasets: highly imbalanced (standard) and normalized (derived) datasets. Result shows the 99% of accuracy for highly imbalance dataset and 96% of the accuracy for the normalized dataset using 248 features. This in turn getting almost comparable result using fewer features when compared to all 248 features.

Keywords-Internet Traffic; Feedforward Network; Feature Selection; Classification.

I. INTRODUCTION

Now a day, more and more business and other applications are promoted over the Internet [1]. It is growing like an essential part of the current applications which relies on the internet [2]. As an outcome, monitoring and controlling an internet traffic has attracted from the past few years. As part of security concern, malicious traffic identification and avoiding attackers will help in controlling the security. On the other hand, from the perspective of Quality of Service (QoS), accurate classification of various categories of internet traffic helps in determining the types of applications and effective usage of resources. In addition, network service providers can trace the increase in the different applications. Hence, internet traffic classification helps in network provision to hold different needs of user community [3]. There are many machine learning methods applied in the literature to classify the internet traffic. From the previous work, we found that, machine learning methods gave satisfactory results. Internet traffic classification is necessary to solve network management problem for Internet Service Providers (ISP). Further, it is a major part of automated system which is capable of detecting intrusions [4, 5], allocation of resources based on demand, identifying the patterns of Denial of Service [6], lawful interception [7] etc.

The rest of this paper is formulated as follows: Section II illustrates previous works carried out on Internet traffic. Section III describes feature selection and classification approach. Section IV discusses the experimental results and finally, the conclusion is drawn in section V.

II. LITERATURE SURVEY

Web is a buzzword due to increase in the usage of internet to get various types of services. Services may include in the academic, business, entertainment, social media etc. There is an always demand in the usage of internet in the computer network. Hence, internet traffic management requires prioritization of various applications and delivery of Quality of Service (QoS). With this concern, various machine learning approaches have been used in the literature. Researchers have used algorithms viz., Naive Bayes, C4.5, Naive Bayes Tree (NBT) and Bayesian Network in [8]. Outcome of this experiments shows that C4.5 achieves quick classification. Further, Naive Bayes Kernel is slower compared to rest of the algorithms. The average classification accuracy obtained is 95% considering all group of algorithms. In [9], along with Bayesian Networks and Decision Tree, Multilayer perceptrons also used to compare and measure the performance of an algorithm.

Overall outcome presents, the decision tree achieves satisfactory classification accuracy with good computing speed but multilayer perceptron is with less accuracy. Machine learning algorithms viz., Adaboost Genetic programming and C5.0 is used to recognize VOIP data traffic in the Skype applications [10]. Researchers have applied equal sampling by using random selection process and achieved a better accuracy. Multi Objective Evolutionary Fuzzy Classifiers is proposed in [11]. Proposed method is based on Fuzzy Rule Based Classifiers (FRBC) and achieved a satisfactory result. Identification of individual applications instead of group is proposed using J48, Random Forest, Bayes Net and KNN. These algorithms use complete 111 features of UNBISCX standard dataset [12]. The result obtained shows an accuracy of 93.44% using KNN for ISCX datasets while Random Forest (RF) achieved an accuracy of 90.87% for internal datasets. Second set of experimentation is conducted by reducing the number of set of features from 111 to 12 features with increase accuracy by 2% for the internal dataset. Extreme Learning Machine (ELM) approach is proposed for internet traffic classification. Kernel Based Extreme Learning Machine is tested on standard Cambridge dataset and implemented as an application based on genetic algorithm to choose required parameters. Based on the developed tool, an accuracy of 95% is achieved [13]. Multi-class imbalance is again another issue in the internet traffic classification. An attempt is made to evaluate the performance using cost-sensitive method based on Flow rate based Cost Matrix (FCM) and Weighted Cost Matrix (WCM) [14]. Result shows that, WCM performs better when compared with Flow rate based Cost Matrix (FCM) in terms of stability. Imbalanced Data Gravitation-based Classification (IDGC) based model is implemented to solve class imbalance problem using conventional algorithms and imbalanced algorithms for experimentation. The result shows that conventional classification models are not as effective as IDGC to classify imbalance internet traffic data. On the other hand, C4.5CS performs equally effective when compared to IDGC [15].

