Developing Machine Learning Based Framework for Network Traffic Prediction

Ranjdr M. Rafeeq

Computer Technical Engineering, Al-Qalam university College, Iraq
Email: rang.rafeeq@alqalam.edu.iq

ABSTRACT

Network traffic analysis is a crucial step in developing efficient congestion control systems and identifying valid and malicious packets. Because network resources are apportioned based on predicted usage, these solutions reduce network congestion. For a variety of reasons, including dynamic bandwidth allocation, network security, and network planning, the ability to forecast network traffic is critical. Machine learning (ML) techniques to network traffic analysis have received a lot of interest. This article outlines an approach for analyzing network traffic. Three machine learning-based methodologies make up the methodology. The experimental investigation employed the NSL KDD data set. On the basis of accuracy and other criteria, KNN, Support vector machine, and nave bayes are compared.

Keywords: Machine Learning, Classification, Prediction, Accuracy, Network Traffic Analysis

1. Introduction

Network traffic analysis [1] [2] is critical for ensuring the security of sensitive data, especially in the e-commerce, banking, and commercial sectors. Therefore, it goes without saying that network traffic analysis is critical. While network traffic analysis and forecasting are reactive in nature, they are proactive in that they monitor the network for security breaches to guarantee that they do not occur. In order to build effective congestion management systems and identify legitimate and malicious packets, network traffic analysis is a critical step. With these strategies, network congestion is avoided because network resources are distributed based on anticipated traffic. It's crucial to be able to estimate network traffic for a variety of reasons, including dynamic bandwidth allocation, network security, and network planning. There are two types of predictions: long-term and short-term. Long-term traffic forecasting allows for more minute planning and better judgments since it provides a precise forecast of traffic models for evaluating future capacity requirements. Dynamic resource allocation is connected to short-term prediction (milliseconds to minutes). It may be used to enhance QoS processes, control congestion, and manage resources optimally. Packet routing is another usage for it. For network traffic analysis and prediction, a variety of methods are utilized, including time series models, current data mining techniques, soft computing approaches, and neural networks.

In network traffic analysis, machine learning (ML) [3] [4] approaches have gotten a lot of attention. As noted below, ML approaches may be broken down into four categories. There are four types of learning: supervised, semi-supervised, unsupervised, and reinforced. When it comes to large data analytics and

knowledge discovery, Deep Learning (DL) is a critical stage in many machine learning methodologies. Deterministic learning has found applications in a wide range of industries, from the computer vision and healthcare industries to transportation and smart farming. Apart from that, technology-based firms have begun to pay attention to DL (TBC). There are several significant firms that generate enormous volumes of data every day, such as Twitter, YouTube or Facebook [5]. Because processing such a large volume of data using typical data processing techniques is nearly unfeasible, deep learning algorithms are used to evaluate and extract useful information from the generated data. Using network traffic analysis and machine learning, this study sheds light on the intersection of two rapidly developing fields.[6]

2. Literature Review

No previous study has examined how NTMA and deep learning are related, much alone examined how deep learning models apply in NTMA. It's a significant addition to the discussion. NTMA's data mining and conventional machine learning approaches have been the subject of a few academic studies. A few research have provided some DL models for NTMA applications, such as traffic classification. Rezaei et al. [7] studied categorized encrypted traffic using DL models. Several DL-based traffic classification models were examined as part of this research project. However, it did not look at other NTMA applications, which are the focus of this paper. [8] examined supervised, unsupervised, and semi-supervised malware analysis learning algorithms. Aside from that, this research also addresses the underlying issues and concerns. However, the authors did not conduct any study on the importance of DL in malware analysis and detection.

Researchers Conti et al. [9] claim to have undertaken a thorough investigation of network traffic analysis. There are three criteria on which relevant works may be categorized, according to them: (1) analysis goal, (2) traffic monitoring location, and (3) chosen mobile operating system platform. (s). Naive Bayes and C4.5 decision trees, as well as random forests and k-means were among the algorithms investigated. The study compares mobile device analysis methodologies, validation procedures, and results. As a result, whereas [9] focused on typical machine learning approaches, our study focuses on deep learning models.

