Network Traffic Classification using Machine Learning for Software Defined Networks

J.K.Menuka Perera, Kandaraj Piamrat, and Salima Hamma
LS2N/University of Nantes, France
jkmenukaperera@gmail.com/ firstname.lastnameg@univ-nantes.fr

Abstract – The recent development in industry automation and connected devices made a huge demand for network resources. Traditional Networks are becoming less effective to handle this large number of traffic generated by these technologies. At the same time, Software defined networking (SDN) introduced a programmable and scalable networking solution that enables Machine Learning (ML) applications to automate networks. Issues with traditional methods to classify network traffic and allocate resources can be solved by this SDN solution. Network data gathered by the SDN controller will allow data analytics methods to analyze and apply machine learning models to customize the network management. This work has focused on analyzing network data and implement a network traffic classification solution using machine learning and integrate the model in SDN platform.

Keywords: Machine Learning, Classification, Network Traffic, Software-Defined Networking.

I. INTRODUCTION

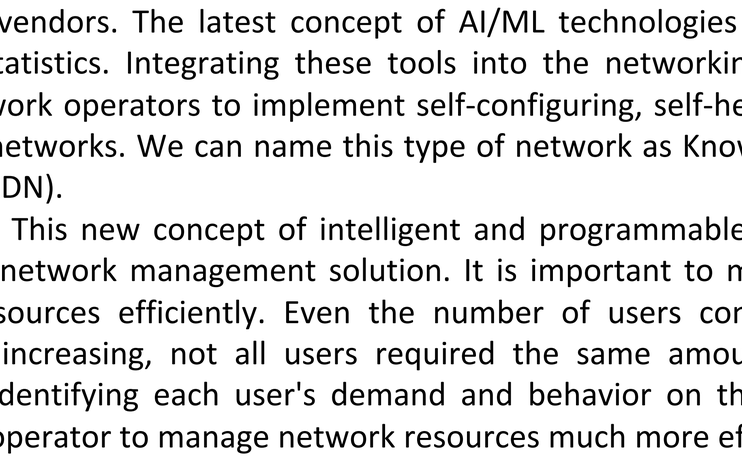
Recent advances in software defined networking and machine learning techniques have created a new era of network management. This new concept has combined network intelligence and network programmability to create autonomous high performing networking, which will expand 5G (5th Generation) capabilities. With the recent improvements in Internet of Things (IoT), cloud computing self-driving vehicles, etc., the demand for bandwidth consumption has increased exponentially and pushed network operators the ability to search for new concepts of network management. Software defined networks provide a programmable, scalable and highly available network solution. This solution has a global view of the network and enables the network operator to program their policies rather than depending on network equipment vendors. The latest concept of AI/ML technologies are developed based on statistics. Integrating these tools into the networking industry will enable network operators to implement self-configuring, self-healing, and self-optimizing networks. We can name this type of network as Knowledge Defined Networks (KDN).

This new concept of intelligent and programmable network is an end-to-end network management solution. It is important to manage existing network resources efficiently. Even the number of users connected to the network is increasing, not all users required the same amount of network resources. Identifying each user's demand and behavior on the network will enable the operator to manage network resources much more efficiently. When it comes to network traffic classification, ML algorithms depend on a large number of network features. And software defined networking will enable ML algorithms to control the network and can become automatic resource allocation process.

Therefore, in this study, ML-based traffic classification solution was introduced for SDN. The proposed architecture uses existing network statistics and an offline process for understanding network traffic patterns with a clustering algorithm. For the online process, a classification model is used to classify incoming network traffic in real-time.

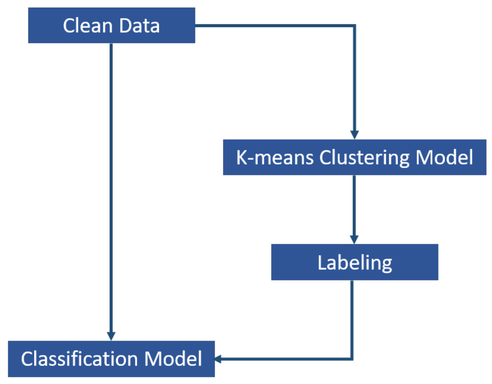
II. PROPOSITION

This proposed solution was divided into two sections. One was to train the machine learning algorithm and the other was to create a network experiment to run the trained ML model on an SDN platform as a proof of concept.