Feature selection method plays an important role in classification using machine learning approaches. Feature selection method is used to determine subset of significant features and discard redundant features from each of the feature subset. In machine learning, numerous methods are proposed to solve internet traffic classification problem. Extraction of real time feature subset is proposed in [16] and performance is evaluated using various machine learning algorithms based on decision tree algorithm. Hybrid approach is proposed based on discretization, filtering and classification methods [17]. The result shows that, hybrid method achieves better performance. Weighted Symmetrical Uncertainty Area Under ROC Curve which is also an hybrid approach is presented in [18]. Further, Selecting Robust and Stable Feature (SRSF) is applied to evaluate the internet traffic data for various datasets. Experimentation results show that the proposed method gives better result using C4.5 in terms of accuracy and speed. Selecting features based on rough set theory is proposed in [19]. This method minimizes the features from 6 to 10. Using these selected features; Bayesian network is used to evaluate the performance. Result shows an improvement in the classification accuracy.

III. PROPOSED METHODOLOGY

In this section, we are using a feedforward neural network (FNN) to classify the IP traffic. The main reason to use FNN is that, it has tremendous learning rate for the non-linear data [20]. Hence, we normalize data before inputting it to the network model. Normalization is achieved by using equation (1):

$$f(y) = (y - \mu) / \sigma \quad (1)$$

where, y refers features, μ refers mean of y and σ refers standard deviation of y.

Later, the new features fall in the range of [-1, 1]. The proposed method make use of feedforward neural network having four layers. First layer is input layer, the other two are hidden layers and the last layer is output layer with softmax activation function.

In Fig.1, the feedforward neural network architecture consists of 248 neurons assigned to the input. In hidden layer 1, we have 124 neurons which is half of the input layer. All the neurons of preceding layers are associated to all the neurons of next layer via the dot product of weights given in equation (2).

$$f(x) = wx + b \quad (2)$$

where, w refers weights of the current layer, x refers output of the previous layer and b refers bias.

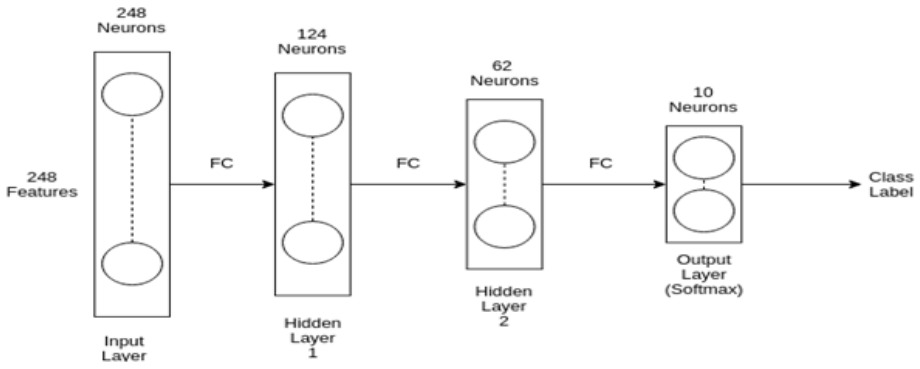

Fig.1: Proposed Feedforward Neural Network with 248 features. (FC: Fully Connected)

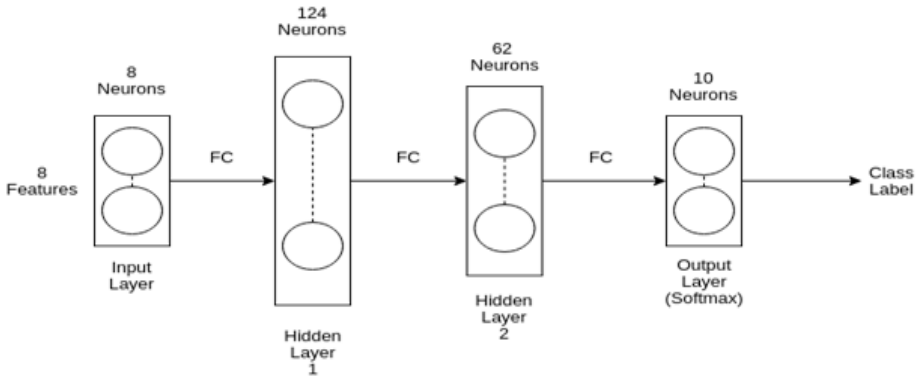

Fig. 2: Proposed Feedforward Neural Network with 08 features. (FC: Fully Connected)

The feedforward neural network uses Rectified Linear Unit (ReLU) presented in equation (3) as the activation function [21] that suppress the negative input value to zero and retains positive input value constant.

$$f(x) = \max(0, x) \quad (3)$$

where, x refers to input

In the same way, we have 62 neurons in the second hidden layer which is half of the previous hidden layer1 and the procedure is same as first hidden layer. The last layer is the output which consists of 10 neurons for 10 categories which has to be classified. Last layer uses softmax, which is a loss function and is given by

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_i}} \quad (4)$$

where, y_i refers to score of class j and e refers to standard mathematical constant $\simeq 2.718$

Back propagation is used over neural network that updates weights through adam optimizer [22]. Fig. 2 consists of 8 neurons in the input layer as we are using 08 selected features using ensemble feature selection method based on tree based approach. The rest of the process is same as shown in Fig. 1.