Fadlullah and colleagues investigated DL models and architectures for network traffic control systems in [10]. In contrast to our study, which focused on NTMA's network architecture, this paper examines how deep models are used in the network.

When it came to the NTMA, D'Alconzo et al. used a big data strategy. Researchers studied prior studies that used big data approaches to better understand network traffic statistics. Also, for four essential NTMA applications, such as traffic classification, traffic prediction, fault management, and network security, the researchers took a cursory look into huge data analytics (such as conventional machine learning). The key difference between this study and others is that no consideration was given to DL models.

Finally, Verma et al. [11] looked at the real-time processing of enormous amounts of IoT data. Real-time IoT data analytics approaches using network data analytics were investigated by the authors of this research. Real-time IoT analytics are also examined in the study's application cases and software platforms. This study, in contrast to earlier ones, did not investigate data analytics using DL models.

In data mining from large datasets, outlier detection, as presented by [12] is a major research subject and an important issue in several disciplines. In data mining applications, outliers, traditionally known as noisy data, have emerged as a prominent focus. Unrecognized and unexpected data may be detected with the help of outlier detection. When it comes to data quality, data preparation focuses on issues like outliers and noise. This step's major objective is to eliminate obstacles to data analysis.

Using the outlier score, authors in [13] described two approaches for discovering and removing outliers in a health care dataset: the

Distance-Based outlier detection method and the Cluster-Based outlier algorithm. It was discovered that the cluster-based outlier detection algorithm was more accurate when used with three integrated health care datasets than the distance-based outlier detection technique. Authors [14] Provided a survey that gave a comprehensive and articulated review of outlier classifications in various temporal data systems, methodologies employed to identify and remove them from the database, and setups with appropriate detection techniques implemented in specific applications.

Work in [15] came up with Feature-Rich Interactive Outlier Detection (FRIOD). When using the proposed outlier detection approach, users will be able to provide feedback at every stage of the process. Dense cell selection was part of the process, as were distance thresholding and top outlier verification. Data clustering is a data mining technique with a variety of uses, including the detection of outliers.

Data clustering and outlier detection were the primary concerns of the author [16]. When identifying the cluster center, the authors used a modified K-means type technique that incorporated an extra "cluster." Actual and fictitious data improved the algorithm's performance.

Open Computing Language was used by the author in [17] to develop a parallel processing system for fuzzy associative classification. The proposed method employed a CPU-GPU implementation to find outbreaks of infectious diseases, such as influenza, using disease and environmental data that had already been gathered. In addition, the study compared the Hybrid technique's results to those of the other approaches.

Association rule mining on Electronic Medical Records (EMR) was utilized by the author in [18] to find groups of risk factors and their related subpopulations that are linked to patients with a high risk of diabetes development. The high dimensionality of EMRs necessitates the use of association rule mining to generate a large number of rules needed for clinical use. Predicting which patients will acquire heart disease was done by researchers [19] using fuzzy rule-based categorization linguistics. Using the framework, professionals may access existing patient data to make diagnoses and obtain a quick analysis of the decision.

The Radial Basis Function (RBF) neural network was examined by researchers in [20] for its use in general-purpose supervised feed forward neural networks. It's flexible and uses fewer locally modified components than other systems. It was shown that the RBF neural network outperformed both the commonly used Multilayer Perceptron (MLP) network model and the classic logistic regression when using Wisconsin breast cancer data. Logistic regression was shown to have lower predictive power than both neural network models with high sensitivity and specificity. The neural network models outperformed logistic regression on a different dataset as well. Research demonstrates that RBF's neural network has superior prediction abilities and requires less time to run. There are certain downsides of RBF, including its sensitivity to the number of dimensions as well as its inability to handle large datasets.

Three different machine learning methods were used across four different healthcare datasets in [21] to examine the performance of several data mining classification methodologies. The criteria employed are the accuracy and error rate percentages for each classification approach. The 10 fold cross validation technique is used to conduct the trials. Using a dataset as a guide, the strategy with the highest accuracy and lowest error rate is selected. Results show that various classification algorithms respond differently to distinct datasets, depending on the kind and magnitude of their features. A dataset's classification strategy with the highest accuracy and lowest error rate was selected as the best classification approach.

