Analyzing The Classification of Network Traffic Using Machine Learning Technique

Lina P. Talole[1], Swati V. Borle[2], Vaishnavi G. Patil[3], Anjali R. Ingle[4], M. V. Shastri[5]
[1, 2, 3, 4, 5] Dept of Computer Science & Engineering
[1, 2, 3, 4, 5] Padm. Dr. V. B. Kolte. College of Engineering , Malkapur, Buldhana, India

Abstract- *The Internet is consistently growing in size and becoming more complex. the sphere of networking is thus continuously reaching to deal with this monumental growth of network traffic. While approaches such as Software Defined Networking (SDN) can provide a centralized control mechanism for network traffic measurement, control, and prediction, still the quantity of information received by the SDN controller is huge. To process that data, it's recently been suggested to use Machine Learning (ML). during this paper, we review existing proposal for using ML in an SDN context for traffic measurement (specifically, classification)and traffic prediction. we are going to especially concentrate on approaches that use Deep learning (DL) in traffic prediction, which seems to own been mostly untapped by existing surveys. Furthermore, we discuss remaining challenges and suggest future research directions.*

Keywords- Network measurement, software defined networking, machine learning, deep learning, traffic classification, traffic prediction

I. INTRODUCTION

In recent years, with the rapid development of the net, the network traffic data also showed explosive growth, while giving people convenience, but also to the effective network management, security, network environment has brought great challenges, like virus flooding, it's difficult to watch the network of unhealthy content, P2P applications take plenty of network bandwidth and other issues. in sight of the issues in these networks, network researchers have proposed capacity planning, traffic scheduling and other strategies to boost the operational efficiency of the network. However, the premise of those strategies is to classify the traffic, that the network traffic classification technology is more and more by many network research scholars and network service providers. At present, the network traffic classification technology is especially employed in service quality and traffic engineering, network security monitoring, network management technology. The fast development of the web and communication devices has created bigger and more complicated network structures, adapting and developing bigger hubs, routers, switches, etc.

The complexity in networks has introduce an overflow of more amounts of traffic data and also contributed to the challenges in network management and traffic optimization, including traffic measurement (For example : traffic classification) and traffic prediction. In parallel, we are seeing two promising solutions to assist manage networks more efficiently: SDN and Machine Learning. SDN provides a centralized access and control mechanism to all or any networking devices, where the SDN controller cannot only monitor and measure all varieties of network parameters and metrics, but can even make a more informed and efficient decision about resources allocation and routing, since it's a worldwide view of everything within the network. However, the number of knowledge an SDN controller receives may well be overwhelming. While the SDN controller itself is made scalable, as an example by running it in a very cloud, still efficient algorithms are needed to extract the required measurements and data from the received data. Here is where Machine Learning can help. Many of the traffic classification and traffic prediction issues will be performed efficiency by various ML algorithms, improving the system performance while maintaining relative simplicity in design. during this survey, we review existing approaches for traffic classification and traffic prediction which use ML in an SDN context.

We especially specialise in ML's subcategory of Deep Learning (DL), which has not been covered in details by existing surveys. Therefore, our contribution is covering DL methods for traffic prediction, which is usually not covered within the existing surveys, while we also cover some newer works in machine learning and deep learning for both traffic classification and prediction that existing surveys haven't covered. As an example of the importance of network traffic classification, one can think about the asymmetric architecture of today's network access links, which has been designed supported the assumption that clients download quite what they upload. However, the pervasiveness of symmetric-demand applications [such as peer-to-peer (P2P) applications, vox IP (VoIP) and video call] has changed the clients' demands to deviate from the idea mentioned earlier. Thus, to supply a satisfactory experience for the clients, an application-level knowledge is required to allocate adequate resources to such applications. The emergence of latest applications

furthermore as interactions between various components on the net has dramatically increased the complexity and variety of this network which makes the traffic classification a difficult problem

• Traditional Machine Learning Algorithms

ML could be a data analysis method that learns from data to identify patterns within it and make decisions supported the knowledge collected. It generally involves preprocessing, training and testing phases. The preprocessing includes actions like data preparation, filtering, imputation, and tuning for specific purposes. Once the information is preprocessed, ML methods are implemented to coach the information. Then the system makes decisions supported the input received from the training phase. ML algorithms are often studied under supervised or unsupervised learning where the previous is given labeled training data and therefore the latter works with unlabeled training data trying to extract information through clustering consistent with the resemblance within the observation points.