IV. EXPERIMENTAL RESULTS

A. *Dataset*

The dataset used in experimentation is Cambridge university dataset which consists of 248 features. We have used two types of dataset consisting of 10 categories as shown in table I. Dataset01 is highly class imbalance and Dataset02 is derived dataset with less class imbalance. In dataset01, instances are more biased with regard to WWW class and less biased with regard to MULTIMEDIA class. On the other hand, dataset02 is a normalized derived dataset.

TABLE I. PROPERTIES OF DATASETS

Class No.	Class Name	Standard (Dataset01)	Derived (Dataset02)
		Flow Count	Flow Count
1	ATTACK	122	1793
2	DATABASE	238	2648
3	FTP-CONTROL	149	3000
4	FTP-DATA	1319	3000
5	FTP-PASV	43	2688
6	MAIL	4146	3000
7	MULTIMEDIA	87	576
8	P2P	339	2094
9	SERVICES	206	2099
10	WWW	18211	3000
Total Number of Flows		**24860**	**23898**

B. *Feature Ranking and Selection*

The XGBoost tree based feature selection approach assigns weights to each of the feature to obtain the feature ranking. Further, each of the features is evaluated using XGBoost model to select features which gives more accuracy. From the experimentation, we have chosen 08 features out of total 248 features as shown fig. 3 and fig. 4. The detailed explanation can be found in our proposed work in [23].

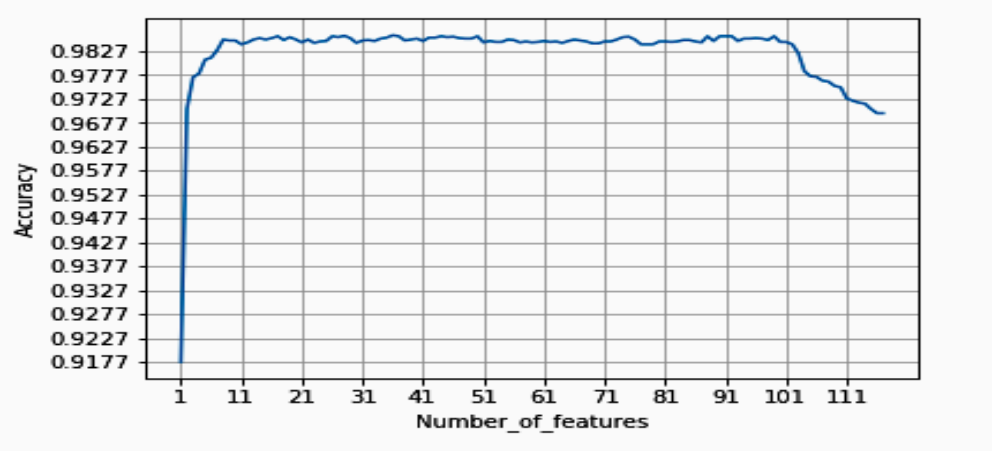

Fig. 3. Accuracy of first 117 features

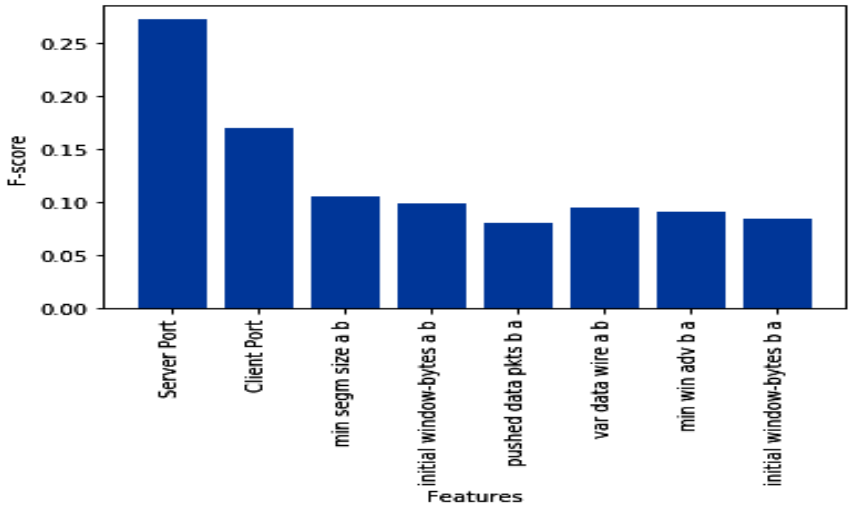

Fig. 4. Features selected based on highest accuracy derived from fig 3.