Using diabetes forecasting as an example, [22] created a Decision Support System that makes advantage of the strengths of both OLAP and Data Mining in order to anticipate future conditions and give relevant information for

optimal decision making. It also compared the ID3 and C4.5 decision tree algorithms' outputs.

A decision tree classifier-CART was tested on multiple breast cancer datasets to see how well it performed with and without feature selection in terms of accuracy, model creation time, and tree size. Preprocessing, such as feature selection, significantly improved classification accuracy, according to the research. There was a significant improvement in classifier accuracy when any of the feature selection approaches were used. Three different breast cancer datasets were subjected to a variety of tests to see if the same feature selection strategy could get the best results across the three datasets. The study's findings suggest that for some breast cancer datasets, a particular feature selection may not yield the best results. The number of attributes, kinds of attributes, and occurrences define the best feature selection strategy for a particular dataset. Since each new dataset must be examined, it's necessary to try out a variety of feature selection procedures before settling on the most effective one for improving classifier performance. Once a dataset's optimum feature selection strategy is identified, it may be used to increase the accuracy of the classifier. [23]

3. Methodology

Methodology consists of three machine learning algorithm. These algorithms are KNN, support vector machine and naïve bayes.

The K-NN classifier has been used in several researches to sort data. Numerous algorithms are employed in pattern recognition to sort out the items and classify them. K-NN is a classification algorithm that uses the closest training samples to classify objects. K-NN is a kind of event-based learning. When utilizing a locally estimated function, calculations are delayed until after classification. This study was carried out by [24]. The KNN classification strategy is the most straightforward when there is minimal prior knowledge of the data's distribution. KNN is a well-known pattern recognition classification technique. There have been several studies done on various data sets that show outstanding outcomes when using the KNN calculation.

When K is 1, the Nearest Neighbor (NN) rule is the most basic sort of KNN rule. To make use of this method, all samples must be grouped together according to how similar they are to one another. Using this method, it is possible to make an educated guess about a sample's categorization if the classification of its closest neighbors is unknown. A training set and a query sample can be used to calculate the distance between samples from the training set and those from the query set. As a consequence, the unknown sample's identity may be determined by comparing it to its categorization.

Each data point in an SVM model is represented by a point in k-dimensional space (where k is number of features). The value of each coordinate is used as the feature's value. Choosing a suitable hyperplane that clearly differentiates the classes is generally followed by categorization. SVM was first developed by Vapnik and has since caught the interest of scientists all around the world. Data is collected and divided into two distinct classes by an SVM classifier, generally speaking. A model is generated for the classification of test data once the classifier has been trained on some training data. Occasionally, a problem referred to as "multiclass classification" would arise. For this, several binary classifiers will be required. Studies on the use of SVM in classification have shown that it is more accurate than competing algorithms in terms of determining classifications [25]. SVMs outperform various traditional classifiers in experiments, according to the results. However, the SVM's performance varies widely depending on the dataset and the settings for the cost and kernel parameters. There are several kernel functions in this algorithm: There are three types of radial basis functions: polynomial, linear, and gaussian. (4) A sigmoid or tangent kernel is required.

The Naive Bayes approach is a simple methodology for choosing problem occurrence class labels from feature value vectors while building classifier models. Instead of relying on a single methodology, a variety of approaches based on the same basic concept are used to train these classifiers. There is no way to tell if

one feature is more important than another if we only have the class variable to go on [26].
When learning under supervision, it is possible to train naive bayes classifiers for certain types of probability models. It's possible to deal with the naive bayes model without using bayesian probability or any other bayesian methods by using the maximum likelihood approach for parameter approximation for naive bayes models in a variety of practical applications. The Bayes theorem is combined with the "naive" assumption of independence between all pairs of attributes in naive Bayes classifiers, which are supervised learning algorithms

4. Result Analysis

The data set used in the experiment is NSL KDD [27]. There are 24 types of attacks in the NSL-KDD dataset, and the data is either tagged as normal or as one of them. Probe, DoS, R2L, and U2R are all types of assaults that fall into one of these four categories.

