Study Traffic and Travel Time Prediction using Machine Learning

Anusha Shetty, Iqra Shaikh

Student, Student,

Bachelor of Science in Information Technology (Bsc.IT)
Bunts Sangha Mumbai's Anna Leela College , Kurla , India.

Abstract : This study has been undertaken to investigate the determinants of time taken to travel from one point to another using machine learning. Travel time prediction has important applicationswithin the field of intelligent transportation, such as vehicle Routing, congestion and traffic management. A challenging task in travel time prediction is obtaining data that is not readily available, as a clear majority of links in roads network are not equipped with traffic sensors.

Index Terms: software defined networking, machine learning, deep learning, traffic classification, traffic prediction.

I. INTRODUCTION:

The population is increasing rapidly with this we need faster and time efficient travelling services. Time prediction can save time, energy as well it can also be helpful for less emission of greenhouse gas such as co2 (carbon dioxide). Industrialists, Export services, and even daily office bearers need accurate travel time prediction, delay or wrong prediction can sometime cause disastrous outcomes accurate prediction can avoid traffic congestion, and loss of high expenditure due to traffic. The main thing needed for prediction is collection of data for a period of time .The sources can be radiators sensors, GPS locators etc. there are many traditional and new upcoming methods which can be used to predict travelling time. One of the method which is used is K nearest neighbor clustering model for short term traffic forecasting in which higher importance was given to the recently collected data to avoid any issues. Support vector regression, linear regression are some the other methods which are used for data collection. Artificial neural network (ANN) uses GPS (global positioning system) to determine time and distance to calculate time that would be needed to travel. While the Google maps use, gps and also locates the autonomous vehicles and also the different routes and least distances which would save time effectively. Deep learning is actually preferred by many localities, industrialists and many academic people because of its wide spectrums as it has image processing and detection with wide range of local languages processing. They process large amount of data in a specific time interval. Traffic prediction needs to have a vast spectrum it is not easy to calculate using these methods as it has more complexity. And it has to be accurate most of the time. We bound to use Genetic, Deep Learning, Image Processing, Machine Learning and also Soft Computing algorithms for prediction of traffic flow since a lot of journals and research paper suggests that they work efficiently when it comes to large amount Of data.

AbbreviationsAndAcronyms:

ML: Machine Learning, DL: Deep Learning, Gps: Global Positioning System, IT: Information Technology, ANN: Artificial Neural Network, CNN: Convolutional Neural Network, AE: Auto Encoders, SDN: Software Defined Network MLP: multilayer Perceptron, SVM: Support Vector Machines.

II. METHODOLOGY :

The methodology used here is Secondary data collection and analysis where we studied the IEEE research papers published by different researchers issued between January 2012 to May2020 and This paper focuses on the concepts of deep learning and machine learning, its basic and advanced architectures, techniques, motivational aspects, characteristics in traffic prediction The paper also presents the major differences between the deep learning, classical machine learning and conventional learn-In approaches and the major challenges ahead in traffic prediction.

III. BACKGROUND

3.1.Traditional methods to predict traffic using machine learning:

Machine learning studies specific pattern to understand any concept and after getting a pattern out of it it trains itself to learn that pattern and give out predictions. There are different supervised and noon supervised methods which can be used to determine traffic. The methods which are used are as follows:

a) Linear Regression: Regression is a technique used to model and analyse the relationships between variables and often times how they contribute and are related to producing a particular outcome together.(supervised)The coefficients of regression lines are functions of current time and prediction horizons. Therefore, the following polynomial is used to predict travel time at t + h, where t is current time and h is the prediction horizon :$T(t+h) = \alpha(t,h) + \beta(t,h) \times T(t)$ [4].

b) Logistic Regression: The intended method for this function is that it will select the features by importance and you can just save them as is own features data frame and directly implemented into tuned model[1].(supervised)

c) Support Vector Machines: data item is plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well [1].(supervised)

d) Naïve Bayes Classification: It is a probabilistic classifier that makes classifications using the Maximum a Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network [2].

e) Ordinary Least Square Regression: The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values. This minimization leads to the estimators of the parameters of the mode [2].

f) K-means k- means is a clustering algorithm that tries to partition a set of points into K sets (clusters) such that the points in each cluster tend to be near each other. It is unsupervised because the points have no external classification [2].K-nearest neighbours is a classification (or regression) algorithm that in order to determine the classification of a point, combines the classification of the K nearest points. It is supervised because you are trying to classify a point based on the known classification of other points. K-means can be applied to data that has a smaller number of dimensions, is numeric, and is continuous. Such as document clustering, identifying crime-prone areas, customer segmentation, insurance fraud detection, public transport data analysis, clustering of IT alerts. (Supervised/unsupervised)

g) Random forest: This method provides in general a good predictive performance, low over fitting, and easy interpretability. They are highly accurate. They generalize better. They are interpretable [3].