The following are the ML and DL algorithms utilized in this survey. Note that each one but Analyzing the classification of network traffic using deep learning technique

• Nearest Centroid (NC):

It computes the centroid for every labeled class. It calculates the space between the observation points and also the centroid. Then it assigns the information points to the category whose centroid has the minimum distance to the observation.

• Naive Bayes (NB):

it's a straightforward probabilistic classifier supported implementation of Bayes' theorem. it's used when the info dimensionality is high since it assumes the info features are independent from one another.

• Decision Tree (DT):

It is one more simple algorithm. It performs a call classifier through a tree-like model with leaf nodes which correspond to the category label, and also the path from the tree roots to the leaf are related to the classification rules.

• Random Forest Tree (RF):

It is an extension of DT that aggregates few DTs and fixes the overfitting problem by randomly selecting a subset of knowledge features.

• Support Vector Machine (SVM):

It is a binary classification and pattern recognition technique which maps the info points in n-dimensional space and plots the hyper plane that separates them into different clusters.

• Multi-Class Support Vector Machine (MCSVM):

In order to segregate the info into quite two classes, SVM is applied as a series of binary problems. However this is often computationally expensive. Therefore new methodologies are developed to mitigate this issue. Analysing the classification of network traffic using deep learning technique

• Laplacian Support Vector Machine (LapSVM):

It is an extension of SVM which regularizes the SVM by a Laplacian graph [2].

• Adaptive Boosting (AdaBoost):

It is a boosting technique that builds more accurate algorithms by creating a hybrid classifier out of weak classifiers.

• Gradient Adaptive Boosting (G-AdaBoost)

It functions in three means : first is optimization of a loss function, 2nd one is predictions from a weak learner, and the last is minimization of the loss function via the hybrid model of the weak learners.

• M5Rules:

It is found under Weka software. It makes decisions for prediction problems by combining decision trees and regression toward the mean.

• Linear Regression:

When used for prediction and fore- casting purposes, it tries to suit a model to data points supported independent variables.

• Polynomial Regression:

It shifts a regression toward the mean model into a curve to raised fit the observation points.

• K-means:

Also remarked as k-means clustering. Unlike other methods mentioned, k-means is an unsupervised learning algorithm. It divides the info into k different clusters within which each datum is assigned to a cluster with the closest average.

Analyzing the classification of network traffic using machine learning technique

• Traditional Deep Learning Algorithms

Deep Learning algorithm uses multiple layered neural networks which are biologically inspire computing systems with input, hidden and output layers consisting of inter-connected neuron like nodes. These nodes contain activation functions. And the Information is fed through the input layer. The pattern recognition process is completed within the hidden layer via activation functions and the solution is present within the output layer. Each layer take the output of the previous layer(s) as input and then apply non linear transformation to extract useful features for classification.

• Convolutional Neural Network (CNN):

It is a kind of NN that's build around three ideas: convolutional layers, weight sharing, and pooling. Convolutional layers and weight sharing function as filters that detect localized features within the data and reduce the quantity of information parameters whereas pooling further reduces the feature size while keeping the invariance of the information.

• Auto encoders (AE):

It is an unsupervised learning algorithm that encodes the info through dimensionality reduction. It trains the network by reconstructing its input. Its variations are sparse, de noising, contractive, convolutional, stacked.

