Vehicular Traffic analysis and prediction using Machine learning algorithms

Andrew Moses
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
andrew2moses@gmail.com

Parvathi R
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
parvathyraman@gmail.com

***Abstract*— With day to day computing becoming difficult. Lot of time and energy is lost in traversing through vehicular traffic. In this paper we propose a solution to predict vehicular traffic using machine learning techniques. US traffic 2015 dataset contains daily volumes of traffic, since this is an actual data and not a simulated one, patterns found in this research can be benefited and made valuable with differing datasets over time. Data driven solutions are proven to breakthrough tough problem statements. Thus in this paper we propose stage by stage machine learning processes to build an efficient model capable of predicting traffic volume based on features which brings out hidden insights in vehicular movements. Inspired by existing research exploration over traffic, in this paper we share our contribution of making traffic predictable thus saving time and energy in day to day life.**

Keywords— vehicular traffic, machine learning, prediction, data analysis, pattern discovery

I. INTRODUCTION

Urban population is getting increased at an exponential pace. Vehicular traffic has become a huge barrier in almost every industry directly or indirectly. To certain extent it even influences the economy of various markets. Individually it can be seen as seemingly insignificant, but definitely it cannot be ignored when it affects business meetings and unplanned disasters at high profile events. Thus research efforts are put in solving this significant problem, at least a step closer to solving.

A. Machine learning Techniques

Machine learning (ML) is being utilized extensively is many fields and the critical factor which decides an efficient model is the data used at its base. Digital techniques have improved traffic management and this facilitates digital data of vehicular movement. High amount of data production empowers machine learning methods. Also the increase in computation power and re-engineered algorithms have paved a way to capture hidden insights from high volumes of data.

This paradigm is used as a shift from explicitly programmed application software. Thus, the ML solutions are data driven and utilizes algorithms which allows the application to predict outcomes based on the input data fed to the model. Varying patterns in data would eventually result in gathering insights which were never seen before.

B. Machine Learning Advantages

Broadly ML can be categorized into two forms: one is where the algorithms learn from data and can predict unseen(futuristic) data. While the other type of algorithms work in grouping up the given data by identifying their types (or) similarities.

On having a closer look at ML algorithms, regression pops up often. Regression is concerned with the relationship between features of the data. Iteratively the model is refined using a measure of error in the predictions. Thus, features contributing towards the target and its amount of correlation with the target makes predictions more aligned with the data. The most popular regression algorithm includes Least Squares regression, Linear regression, Logistic regression, stepwise regression, multivariate adaptive regression splines, Locally estimated scatterplot smoothing, etc. The list goes on.

Another group of ML algorithms focuses on instances of training data. The prominent data points are explored and using similarity measure new data points is labeled to the nearest proximity class. It is more concern on how the training data is represented in the database. The most popular algortihms are K-nearest neighbor, Learning Vector Quantization, Self- organizing map, Locally weighted learning, Support Vector machines, etc.

Decision tree algorithms builds up models with conditions on values of specific attributes of the given data. The input data is represented in tree structure with decisions created based on instances of attributes' values leading to target feature spread out at the leaf nodes. It is useful for both classification and continuous data type problems. Decision trees are deemed to be quick and more accurate. With the decisions derived, it can be translated into rules making insights transparent and more understandable. The following alogrithms are few examples of decision tree types: Classification and regression tree (CART), Iterative dichotomiser 3, C4.5 and C5.0, Chi- squared automatic interaction detection, decision stump, M5 and Conditional decision trees.

Bayesian methods are set of algorithms which utilized bayes' theorem at its base and works on probability principles. Naive bayes, Gaussian Naive bayes, Multinomial naive bayes, bayesian belief network and bayesian network.

Clustering algorithms are typcially representing data using centroid and hierarchial methods. The model doesn't require a target feature and input data groups up the data with the specified number of cluster bins. The centroid values are iteratively adjusted with the input data. Commonly used algorithms include: K-means, K-medians, Expectation max- imisation, Hierarchical clustering.

Finally the artificial neural network models inspired by nodes in human brain. Pattern matching mechanism which

978-1-7281-4141-1/$31.00 ©2020 IEEE

tags parameters adjustable with increase in input data. This methodology gets better and better with varying number of parameters that are allowed to be tweaked and increase in datapoints fed to the model. Preceptron, Multilayer preceptrons, back-propagation, stochastic gradient descent, hopfield network, radial basis function network are few examples of artifical neural network algorithms.