3.2.Deep learning algorithms:

Machine learning cannot deal huge of data efficiently as compared to deep learning, so if we want to deal with huge amount of data with higher complexities with real-time outputs. While using machine learning data features of object is fed manually. But in deep learning it self studies the objects prominent features and tries to generate a pattern out of it. Manual interruption is less in deep learning. This characteristic in deep learning helps it predict and solve problems. It handles a large amount of structured and unstructured data .deep learning can also be used in translations .hence making the predictions easy to understand by the local people. The algorithms which can be used to generate predictions in deep learning are as follows:

a) Artificial neural network (ANN): The neurons in human body is responsible for the reflexes or reactions in a same way ANN contains neural nodes the data from single nodes and generates a summation and then reacts to it as it contains artificial neuron like nodes it is termed as artificial neuron network it has a activation function which generates a particular result/action [3].

b)Convolutional neural network(CNN): Neural networks learn to transform input data such as images into successive layers of increasingly meaningful and complex representations.it can be used in image and video recognition, recommender systems and natural language processing[4].(Supervised).

c) Auto encoders (AE): Auto encoders use this input data as an input and output it is a type of artificial neural network, it used to make data learning efficient in an unsupervised manner. It can be used to reduce dimension. It has three nodes visible input nodes, visible output nodes and encoding/ decoding hidden nodes.

3.3.Software Defined Networking:

As a new network architecture, software defined network (SDN) separates the control plane from the forwarding plane which enables administrators to define and control the network through the method of software programming, provides a new research direction for the next generation of network architecture. At the same time, the machine learning technology has been developed rapidly in recent years and some studies have begun to introduce machine learning methods into SDN to improve the efficiency of network management and conformity, or to solve problems that cannot be solved easily by traditional methods. The paper analyses, summarizes and introduces these researches which used the supervised learning, unsupervised learning or semi-supervised learning methods to solve some specific problems on SDN, and it will help later researchers understand the filed more quickly and promote the development of the machine learning technology in SDN.

TRAFFIC PREDICTION:

The objective in Traffic prediction is to estimate future clog in the system utilizing recorded or potentially continuous traffic information. In expansion to traffic order, traffic forecast likewise plays a noteworthy job in Dissecting the traffic stream to forecast traffic before getting blocked. The consistency of system clog is wanted so as to give and keep up top notch organize correspondence since dependent on the result of the broke down traffic information, the SDN controller guides the streams to the less clogged

connections. Previously basic deep belief architecture and the deep CanWest used in the training process. Then DL based prediction algorithm was coupled with a DL based channel assignment algorithm to intelligently route the traffic [2]. Mestres et. Al, [16] examined neural systems to demonstrate the delays in the systems. They prepared different neural system models under different situations with various basic organize parts, for example, geography, arrange size, traffic force .What's more, steering so as to define guidelines about preparing such neural systems.