• Recurrent Neural Network (RNN)

It is a network with loops that preserves its input because of its internal memory. similar to an individual's behavior, when it makes a choice, it takes into consideration this information it's and former experience gained through loops. Its most well liked implementation is Long remembering (LSTM). It back propagates the errors through layers to be told

II. SYSTEM IMPLEMENTATION

Proposed Work

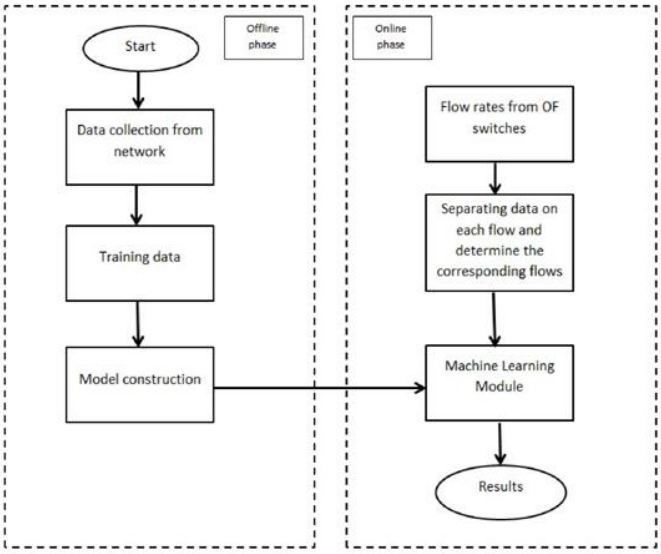


Fig : Proposed System Architechture

In the beginning, collect application traffic. during this step, we capture from various applications through Microsoft Network Monitor 3.4. Application traffic should be captured in appropriate amount. If the quantity of traffic isn't adequate, it'll not only give accurate classification results, but also will take a protracted time.

In the second step, extract the educational features of the collected application traffic. during this step, rather than the prevailing learning feature extraction system, the proposed learning feature extraction system is employed. Few of them learning features are extracted by using the existing system. Where as the proposed learning feature extraction system can solve this problem.

Extracted learning features of every application are divided into plaything and Test Set. plaything may be a set of extracted learning feature to be learned through machine learning. Test Set may be a set of extracted learning feature to be tested whether the classified traffic is correctly classified at the following stage.

In the third step, classified plaything and Test Set are normalized. There are variety of the way to normalize. However, the results greatly rely upon which normalization method is employed. Therefore, it's also important to use

sophisticated and appropriate normalization method during this step.

Finally, conduct several experiments with normalized toy and Test Set. we are going to use the varied classification methods of Tensor flow.

Initially extracting methods not only complicated but also few learning features are extracted. Therefore, we propose a system that may extract learning features of traffic more efficient and various than other extracting methods to take care of the anomaly and normal traffic in network.

Separation of knowledge plan and control plan, gives ability to network administrators to form programmable policies and simply manage data plan via the controller. SDN also makes it easy to possess a dynamic management, configuration, troubleshooting and even testing new protocols and ideas within the network without troubles. a crucial case in network management for having high availability and efficiency is traffic classification.

There are many methods to applying on traffic classification in networks:

• Using port numbers to work out application and application layer protocols. However, these methods don't seem to be completely accurate.
• Deep Packet Inspection (DPI) is employed. These methods have high accuracy, however, there are some issues regarding its implementation while dynamic ports and encrypted traffics aren't supported in current networks yet. And also it causes high overhead to the system but also the violates user privacy.
• These methods have their own problems, so, researches are recently that specialize in machine learning techniques, which profit of statistical properties for traffic classification.
• Although there are many challenges in current networks for traffic classification, global view of controllers in SDN improves network management while its concept is easy and straightforward to use for extracting the statistical data from network traffic with the help of switches.
• There are many techniques accustomed analyze network traffic, like self-similarity and TES, which are supported communication system analysis and attacks discovery. Meanwhile, flow analysis is predicated on the identification of anonymity networks. to stay up with the evolving attacks, many enterprises use machine learning to uncover various forms of malicious behaviors including mass registration of pretend accounts fraudulent transactions and fraud, where security will be achieved through alleged machine learning applications and method. Machine learning (ML) means the pc can work out an answer without being specifically programmed. That is, machines are able to continuously learn and pander to huge datasets using classifiers and algorithms.