- **Machine Learning Algorithm**

For this work, "IP Network Traffic Flows, Labeled with 75 Apps" dataset from Kaggle database was used. This dataset was created by collecting network data from Universidad Del Cauca, Popayn, Colombia using multiple packet capturing tools and data extracting tools. This dataset is consisting of 3,577,296 instances and 87 features and originally designed for application classification. But for this work, only a fraction of this dataset is needed. Next the standard data cleaning process was done in order to avoid errors and biasing during the model training. Even though the data was clean enough to train ML models, data was not labeled. Classification process is a supervised learning algorithm that need labeled data for the training process. Understanding the traffic patterns in the dataset is a complicated and time-consuming task. Since the dataset is very large, it is very hard to label traffic flows manually. To avoid manual labeling, an unsupervised learning model can be used.

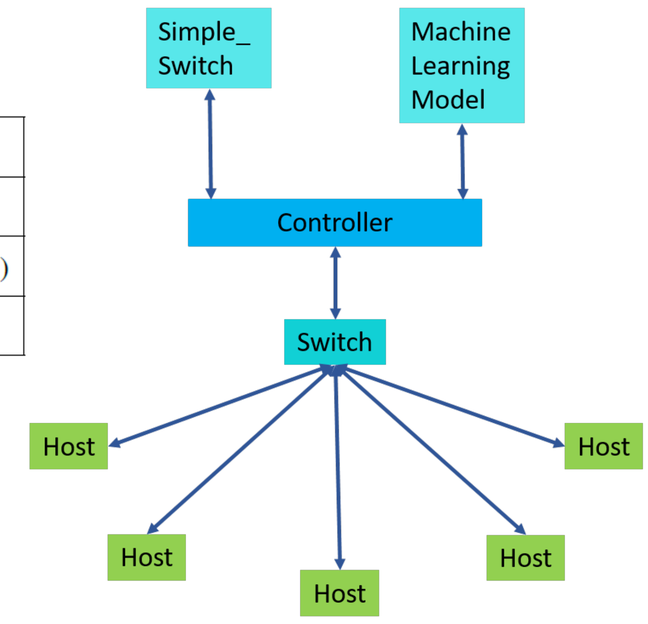


Labeled data was used to train classification models. There are multiple classification models available and each and every model classify data with different mathematical models. In other word, it is better to train and test multiple classification models to find out which model fit better for the project.

- **Network Application**

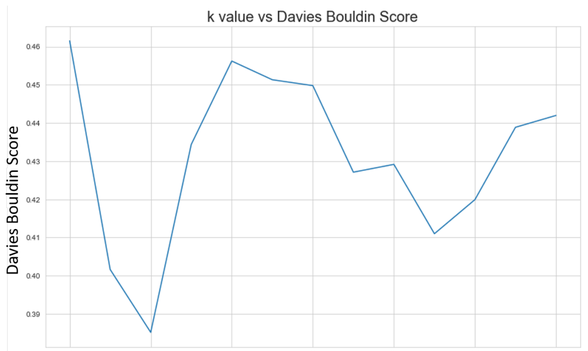
For the simulation testbed, a simple virtual network was created on Mininet network emulator with five hosts, one OpenFlow enabled open vSwitch and one SDN controller (RYU). For the simplicity of this research, tree topology was used.

System OS	Ubuntu(18.10)
SDN Controller	RYU(4.30)
Switches	Open vSwitch(2.11)
Network Emulator	Mininet(2.2.2)



III. Performance Evaluation

In the Kmeans clustering results, the number of clusters (k value) will be varied from 2 to 15 and calculate the Davies-Bouldin score for each k value. Based on the score, there are four types of traffic behaviors that can be identified from this dataset. The data set was labeled and five supervised learning models were trained and evaluated. Following are the accuracies of trained models.



Model	Accuracy
SVM (Linear)	96.37%
SVM (RBF)	70.40%
Decition Tree	95.76%
Random Forest	94.92%
KNN	71.47%

The trained classication model was integrated with the network application and evaluated the real-time network traffic classification by generating network traffic in the testbed using D-ITG tool. For this evaluation, 50 traffic flows were generated considering cluster characteristics identified by the clustering algorithm. Generated traffic were compared with its characteristics and classification outputs. And the results of the complete system has 96% overall accuracy.

IV. CONCLUSION

This work has been carried out as a proof of concept while combining machine learning with software defined networking, in particular, for network traffic classification. It can be seen that traffic classication using machine learning algorithms provides good results within SDN environment. This is possible thanks to the ability of collecting information in this type of architecture. It is clear that this is a promising solution. In the near future, these high performing, intelligence-based networking concepts will enhance or even replace conventional networking management.