1. Traffic Classification:

Traffic classification is crucial in optimizing internet access and user experience. Since the available bandwidth is limited, by classifying traffic we make the best use of the bandwidth and internet service providers can manage the resources by prioritizing the flow of packets. Traffic classification can be achieved by identifying the network applications or group of applications. One of the basic approaches is port-based. However it is not practiced anymore due to unsatisfactory classification results since modern applications run on dynamic ports. The alternative to port based is payload-based approach which is often referred to as deep packet inspection (DPI). DPI identifies the application by inspecting the content of the packet and yields better classification results. Nonetheless, it introduces several challenges. First, it consumes resources since the packets are treated as stacks and identifying a pattern within a packet is computationally expensive. Second, it cannot recognize encrypted traffic, which is quite prevalent these days. Hence, flow-based approaches using ML and DL are used to overcome the limitations of classification. ML methods try to detect patterns within the applications based on the selected feature sets. They can classify the encrypted traffic and work with a lower computational cost. Table II, shows a summary of the surveyed papers for traffic classification. Xiao et al. [10] presented a low cost learning method to catch elephant flows in real time. The proposed strategy includes two-stage elephant flow detection. At first, suspicious elephant flows are distinguished from mice flows. At the second stage, after using a feature selection module, a correlation-based filter creates the optimum features in the dataset. At that point elephant streams are utilized for improving the characterization exactness while choice trees arrange them as genuine elephant streams or dubious streams. Da Silva et al., [8] introduced a structure which identifies assault based traffic, classifies traffic oddities by utilizing ML calculations, for example K-means and SVM. After the classification, they performed specific activities dependent on the data.
Gathered. This moderation technique incorporates activities such as forestalling the traffic of dubious streams.
Wang et al., [9] proposed a plan for SDN which sorts traffic continuously dependent on classes with various QoS conditions. So as to accomplish higher exactness in order, DPI and semi-regulated ML; i.e., Laplacian SVM also, k-implies approaches, were utilized together. Wang et al., [7] planned a plan for supporting the SDN controller continuously traffic observing and characterization of scrambled traffic streams by executing MLP, stacked AE, and CNN. They first preprocessed the information and the subsequent Meta information was taken care of back to make the proposed structure.

IV. EXPERIMENTATION SURVEY STUDY:

This process was done by US Department of Transportation and carried with the dataset. It provides Open Data Policy and publicly available for all. It is useful for the various stations of US and consist of daily volumes of traffic, binned in fashion hourly [3]. Here there are different facilities provided such as information on flow direction and sensor placement on other key is used for the same. There are two types of datasets available they are primary datasets and secondary datasets. In primary datasets it has various stations ids, location information such as latitude and longitude, flow directions for traffic control s that it can be used for the users to judge the right place to choose for travelling. Now the secondary dataset consist of the deep information about the individual stations and observe it. Both the datasets have huge information about the traffic control across multiple stations which are spreader over multiple stations into different states. Proper specification is done through the datasets within the location given. For this the attributes of string types was encoded into numeric. Also, the detailed description to codes were removed. Example, functional classification of road types rural or urban sub category information was one of the key attributes which was pre-prepared and processed to make it usable to apply machine learning model. At the same time, the station id was processed. Except the critical attributes such as, date, direction of flow and volume were numeric and hence no explicit transformations were made. To explore a geo location area both primary and secondary datasets are merged upon to select state codes it shows the visual feel of traffic flow. Also additional filters were created with specific range of time period. Basically, it was created to analyze the congestion points and to generate a heat map. Hence, the traffic volume attribute had to be enumerated with its corresponding latitude and longitudinal information. This process was done to judge the individual data points to generate matrix structured transformed data for iterating them. To visualize python transformed data a folium named wrapper was used for the capability. It uses the leaflet.js library to mark manipulated points on maps such as open street map, map box and stamen. Enumeration transformation was carried out to make the data points acceptable by folium to be visualize over open street maps. Heat map functions of folium made it straight forward get results integrated within Jupiter notebook. As the proposal was to utilize the machine learning models to predict traffic flow, the dataset was split into train and test parts with seventy thirty ratio respectively. The entire datasets was randomly splinted and made such that one part of it contained seventy percentage of the data points and other part contained the remaining thirty percent. Support Vector regression based on LIBSVM (an integration application for support vector classification) was firstly evaluated. In SVM regression, the input is first mapped into an m-dimensional notation. The performance for the model was evaluated using Mean Squared error method. Linear regression model was built to scalar response and the independent variables. Decision tree learning was explored and the algorithm was used to generate a decision tree from the datasets. Above all through this algorithms any of the one can use the appropriate specifications.