• Classifiers, which categorize observations, are considered the backbone of ML. Meanwhile, other ML algorithms are build models of behaviors and use those models as a basis for creating future predictions supported new input file. the facility of machine learning tools lies in detecting and analyzing network attacks without having to accurately describe them as previously defined. Machine learning can aid in solving the foremost common tasks including regression, prediction, and classification within the era of extremely great amount of knowledge and cybersecurity talent shortage. Machine-learning techniques are applied in many aspects of network operation and management, where the system performance will be optimized and resources is better utilized.

• Besides, clustering and the classification extract patterns out of these information packets which might be employed in the many applications as security analysis and the user profiling. What'smore, there are many applications for the analyzing of traffic supported the machine learning algorithms like determining the anomalies through discovery-based workbooks otherwise features which describe user behaviour. the mix of ML algorithms and traffic analysis could be the hot topic of thanks to the facility of machine learning tools that lies in between the detecting and analyzing network attacks unescorted by having to accurately describe them as previously defined.

ML may be employed in various sector of cybersecurity to provide analytical approaches to detect and answer an attack. It also able to enhance safety operations. Following are some frequently used machine learning mechanism in traffic analysis.

• Network-based defense:

This method protects the network by attempting to scale back the chance of attacks by providing an extra layer of security, because each layer has established policies and controls in situ to define users who are authorized to access the network. As a primary step, a framework was developed to gather data and filter traffic. This basis then uses ML techniques in defense against attacks. The results can help in determining the probability and impact of the attack in regard to a particular network area. Therefore, it can help different organizations to cut back the danger of victim exposure.

• Intrusion detection:

Machine learning algorithms may be implemented in applications to spot and reply to cyber-attacks before they go. Network monitoring system for a malicious activity where the behavior varies between hackers and also the normal user, and thus, they will determine the identity. this can be usually achieved employing a model developed by analyzing big data sets of security events and identifying the pattern of malicious activities. As a result, when similar activities are detected, they're automatically addressed

• Requirements and time:

Classifying network traffic requires periodic updating and training of the model, additionally thereto it has to define the task within the automated model. the method is big in terms of harmful traffic and alter in types and features of cyber-attack. This takes it slow to guage the information and its scope through careful data management Data integrity and data exploration until you get clear data. Therefore, it is difficult for devices that detect anomalies in a very timely manner.

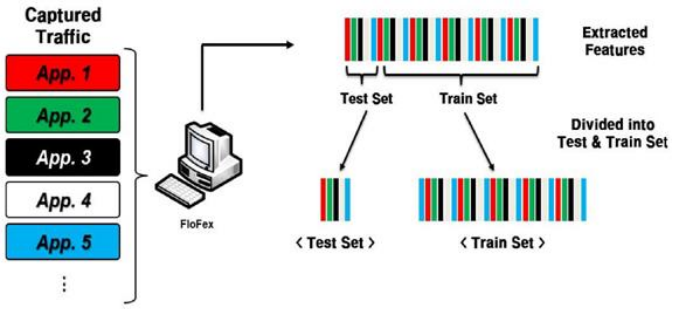
• Complexity:

This process is large with relation to harmful traffic and therefore the change in types and features of cyber-attack. Moreover, there are many criteria to contemplate when determining how effective roads are in detecting the sort of attack, and whether it's impossible to form one recommendation for every method. additionally, each technical layer has different subcategories; it's difficult to use the identical algorithms to them.