II. LITERATURE SURVEY

Previous research works and its aspect towards traffic problems were reviewed. To start with approaches like Intelligent Transport Information System[1],[5],[6]: vehicular movement were captured with dedicated Closed Circuit Television (CCTV) cameras, Variable Message Signs (VMS), vehicle detection stations. These devices were interconnected and managed centrally. Taking another step forward, the analyzed data were derived into report and sent to the public via Short Message Service (SMS). Also data collection were made with probe vehicles equipped with Global Positioning System (GPS) to understand the intensity of traffic at ground level. Unfortunately the above said approach was not well received by the downstream operators like private agencies to integrate it with the public usability. Thus giving a requirement to build a more public oriented system which can be integrated with multiple agencies.

Crowdsourcing based Traffic Information system[3],[7],[8] (CroTIS) was proposed and its architecture followed its plat- form to be put as a base. A place were collaboration was given importance, public would be given the ability to volunteer for the common good. Sensors in users mobile who register themselves into the application would serve critical traffic data which is processed further. The mobile friendly application would exchange real time data with the intelligent server application to get info on congestion and traffic causing events like accidents or road closure.

Geosocial networks was based on the increasing social footprint of the general public [11],[13]. A dedicated crawler which would gather publicly posted information on traffic, events, accidents and road closure data. Processed and auto- mated report generation to understand the traffic movements. It is a data mining software based on collaborative community driven navigation applications.

Open Traffic system[4],[9],[10], an open source platform for analyzing traffic movement. Integrated with open street map, the application provides a platform for research work on traffic analysis to be carried out with open source tools. Both historical and real time data can be analyzed in this system. Mainly consists of traffic engine, traffic data pool and OSM-linked traffic dataset which builds the bases of understanding the traffic movement. The framework proposes estimated arrival time to a said destination with comparisons of alternate routes.

III. EXPERIMENTATION

The proposed research was carried out with the dataset compiled by US Department of Transportation. The dataset is publicly available and meets the requirements of Open Data policy. It consists of daily volumes of traffic across various stations in US, binned in hourly fashion. Information on flow direction and sensor placement are other key info which is provided. The primary dataset consists of the various station ids, location information such as the latitude and longitude, traffic flow direction and the type of road in which the vehicles are travelling. A secondary dataset provides deeper info on individual observation stations. Fig. 1 shows a slice of traffic volume on a specific date during peak hours.

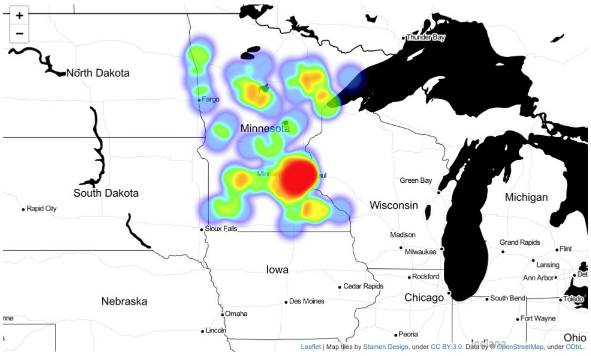

Fig. 1. Folium Heatmap

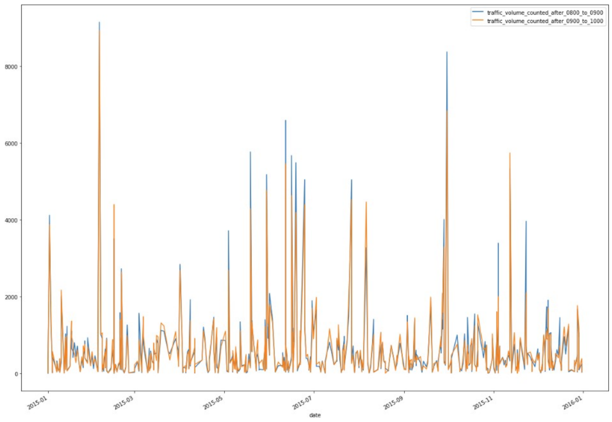

Fig. 2. Traffic Trend

The dataset had huge information about traffic flow across multiple stations and spread across different states. Fig. 2 shows the traffic trend for the entire year 2015 during rush hours. To get an insight, focus was given to specific state and a particular hourly bin to visualize the geographic dataset. Attributes of string types were encoded into numeric. Also, the attributes which served as detailed description to codes were removed. Example, functional classification of road types (i.e.,) rural or urban with sub category information was one of the key attributes which was pre-processed to make it usable to apply machine learning model. Similarly, station id attribute was processed. The critical attributes such as date, direction of flow and volume were numeric and hence no explicit transformations were made.