C. Results and Discussion

Following sections shows the result for two set of features: 248 and 08 features. Main aim of this work is to overcome the issue of class imbalance problem. The proposed model gives satisfactory result by avoiding bias. Further, it also gives better result for dataset02.

1) Results on Dataset01 (Standard)

TABLE II. PRECISION, RECALL AND F_1 SCORE FOR ALL THE CLASSES OF DATASET01

Class	Using 248 Features			Using 08 Features		
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score
1	0.90	0.50	0.64	0.64	0.47	0.54
2	1.00	0.97	0.98	0.97	0.93	0.95
3	1.00	0.97	0.98	0.95	0.93	0.94
4	0.99	0.99	0.99	0.99	0.99	0.99
5	0.75	0.81	0.78	0.88	0.72	0.80
6	0.99	0.99	0.99	0.99	0.99	0.99
7	0.53	0.54	0.53	0.61	0.70	0.65
8	0.87	0.81	0.84	0.86	0.81	0.84
9	1.00	1.00	1.00	1.00	1.00	1.00
10	0.99	0.99	0.99	0.99	0.99	0.99
Average	**0.90**	**0.86**	**0.87**	**0.89**	**0.85**	**0.87**

We obtained an accuracy of 99% for class imbalanced dataset using 248 features. Class 1 belongs to ATTACK category which is almost 50% of classification accuracy compared to remaining 9 classes. Further observation is that, average precision is 0.90% recall 0.86 % and F_1 score is 0.87% is presented in table 3. Hence, the model is successful in handling the high imbalance data. Further, we can observe that 08 features which are selected using XGboost tree based feature selection method gives competitive and comparable accuracy of 98.78%. In addition, average Precision, Recall and F_1 score is 0.89, 0.85 and 0.87 is achieved respectively. Hence, 08 features gives approximately same result which is obtained using 248 features as shown in table II.

2) Results on Dataset02 (Derived)

TABLE III. PRECISION, RECALL AND F_1 SCORE FOR ALL THE CLASSES OF DATASET02

Class	Using 248 Features			Using 08 Features		
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score
1	0.93	0.80	0.86	0.92	0.80	0.86
2	0.98	0.99	0.99	0.99	1.00	0.99
3	0.96	0.99	0.98	0.97	0.98	0.98
4	0.98	0.99	0.99	0.99	0.99	0.99
5	0.99	0.99	0.99	0.99	0.99	0.99
6	0.98	0.95	0.97	0.98	0.97	0.97
7	0.91	0.77	0.84	0.96	0.76	0.85
8	0.90	0.94	0.92	0.90	0.95	0.93
9	0.98	0.98	0.98	0.988	0.99	0.98
10	0.88	0.94	0.91	0.88	0.94	0.91
Average	**0.95**	**0.93**	**0.94**	**0.96**	**0.94**	**0.94**

Experimentation is carried out on the derived dataset (Dataset02) which is normalized dataset. From the experimentation, we have achieved an accuracy of 96.08% using 248 features. We also note that, average Precision, Recall and F_1 score are 0.95, 0.93 and 0.94 is achieved respectively. Further, using only 08 features, we have achieved an accuracy of 96.30% which is higher compared to 248 features. The higher Precision, Recall and F_1 is achieved with an average score of 0.96, 0.94 and 0.94 respectively as presented in table III. Therefore, the proposed method performs better even to the less imbalance data.

TABLE IV. OVERALL ACCURACY ON DATASET01 AND DATASET02

Sl. No.	Dataset	248 Features	08 Features
		Accuracy in %	Accuracy in %
1	Dataset01	99.00	98.78
2	Dataset02	96.08	96.30

Table IV presents the comparative performance for both the dataset using 248 and 08 features. From the experimentation, it is clear that, the feedforward neural network performs better using only 08 features in both the datasets.

V. CONCLUSION

Efficient and effective management of network is most important through network traffic analysis. Analysis helps in traffic classification to provide Quality of Service (QoS). The major issue with the multi class imbalance problem is to handle biasing. At the same time, reducing the dimensions to cope up with available limited computational resource is also most important. Keeping these issues in mind, in this paper we used a Feedforward Neural Network model and experimentation is carried using 248 and the reduced 08 features. The feedforward neural network gives satisfactory results on both the datasets. On the other hand, feedforward neural network method achieves comparable result using only 08 features which are selected from our previous work. In future, we are planning to work on large volume of traffic data over real time networks.