III. METHODOLOGY

The collected application traffic and extract learning features by using system network. Then extracted learning feature is split into toy and Test Set. Because each of two Sets contains a different of role. plaything may be a Set for learning. It occupies for 70 ~ 80% of the extracted learning features. Test Set could be a Set to verify learning. Test Set occupies the remainder of the extracted learning features. Next we normalize learning features which are extracted from system network. Because the difference between each values of extracted learning features is simply too large to classify. If we don't normalize these values, it classified inaccurately. Therefore, it's important to use an appropriate normalization method. Although there are many methods of normalization, we use Min-Max Normalization. Because, it's more simple and efficient than other methods. within the final step, the experiment is performed with the normalized toy and Test Set.

First, train with the plaything and test with the Test Set. There are some ways to learning and classifying by using Tensorflow. Among them, we are going to use softmax regression. Softmax regression could be a classification method that's used mainly for 3 or more classification objects (Multinomial Classification). When there are 2 classification objects, we should always use logistic regression (Binary Classification). it's more simple method than softmax regression. However, the amount of application traffic which we should always classify is quite 2. Therefore, it's suitable to use softmax regression.



Second, test with plaything and check its accuracy. during this case, we must always conduct many experiments to get the optimal learning rate and therefore the number of loops with the best accuracy. And apply them to the foremost optimal extracted learning features with the obtained learning rate and therefore the number of loops. Then, we make an application traffic classification model supported machine learning so as to prove the validity of the proposed system, we conduct many experiments. We collected the net browser application traffic. the online browsers utilized in this experiment are Chrome, Firefox, Internet Explorer, Swing, and Whale. All of those 5 web browsers application traffic are collected in same conditions. the amount of loops is that the number of learning. Usually if the quantity of loops is greater, the educational result's better. But it takes plenty of your time to run many loops. Learning rate indicates the speed of learning. If the training rate is simply too small, then it'll spend plenty of your time. On the opposite hand, if the educational rate is just too big, then spending time are going to be shortening, but the results of accuracy are going to be poor. Learning Feature List Structure Extracted from system network was important to search out the optimal learning rate through experiments. Cost shows the difference between the particular value classified and also the predicted value supported learning. Cost is that the better when it converges to 0. Accuracy, on the opposite hand, represents the proportion of correctly classified traffic among the analyzed traffic.

It shows that, just in case of full feature dataset, Bayes Net Classifier provides the higher accuracy which is 82.33 %, C4.5 provides the upper accuracy which is 93.33%. These algorithms as Bayesian Network, Naïve Bayes, Nearest Neighbor, C4.5, and RBF Decision Tree are shown in Figure1. Among these algorithm C4.5 decision tree gives high accuracy (93.55%) for traffic classification. The Training Time of this algorithm is 0.05 seconds in comparison with other algorithms.

Working of the Machine Learning C4.5 Algorithm
For each and every attribute a, find the normalised information gain ratio from splitting on a.
Let a_best be the attribute with the best normalized information gain.
Create a call node that splits on a_best.
Repeat the sublists acquired by the splitting on a_best, and add these nodes as child of node.
Advantages of C4.5 over other Decision Tree systems
These algorithm fundamentally employs Single Pass Pruning Process to Mitigate overfitting.
It can cooperating not only with the Discrete but also Continuous Data
C4.5 can handle the problem of incomplete data fine

In the opening move, collect application traffic. during this step, we capture from various applications through Microsoft Network Monitor 3.4. Application traffic should be captured in appropriate amount. If the quantity of traffic isn't adequate, it'll not only give accurate classification results, but will take an extended time.

In the second step, extract the training features of the collected application traffic. during this step, rather than the prevailing learning feature extraction system, the proposed learning feature extraction system is employed. Few of them learning features are extracted with the help of the existing system. Although the proposed learning feature extraction system can solve this problem. Extracted learning features of every application are divided into plaything and Test Set. toy could be a set of extracted learning feature to be learned through machine learning. Test Set could be a set of extracted learning feature to be tested whether the classified traffic is correctly classified at the following stage.

In the third step, classified plaything and Test Set are normalized. There are variety of the way to normalize. However, the results greatly depends on which normalization method is employed. Therefore, it's also important to use sophisticated and appropriate normalization method during this step.