V. RESULTS AND DISCUSSION :

The study of research shows the result in identifying an optimal model to the publicly available datasets. Random forest is a part of an ensemble learning method. It can be operated by constructing multitude of decision trees at training time. Individual trees are predicted by mean outputting the class. In the other evaluation metric cross validation score was generated for the different models which are discussed. Quite apart, a technique which involves reserving a particular sample of a dataset on which the model was not trained. Later, the model is tested on this reserved sample before finalizing it. The effectiveness of the models performances

helps in gauge. The K-fold cross validation was followed, with ten as its parameter. To use decision trees random forest is the best way of over fitting to their training sets. Whereas, deep learning and the algorithm is an important problem in data analysis. Extensively Machine learning does not belong to a dealt community. The underlying domain expertise and data distribution plays a very important role in picking up approaches in solving problems which are data driven. The framework can be iterated again by modifying the dataset. Also the proposal is to take at large the solution to different geo locations to prove its efficiency. Further research is expected to make the model generic enough to get it integrated with existing agencies in solving the traffic problem in real time [3].

Conclusion:

In this article, we studied the ML and DL techniques utilized for traffic forecast, next, we summed up the current works. We at that point introduced our study lastly we tended to the difficulties and future work that are dataset qualities, information volume, strategy of applying DL Since utilizing ML and DL calculations for traffic forecast is very new, more issues may be recognized practically speaking that can't be anticipated at this point

Acknowledgment:

Foremost we would like to thank our Professor guide, “Mrs. Rupa Patel Karnik” then we would like to thank the institution we are studying in “Bunts Sangha’s Annaleela Shobha Jayaram College” for always motivating us. Then we would like to propose a thanks to “IICCTCMS 2020”, for giving us an opportunity to participate and present our paper.

References

[1] Gaurav Meena ,Deepanjali Sharma ,Mehul Mahrishi 07-08 February 2020, Traffic Prediction for Intelligent Transportation System using Machine Learning, IEEE Conference Record 48199.

[2] Ays¸e Rumeysa Mohammed, Shady A. Mohammed, and Shervin Shirmohammadi, Machine Learning and Deep Learning Based Traffic Classification and Prediction in Software Defined Networking.

[3] Andrew Moses, Parvathi R, Vehicular Traffic analysis and prediction using Machine learning algorithms.

[4] F Guo, R Krishnan, J W Polak, Short-term traffic prediction under normal and incident conditions using singular spectrum analysis and the k-nearest neighbor method IET.

[5] Yi-ming Xing Sch. of Comput. & Commun. Eng., Univ. of Sci. & Technol. Beijing Beijing, Beijing, China
; Xiao-juan Ban; Ruoyi Liu, A Short-Term Traffic Flow Prediction Method Based on Kernel Extreme Learning Machine.

[6] 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.

[7] P. Wang, F. Ye, X. Chen, and Y. Qian, “Datanet: Deep learning based encrypted network traffic classification in sdn home gateway,” IEEE Access, vol. 6, pp. 55 380–55 391, 2018**.**

[8] A. S. da Silva, J. A. Wickboldt, L. Z. Granville, and A. Schaeffer-Filho, “Atlantic: A framework for anomaly traffic detection, classification, and mitigation in sdn,” in NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2016, pp. 27–35**.**

[9] P. Wang, S.-C. Lin, and M. Luo, “A framework for qos-aware traffic classification using semi-supervised machine learning in sdns,” in 2016 IEEE International Conference on Services Computing (SCC). IEEE, 2016, pp. 760–765.

[10] P. Xiao, W. Qu, H. Qi, Y. Xu, and Z. Li, “An efficient elephant flow detection with cost-sensitive in sdn,” in 2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom). IEEE, 2015, pp. 24–28.

[11] J. Suarez-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.

[12] 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.

[13] M. Amiri, H. Al Osman, and S. Shirmohammadi, “Game-aware and sdn-assisted bandwidth allocation for data center networks,” in 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018, pp. 86–91.

[14] Z. Fan and R. Liu, “Investigation of machine learning based network traffic classification,” in 2017 International Symposium on Wireless Communication Systems (ISWCS). IEEE, 2017, pp. 1–6.

[15] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, and K. Han, “Enhanced network anomaly detection based on deep neural networks,” IEEE Access, vol. 6, pp. 48 231–48 246, 2018.

[16]A. Mestres, E. Alarcon, Y. Ji, and A. Cabellos-Aparicio, “Understanding ´the modeling of computer network delays using neural networks,” in Proceedings of the 2018 Workshop on Big Data Analytics and MachineLearning for Data Communication Networks. ACM, 2018, pp. 46–52.