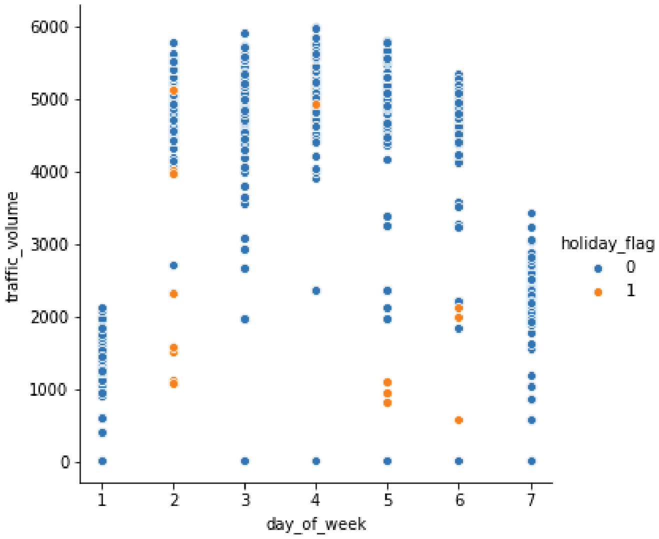

Fig. 3. Distribution of Traffic volume

To get the visual feel of traffic flow, both primary and secondary datasets were merged upon select state codes with focus on exploring a specific geo location area. Also additional filters were created with specific range of time period. The plan was to generate heatmap to analyze the key congestion points. Hence, the traffic volume attribute had to be enumerated with its corresponding latitude and longitude information. This step was carried out by iterating over individual datapoints to generate matrix structured transformed data.

Folium is a wrapper which has the capability of visualizing pythonic transformed data. It utilizes leaflet.js library to mark manipulated points on maps such as openstreet map, mapbox and stamen. The above enumeration transformation was car- ried out to make the datapoints acceptable by folium to be visualized over openstreet map. Heatmap functions of folium made it straightforward get results integrated within jupyter notebook.

Distribution of data with respect to various attributes were visualized to get deeper understanding about the raw data- points. Fig. 3 shows the distribution of raw data with respect to holidays. In order to get the entire year's traffic movement, hourly bins of rush hours were filtered and visualized with multiple overlapping line graphs. With the intuition of differential traffic movement during holiday time, a list of holiday dates were included and marked, to flag out the dates in the analyzed dataset. Scatter plot visualization of holidays (flagged data point) spread across days of the week gave the deeper understanding of the intuition made earlier.

As our 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 dataset was randomized and a split was made such that one part of it contained seventy percentage of the data points and the other part contained the remaining thirty percent. This step is always carried out to evaluated the built model. The thirty percent test part would not be utilized to train the model, hence its like the unseen data points to the models which is going to be built.

Support Vector regression based on LIBSVM (an integration application for support vector classification) was firstly evaluated. In SVM regression, the input is first mapped onto a m-dimensional feature space using nonlinear mapping, and the linear model is constructed in this feature space. Using mathematical notation, the linear model (in the feature space) is shown in equation (1).

$$f(x,\infty)=\sum_{j=1}^{m}\infty_j g_j(x)+b \qquad (1)$$

where,

$f(x,\infty)$ = Feature space

$g_j(x)$ = set of nonlinear transformations

b = bias

The model was trained and evaluated. The performance of the model was evaluated using Mean Squared error method. It is an estimator procedure to measure the average squares of the errors. A risk function corresponding to the expected value of the squared error loss, as shown in equation (2).

$$MSE = {}^{1}\!/_{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \qquad (2)$$

where,

n = predictions generated from the training data points.

Y_i = vector of observed values of the variable being predicted.

$\hat{Y}_i$ = predicted values.

Linear regression model was built to scale the relation between the scalar response and the independent variables. The model depend linearly on their unknown parameters, the resulting estimators are determined, as shown in equation (3).

$$y_i=\beta_0+\beta_1x_{i1}+\ldots+\beta_px_{ip}+\varepsilon_i=x_i^T\beta+\varepsilon_i \quad (3)$$

where,

i = 1,...,n

$x_i^T\beta$ = inner product between vectors

Decision tree learning was explored. The algorithm was used to generate a decision tree from the dataset. It iterates through the attributes of datapoints to calculates the entropy or the information gain. Then the appropriate attribute with smallest entropy or largest information gain value is selected. Throughout this iterative process, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch, as shown in equation (4).