Finally, conduct several experiments with normalized plaything and Test Set. we'll use the assorted classification methods of Tensorflow. To begin with extracting methods not only complicated but also few of them learning features are extracted. Therefore, we propose a system which will extract learning features of traffic more efficient and various than other extracting methods.

We gathered application traffic and the extract learning features by using the FloFex. Then extracted learning feature is split into toy and Test Set. Because each of two Sets encompasses a different of role. plaything could be a Set for learning. It occupies for 70 ~ 80% of the extracted learning features. Test Set could be a Set to verify learning. Test Set occupies the remainder of the extracted learning features. In the next step, we normalized the learning features which are extracted from FloFex.

Because the difference between each values of extracted learning features is just too large to classify. If we don't normalize these values, it classified inaccurately. Therefore, it's important to use an appropriate normalization method. Even if there are many methods of normalization, but from that we use only the latest Min-Max Normalization. Because, it's more simple and efficient than other methods. within the final step, the experiment is performed with the normalized plaything and Test Set.

IV. CONCLUSION

In this way, the ML and DL methods used for classification and prediction in SDNs. First, it explained the algorithms and therefore the SDN architecture. Next, we summarized the prevailing works. We then presented our survey and at last we addressed the challenges and future work that are dataset characteristics, data volume, methodology of applying DL, security related issues because of SDN structure, and flow encryption. Since employing ML and DL algorithms for classification and prediction in SDN is sort of new, more problems can be identified in practice that can't be predicted now.

In addition, there's no rigorous theoretical framework to style and analyze such networks. If there's some progress in these matters, it'll have direct impact on proposing better deep neural network structures specialized for network traffic classification. Along the identical line, one in every of the opposite important future direction would be investigating the interpretability of our proposed model. this may include analyzing the features that the model has learned and also the process of learning them.

V. ACKNOWLEDGMENTS

We would like to express our deep gratitude and thanks to Mr. M.V.Shastri (guide) Padm. Dr.V.B.K.Coe, Malkapur. giving us an opportunity to work under his guidance for our review of research papers and his consistent motivating & direction in this regard. We extend our sincere thanks to Mr. M. V. Shastri. HOD, Padm. Dr.V.B.K.Coe, Malkapur for his continuous support and encouragements throughout the course work. Last but not least I would like to thank my parents & family who always inspired and directed me. I would like to thank the all (Principal, Hod ,Faculty members) people who were involved directly or indirectly to complete our review paper work.

REFERENCES

[1] B. Raghavan, M. Casado, T. Koponen, S. Ratnasamy, A. Ghodsi, and S. Shenker, "Software-defined internet architecture: decoupling architec- ture from infrastructure," in Proceedings of the 11th ACM Workshop on Hot Topics in Networks.

[2] T. Koponen, M. Casado, N. Gude, J. Stribling, L. Poutievski, M. Zhu, R. Ramanathan, Y. Iwata, H. Inoue, T. Hama et al., "Onix: A distributed control platform for large-scale production networks." in OSDI, vol. 10, 2010, pp. 1–6.

[3] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: enabling innovation in campus networks," ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, pp. 69–74, 2008.

[4] J. Yan and J. Yuan, "A survey of traffic classification in software defined networks," in 2018 1st IEEE International Conference on Hot Information-Centric Networking (HotICN). IEEE, 2018, pp. 200–206.

[5] N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, "Survey on sdn based network intrusion detection system using machine learning approaches," Peer-toPeer Networking and Applications, pp. 1–9, 2018.

[6] J. Su a´rez-Varela and P. Barlet-Ros, "Sbar: Sdn flow-based monitoring and application recognition," in Proceedings of the Symposium on SDN Research. ACM, 2018, p. 22.

[7] A. Abubakar and B. Pranggono, "Machine learning based intrusion detection system for software defined networks," in 2017 Seventh International Conference on Emerging Security Technologies (EST). IEEE, 2017, pp. 138–143.