A Survey of Machine Learning Algorithm in Network Traffic Classification

Supriya Katal[1], Asstt. Prof. Hardeep Singh [2]

[1](Department of CSE/Lovely Professional University, INDIA)
[2](Department of ECE/Lovely Professional University, INDIA)

ABSTRACT: *Network Traffic Classification is an emerging research area and now a day the research is widely used in various activities such as intrusion detection system and for security purpose. Many of the protocols and proposed application have been investigated and developed by using machine learning algorithms. We also focused on traditional and statistical methods. The previous used techniques and recent techniques have been compared using machine learning. In this paper the survey of different machine learning techniques are done and I identify better techniques to improve the performance.*

Keywords: *Network Traffic Classification, machine learning, supervised classification, feature set*

I. INTRODUCTION

Network Traffic classification has extensively researched in recent years and many techniques has been proposed including Flow-Based technique, Host-Based technique and Graph-Based technique [1]. Some of them were under research but many of them had achieved great success in the area of research.

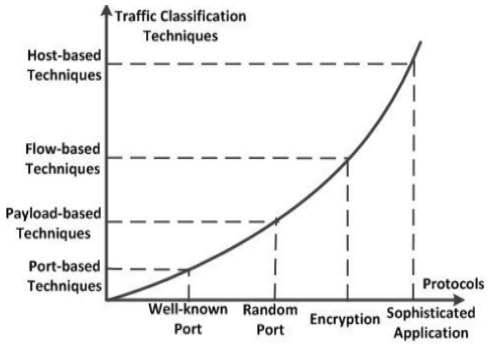

Figure 1. Evolutions of Protocols and Classification techniques

Now a day, flow-based techniques are in great progress. In flow-based, we light on feature selection and in machine learning, we named it as variable selection or attribute selection. It contains many redundant or irrelevant data and the extract new features from their currently selected data set. Previously, we use P2P, VoIP, and Bit Torrent and now new applications had taken place like Google Talk, Face book, Cloud computing, big data, Hadoop and yahoo messenger for extracting feature data set from traffic flow. Machine learning [2](Rostmizadeh, 2012) deals with data mining also whether it is an artificial intelligence tool and we do studies related to learning so that we can learn from known applications of data. It also deals with generalization and representation of data in data mining. Machine learning and data mining in most aspects are the same because machine learning predicts data based upon known

properties whereas data mining is discovery of unknown properties.

Machine learning also gave some of the algorithms like data mining such as:

- Supervised learning(Classification)
  - Decision tree
  - K-Nearest Neighbor
  - Linear regression
  - Naive Bayes
  - Neural networks
- Unsupervised learning (Clustering)
  - K-Mean
  - Expectancy Mean
  - DBSCAN
  - OPTICS
- Semi Supervised learning
  - Support Vector machine

This paper is organized as follows in section II literature review of network traffic is explained. In section III process of network traffic and conclusion in section IV.

II. LITERATURE REVIEW

Bin Hu and Yi Shen [3] describes Machine learning based Network traffic Classification regarding QoS, accounting and intrusion detection. Previously, we had traditional methods like port match method and after having enhancements in internet we shifted to payload analysis. This analysis is not so popular because payload is unable to encrypt traffic so researchers moved to statistical feature based approach. In this paper, I focused on two techniques of machine learning: Supervised and Unsupervised. We focused on to identify the flow classification using statistical approach for calculating the better performance of flow.

Wang Ruoyu [4] gave a new resampling method for network traffic classification using Supervised Machine Learning algorithm. By resolving the limitations of old techniques in machine learning we had gone through flow based feature technique which comes under statistical technique. In this paper we explore the information about traffic clustering with constraints using correlation information, by using K-mean algorithm. When we gone through this paper we came to know that unsupervised is the best technique for traffic classification. This paper showed us that not only the convergence speed got improved but also the quality of clusters. We believe that UDP (User datagram protocol) is not better than TCP (Transport protocol) so we use the more dataset as user protocol.

Kuldeep Singh and Sunil Agrawal [5] gave a comparative analysis of five machine learning algorithms for IP traffic classification. In this paper, we had taken five algorithms MLP(Multilayer Perceptron), RBF(Radial Basis Function Neural Network), C4.5, Bayes net and Naive Bayes for IP traffic classification with these datasets such as packet capturing tool and attribute selection algorithm. In previous research Bit Torrent is used at a large extent and now a days, YouTube, yahoo messenger and Google talk are rising in IP traffic. Bayes Net gives better result on the basis of classification.