Table 1: Result Comparison of Classifiers

Performance Metrics	Naïve Bayes	Support Vector Machine	KNN Classifier
Sensitivity	0.85	0.57	0.91
Specificity	0.42	0.99	0.99
Accuracy	0.51	0.89	0.98

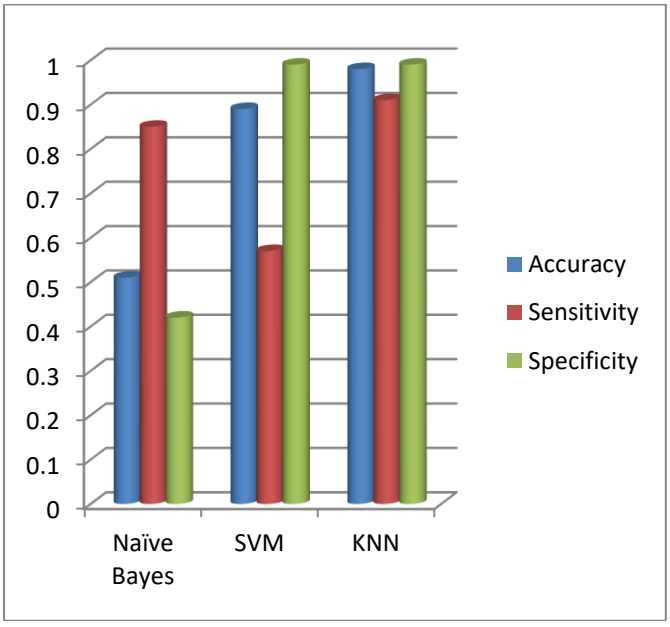

Figure 1: Result Comparison of Classifiers

Table 1 and figure 1 compare the performance of several classifiers. The comparison research utilizes the three dimensions of accuracy, specificity, and sensitivity. When it comes to accuracy, KNN is the clear winner. The specificity of SVM and KNN models is similar, despite their different architectures. When it comes to sensitivity, KNN outperforms SVM

5. Conclusion

Forecasting network traffic is crucial for a variety of reasons, including dynamic bandwidth allocation, network security, and network design. The use of machine learning (ML) techniques to network traffic analysis has sparked a lot of attention. This article describes a method for studying network traffic. The technique is made up of three machine learning-based methodologies. The NSL KDD data set was used in the experimental inquiry. KNN, Support vector machine, and nave bayes are compared in terms of accuracy and other factors.

References

1. D Alconzo Alessandro, Drago Idilio, Mori chetta Andrea, Mellia Marco, Casas Pedr o, A survey on big data for network traffic monitoring and analysis IEEE Trans. Netw. Serv. Manag., 16 (3) (2019), pp. 800-813
2. Shahraki Amin, Taherkordi Amir, Hauge n Øystein, Eliassen Frank, Clustering objectives in wireless sensor networks: A survey and research direction analysis Comput. Netw., 180 (2020), Article 107376
3. Boutaba Raouf, Salahuddin Mohammad A, Limam Noura, Ayoubi Sara, Shahriar Nashid, Estrada-Solano Felipe, Caicedo Oscar M A comprehensive survey on machine learning for networking: evolution, applications and research opportunities J. Internet Serv. Appl., 9 (1) (2018), p. 16
4. Sivarajah Uthayasankar, Kamal Muham mad Mustafa, Irani Zahir, Weerakkody V ishanth Critical analysis of big data challenges and analytical methods J. Bus. Res., 70 (2017), pp. 263-286