IV. RESULT AND ANALYSIS

An ensemble learning method - Random forest. Operates by constructing a multitude of decision trees at training time and outputting the class that is the mean prediction of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. The general technique of bootstrap aggregating is applied to the training dataset.

Another evaluation metric - Cross validation score was generated for the different models which are discussed above. 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. It helps in gauge in the effectiveness of the model's performance. K-fold cross validation was followed, with ten as its parameter value. This means the dataset was randomly split into ten folds.

As model's performance started to change with varying hyper parameters. Optimal hyper parameter discovery was focused. Grid search methodology was utilized to overcome this hurdle. It is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. Grid search was extensively used to fine tune the random forest model. Parameters like number of trees generated in the forest, number of features to be considers for the decision splits and to build each tree with entire dataset without bootstrap were considered with varying values. Grid search process yielded the optimal best parameters which were used further for model evaluation, as described in Table 1.

TABLE I. COMPARISON BETWEEN MACHINE LEARNING MODELS

Model	Parameters	MSE[a]
Support Vector Regression	Gamma='scale', C=1.0, epsilon=0.2	871.03
Linear Regression	Fit intercept=True	782.48

Decision Tree Regression	Criterion='mse',
	Min_samples_split=2,
	splitter='best'
		352.67

Random forest	Max features='auto', n estimators=10, min samples split=2	312.16
Random forest (with optimal parameters using Grid search)	Max features=9, n estimators=60	287.60

[a] Mean Squared Error.

V. CONCLUSION

It is evident that varying algorithms produce vivid results. The underlying domain expertise and data distribution plays key roles in picking approaches in solving problems which are data driven. This research has resulted in identifying an optimal model to the publicly available dataset. 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 real time.

REFERENCES

[1] World Combining Taxi GPS Data and Open-Source Software for Evidence-Based Traffic Management and Planning, September 2015.

[2] Vehicular Traffic Analysis From Social Media Data. International Con- ference on Advances in Computing, Communications and Informatics (ICACCI'16), September 2016.

[3] Urban Traffic Analysis Using Social Media Data on the Cloud. Interna- tional Conference on Utility and Cloud Computing Companion (UCC Companion), December 2018.

[4] Road Traffic Prediction Using Artificial Neural Networks. 2018 South- Eastern European Design Automation, Computer Engineering, Com- puter Networks and Society Media Conference (SEEDA CECNSM), November 2018.

[5] City traffic prediction based on real-time traffic information for Intelli- gent Transport Systems. ResearchGate, November 2013.

[6] Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs. The National Center for Biotechnology Information, May 2019.

[7] World Health Organization Road Traffic Injuries. [(accessed on 27 November 2018)]; Available online: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[8] El Hatri C., Boumhidi J. Traffic management model for vehicle re-routing and traffic light control based on Multi-Objective Particle Swarm Optimization. Intell. Decis. Technol. 2017;11:199–208. doi: 10.3233/IDT-170288

[9] Chrobok R., Wahle J., Schreckenberg M. Traffic forecast using simula- tions of large scale networks; Proceedings of the ITSC 2001, 2001 IEEE Intelligent Transportation Systems, Proceedings (Cat. No. 01TH8585); Oakland, CA, USA. 25–29 August 2001; pp. 434–439

[10] Smith B.L., Demetsky M.J. Traffic flow forecasting: Comparison of modeling approaches. J. Transp. Eng. 1997;123:261–266. doi: 10.1061/(ASCE)0733-947X(1997)123:4(261)

[11] Prediction based traffic management in a metropolitan area. Journal of Traffic and Transportation Engineering (English Edition), July 2019

[12] Chavhan and Venkataram, 2019a S. Chavhan, P. Venkataram Emergent intelligence: a novel computational intelligence technique to solve problems The 11th International Conference on Agents and Artificial Intelligence, Prague, 2019

[13] He and Peeta, 2015 X. He, S. Peeta A marginal utility day-to-day traffic evolution model based on one-step strategic thinking Transportation Research Part B: Methodological, 84 (2015), pp. 237-255

[14] Analysis of vehicular traffic flow in the major areas of Kuala Lumpur uti- lizing open-traffic. AIP Conference Proceedings 1883, 020013. Septem- ber 2017