References

[1] Xu, L., "Advances in Intelligent Information Processing", Expert Systems, vol. 23, no. 5, pp. 249-250, 2006.

[2] Li, W., Zheng, W. B., and Guan, X. H., "Application controlled caching for web servers", Enterprise Information Systems, vol. 1, no. 2, pp. 161-175, 2007

[3] Sperotto, A., Schaffrath, G., Sadre, R., Morariu, C., Pras, A., and Stiller, B., "An overview of IP flow-based intrusion detection", IEEE communications surveys & tutorials, vol. 12, no. 3, pp. 343-356, 2010.

[4] Roesch, M., "Snort: Lightweight intrusion detection for networks", In Lisa, vol. 99, no. 1, pp. 229-238, 1999.

[5] Paxson, V., "Bro: a system for detecting network intruders in real-time", Computer networks, vol. 31, no (23-24), pp. 2435-2463, 1999.

[6] Stewart, L., Armitage, G., Branch, P., and Zander, S., "An architecture for automated network control of QoS over consumer broadband links", In TENCON, 10th IEEE Conference, 2005.

[7] Baker, F., Foster, B., and Sharp, C., "Cisco architecture for lawful intercept in IP networks", (No. RFC 3924), 2004.

[8] Williams, N., Zander, S., & Armitage, G., "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification", ACM SIGCOMM Computer Communication Review, Vol. 36, No. 5, pp. 5-16, 2006.

[9] Soysal, M., and Schmidt, E. G., "Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison", Performance Evaluation, Vol. 67, No. 6, pp. 451-467, 2010.

[10] Alshammari, R., and Zincir-Heywood, A. N., "Identification of VoIP encrypted traffic using a machine learning approach", Journal of King Saud University-Computer and Information Sciences, Vol. 27, No. 1, pp.77-92, 2015.

[11] Ducange, P., Mannarà, G., Marcelloni, F., Pecori, R., amd Vecchio, M., "A novel approach for internet traffic classification based on multi-objective evolutionary fuzzy classifiers", In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-6, 2017.

[12] Yamansavascilar, B., Guvensan, M. A., Yavuz, A. G., and Karsligil, M. E., "Application identification via network traffic classification", In International Conference on Computing, Networking and Communications (ICNC), pp. 843-848, 2017.

[13] Ertam, F., and Avcı, E., "A new approach for internet traffic classification: GA-WK-ELM", Measurement, Vol. 95, pp. 135-142, 2017.

[14] Zhen, L. I. U., and Qiong, L. I. U., "Studying cost-sensitive learning for multi-class imbalance in Internet traffic classification", The Journal of China Universities of Posts and Telecommunications, Vol. 19, No. 6, pp. 63-72, 2012.

[15] Peng, L., Zhang, H., Chen, Y., and Yang, B., "Imbalanced traffic identification using an imbalanced data gravitation-based classification model", Computer Communications, Vol. 102, pp. 177-189, 2017.

[16] Zhao, J. J., Huang, X. H., Qiong, S. U. N., and Yan, M. A., "Real-time feature selection in traffic classification", The Journal of China Universities of Posts and Telecommunications, Vol. 15, pp. 68-72, 2008.

[17] Bolon-Canedo, V., Sanchez-Marono, N., and Alonso-Betanzos, A., "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset", Expert Systems with Applications, Vol. 38, No. 5, pp. 5947-5957, 2011.

[18] Zhang, H., Lu, G., Qassrawi, M. T., Zhang, Y., and Yu, X., "Feature selection for optimizing traffic classification", Computer Communication, Vol. 35, No. 12, pp. 1457-1471, 2012.

[19] Sun, M., Chen, J., Zhang, Y., and Shi, S, A new method of feature selection for flow classification, Physics Procedia, 24, 2012, pp. 1729-1736.

[20] Panchal, G., Ganatra, A., Kosta, Y. P., and Panchal, D., "Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers", International Journal of Computer Theory and Engineering, vol. 3, no. 2, pp. 332-337, 2011.

[21] Agarap, A. F., "Deep learning using rectified linear units (relu)," arXiv preprint arXiv:1803.08375, 2018.

[22] Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization". arXiv preprint arXiv:1412.6980, 2014.

[23] Manju, N., Harish, B. S., & Prajwal, V. "Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier", International Journal of Computer Network and Information Security, Vol. 11, No. 7, PP.37-42, 2019.