5. Shahraki Amin, Geitle Marius, Haugen Ø ystein, A comparative node evaluation model for highly heterogeneous massive-scale internet of things-mist networks Trans. Emerg. Telecommun. Technol., 31 (12) (2020), Article e3924
6. Shahraki Amin, Haugen Øystein An outlier detection method to improve gathered datasets for network behavior analysis in IoT Academy Publisher (2019)
7. Rezaei Shahbaz, Liu Xin, Deep learning for encrypted traffic classification: An overview IEEE Commun. Mag., 57 (5) (2019), pp. 76-81
8. Ucci Daniele, Aniello Leonardo, Baldoni Roberto Survey of machine learning techniques for malware analysis Comput. Secur., 81 (2019), pp. 123-147
9. Conti Mauro, Li QianQian, Maragno Albe rto, Spolaor Riccardo The dark side(-channel) of mobile devices: A survey on network traffic analysis (2017)
10. Fadlullah Zubair Md, Tang Fengxiao, Ma o Bomin, Kato Nei, Akashi Osamu, Inoue Takeru, Mizutani Kimihiro, State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems, IEEE Commun. Surv. Tutor. 19 (4) (2017), pp. 2432-2455
11. Verma Shikhar, Kawamoto Yuichi, Fadlu llah Zubair Md, Nishiyama Hiroki, Kato Nei, A survey on network methodologies for real-time analytics of massive IoT data and open research issues IEEE Commun. Surv. Tutor., 19 (3) (2017), pp. 1457-1477
12. Rashi Bansai, Nishant Gaur & Shailendra Narayan Singh 2016, 'Outlier Detection: Applications and techniques in Data Mining,' IEEE Conference on Cloud System and Big Data Engineering, pp. 373- 377.
13. Christy, A & Meera Gandhi, G 2015, 'Cluster Based Outlier Detection Algorithm for Healthcare Data', Elsevier, vol. 50, pp. 209-215.
14. Manish Gupta, Jing Gao, Charu C Aggarwal & Jiawei Han 2014, 'Outlier Detection for Temporal Data: A Survey', IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2250-2267.
15. Xiaodong Zhu, Ji Zhang, Hongzhou Li, Philippe Fournier-Viger, Jerry Chun-Wei Lin & Liang Chang 2017, 'FRIOD: A Deeply Integrated Feature-Rich Interactive System for Effective and Efficient Outlier Detection', Access IEEE, vol. 5, pp. 25682-25695.
16. Guojun Gan & Michael Kwok-PoNg 2017, 'k-means clustering with outlier removal', Pattern Recognition Letters, vol. 90, pp. 8-14.
17. Erhan Guven & Anna L Buczak 2013, 'An OpenCL Framework for Fuzzy Associative Classification and its Application to Disease Prediction', Procedia Computer Science, vol. 20, pp. 362-367.
18. Gyorgy J Simon, Caraballo, J, Terry M Therneau, Steven S Cha, Regina Castro, M & Peter W Li 2015, 'Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus', IEEE Transactions on Knowledge And Data Engineering, vol. 27, no. 1, pp. 130-141.
19. Sanz, J, Galar, M, Jurio, A, Brugos, A, Pagola, M & Bustince, H 2014, 'Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system', Appl. Soft Computing, vol. 20, pp. 103-111.
20. Padmavathi, J 2011, 'A Comparative study on Breast Cancer Prediction Using RBF and MLP', International Journal of Scientific & Engineering Research, vol. 2, no. 1, ISSN 2229-5518
21. Shelly Gupta, Dharminder Kumar &Anand Sharma 2011, 'Data mining Classification techniques applied for breast cancer diagnosis and prognosis', Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166, vol. 2 no. 2.
22. Rupa Bagdi & Pramod Patil 2012, 'Diagnosis of Diabetes Using OLAP and Data Mining Integration', International

Journal of Computer Science & Communication Networks, vol. 2, no. 3, pp. 314- 322.
23. Lavanya, D & Usha Rani, K 2011, 'Analysis of Feature Selection With Classfication: Breast Cancer Datasets', Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166, vol. 2, no. 5 .
24. Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient k NN classification algorithm for big data. *Neurocomputing*, *195*, 143-148. doi: 10.1016/j.neucom.2015.08.112
25. Mondal, D., Kole, D., & Roy, K. (2017). Gradation of yellow mosaic virus disease of okra and bitter gourd based on entropy based binning and Naive Bayes classifier after identification of leaves. *Computers And Electronics In Agriculture*, *142*, 485-493. doi: 10.1016/j.compag.2017.11.024
26. Suryawati, E., Pardede, H., Zilvan, V., Ramdan, A., Krisnandi, D., & Heryana, A. et al. (2021). Unsupervised feature learning-based encoder and adversarial networks. *Journal Of Big Data*, *8*(1). doi: 10.1186/s40537-021-00508-9
27. http://nsl.cs.unb.ca