Thuy T. T. Nguyen [6] proposed timely and continuous Machine Learning based Classification for interactive IP traffic and it relied on QoS statistic for few packets only. Even the quality of service is required for timely and continuous based classification. Firstly define the sub-flow for classification on the basis of two algorithms: Naive Bayes and C4.5 Decision Tree machine learning algorithms for the identification of first-person-shooter online game and VoIP traffic. The results are classified on the basis of recall and precision for calculating the

identification of IP traffic. It has implemented on two stages, training and testing. In this the concept of sliding window is introduced along with DIFFUSE for packet flows and WEKA tool is implemented for further evaluation. We can allow portability, scalability and stability of flow. The results are calculated on recall and precision for online game and VoIP and data sets are also opted for the packets flow i.e. unidirectional or bidirectional. We are solving only with few packets and same we can do with large number of packets also but the data sets get increase for calculating the large value and we can get the best solution for this problem.

Jun Zhang [7] Network Traffic Classification using Correlation information gave two techniques, supervised classification algorithms and unsupervised clustering algorithms. Recently the work on flow statistical feature based classification methods has not been solved by the researchers and the work is going on by improving the performance of flow. In unsupervised traffic classification, it is difficult to build an application oriented traffic classifier with the help of clustering and moreover without knowing the real traffic classes. The supervised traffic classification can be divided into two categories: parametric and nonparametric classifiers. Parametric classifiers use algorithms such as C4.5 decision trees, Bayesian networks, neural networks and nonparametric classifiers such as K-Nearest Neighbor.

In this paper, the nonparametric approach had given best solution regarding the performance of correlation information by opting Nearest Neighbor algorithm. This solution arises on the basis of both empirical and theoretical perspectives. In this new real-world methods and datasets have been introduced to show the performance under few training samples. The same proposed work we can opt for semi-supervised also as a future work so that an accurate result can be produced by the information.

III. PROCESS OF NETWORK TRAFFIC

Nguyen et al. Described the flow classification process in figure 2 [3]. The process is consisted of two phases:

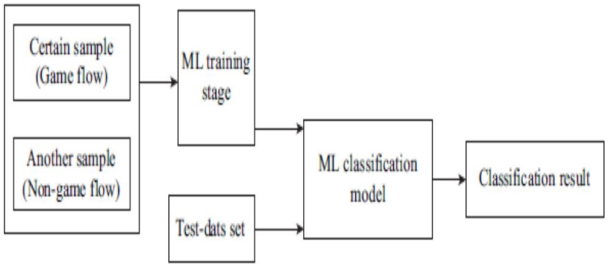

Figure 2. Flow Classification process

1. Training (study) phase: First calculates and then label the flow statistical feature set information from network data then obtain sample set and from that sample chooses the best feature by using classification method in supervised learning for generating the output [3].
2. Testing (classification) phase: first measure and compute flow statistical feature and then submit the feature vector to identify flow type [3].

IV. CONCLUSION

In network traffic, classification is better than clustering so we mainly focus on supervised learning and this is our main step. We can also combine old techniques with new machine learning techniques to improve the performance of network traffic. We also focused on classifiers like NN classifier and methods namely DBSCAN, DIFFUSE for packet identification. We can

perform our result with the help of MATLAB and WEKA for better results. In future we can also use network traffic for cloud computing and Hadoop and big data for security purpose or identification of traffic.

REFERENCES

[1] Yibo Xue, Luoshi Zhang and Dawei Wang, *Traffic Classification: Issues and Challenges*, Journal of Communication *Vol. 8, No. 4*, April 2013.

[2] Rostmizadeh, retrieved from Wikipedia: en.wikipidea.org/wiki/Supervised learning, October 2012.

[3] Bin Hu, Yi Shen, *Machine learning based network Traffic Classification: A Survey*, Journal of Information and Computational science *9: 11(2012) 3161- 3170*.

[4] Wang Ruoyu, a new re-sampling method for network traffic Classification using SML, Project supported by National 973 Program of China, IEEE (2010).

[5] Kuldeep Singh and Sunil Agrawal, Comparative Analysis of five machine learning algorithms for IP traffic classification, IEEE (2011).

[6] Thuy T. T. Nguyen, Grenville Armitage, Timely and Continuous Machine – Learning- Based Classification for interactive IP Traffic, IEEE/ACM Transactions on Networking, Vol. 20, No. 6, Dec 2012.

[7] Jun Zhang, Yong Guan, Network Traffic Classification using Correlation information, IEEE transactions on Parallel and Distributed systems, Vol. 24, No. 1, Jan 2013.