KSII The 14th Asia Pacific International Conference on Information Science and Technology(APIC-IST) 2019.

Traffic Data Analysis and Prediction using Big Data

Dalyapraz Dauletbak and Jongwook Woo
Department of Information Systems, California State University Los Angeles, US
[e-mail: dmanato@calstatela.edu, jwoo5@exchange.calstatela.edu]
*Corresponding author: Dalyapraz Dauletbak

Abstract

The paper adopts traffic dataset to analyze and predict the traffic patterns in Los Angeles County. Data is adopted from a popular platform in the USA for tracking information on the road using the device information and reports shared by the users. The dataset mainly consists of information about jams, traffic incidents, road closure, hazard reported from the application users. The major contribution of this paper is to give a clear view of how the large-scale traffic data set can be stored and processed using the Big Data systems: Hadoop and its ecosystem (Hive). Furthermore, the predictive analysis with the classification models is presented using Azure ML Studio for the sample traffic data set. The process of modeling, as well as results, are interpreted using metrics: accuracy, precision and recall. Besides, it portrays some visuals and analytics using BI tools.

Keywords: Traffic Analysis, Traffic Prediction, Big Data, Machine Learning

1. Introduction

Traffic prediction is an emerging topic and since governments started adopting smart cities' concept for the past decades, traffic prediction attracts more attention. However, traffic prediction can be divided into two major areas: short-term and long-term traffic prediction [1]. Short-term studies concentrate on predicting traffic conditions on the real-time data, developing precise algorithms to capture speed and time for alternative routes and predicting road traffic several minutes to several hours ahead. Whereas, long-term studies go deeper to understand historical data and predict behavioral traffic conditions for weeks and months.

In this paper we have covered long-term study of traffic jams, specifically in Los Angeles County area. US Governments turn to Advanced Traffic Management Systems in order to solve traffic congestions and adopt new transport management plans and resources [2]. Since US City Departments are interested to use information from popular navigation platforms in order to understand and improve traffics, City of Los Angeles provided traffic data set from one of the famous navigator app companies in the state to us for a research study.

2. Related Work

There is a growing interest in traffic prediction systems to support traffic operators in city's decision-making tasks. In 2006 the U.S. Department of Transportation launched the Integrated Corridor Management (ICM) initiative in order to develop new technologies that can operate to improve transportation corridor [3].

Several widely used navigation platforms are willing to help government improve traffic by participating in numerous studies. Google Cloud and Traffic for London arranged hackathon dedicated to traffic simulation using London traffic data set, where one of the companies came

This study was supported by Isaac Engineering and Oracle Cloud Innovation Accelerator. Oracle Big Data Cloud Service was used in the data analysis.

with data flow for processing, visualizing and predicting traffic speed in London [4]. Our work adopts Big Data and Machine Learning for analysis and prediction of traffic jams in Los Angeles. Our work is different in the way of deliverables, since we are focusing on interactive visuals, depth of information, giving more insights of traffic pattern analysis and prediction of traffic congestions.

Waze company has a special program for those who are interested and willing to connect with it for better community - The Connected Citizens Program (CCP), and through such program partners can exchange data with Waze to make data-driven infrastructure decisions and increase the efficiency of incident response [5]. One of the works that is based on Waze company traffic data is available in the form of slides from Summit on Data-Smart Government at Harvard (November 2017) [6]. This study focuses on collaboration of Waze and Louisville City and points out major insights from such partnership. The outcome of this work is analysis of data in the form of animated maps and Excel tables of hot spot traffic [7,8]. Traffic department of Louisville currently have a sustained flow of data and use it on a daily basis. However, our work, apart from analysis of traffics, also explains the flow of big data files management, dynamic geo-maps and further prediction of traffic jams using machine learning.

Another study was conducted in New Haven County, Connecticut. In this research GPS data set was gathered from MapMyRun traffic website and further processed and analyzed using R [9]. The author used sampled small data set for analysis, whereas we present a framework that can be applied to bigger data sets. Also, this work concentrates on clustering the hot areas of traffic, however, our work gives insights to traffic patterns with interactive geo-maps and prediction of jams using classification model.

3. Dataset and Specifications

The dataset was provided by Information Technology Agency of Los Angeles City Department for study purposes and consisted of 5,858 JSON files covering information reported by app users (accidents, jams, road closure etc.) and information captured from users' devices (location, speed, time deviation from original route). Since this database is not publicly open and data is shared upon request only, we were authorized to use a portion of the data only. The dataset is of the size 1.8 GB and covers nine days (Dec 31, 2017 – Jan 8, 2018). However, the data was captured with millisecond difference and is considered a raw dataset from a navigation app. After parsing JSON files into readable CSV format, two major files can be rendered: alerts (information reported by users) and jams (information captured by user's device). Total number of rows (event records) for alerts and jams are 2,170,694 and 16,058,236 rows respectively. The same data processes can be applied to much bigger dataset (as large as 70GB+ annually) as Hadoop systems is linearly scalable.

The below table shows the specification for Oracle cluster we were using for our study.

Table 1. H/W Specification

Number of nodes	6
OCPUs	12
CPU speed	2195.196MHz
Memory	180 GB
Storage	682 GB

4. Method

4.1 Workflow

Initially, the raw dataset, which comprises the details of traffic conditions on specific days in January 2018, was downloaded from a trusted source. The data set is uploaded to Hadoop big data systems and transformed to be analyzed using Hive ecosystems. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel. A large-scale data means a data of giga-bytes or more, which cannot be processed or expensive using traditional computing systems [10]. Hadoop is one of the popular Big Data platform and Hive is one of ecosystems for Big Data analysis. The whole process of date manipulation is shown in the below flowchart (**Fig. 1**).

There are 5858 JSON flat files that has to be parsed into readable tables. This can be done with Python and Pandas library – “pandas.io.json.json_normalize” [11] and further exported in two csv files (alerts and jams). Further, files are uploaded to the Hadoop File System and then HiveQL is used as querying language to create the tables’ schema, clean data, create summary table for analysis and sample dataset for prediction and output the results.

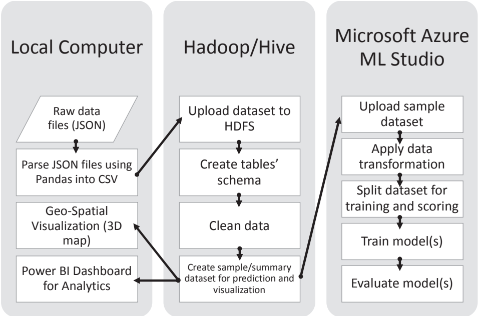
Local Computer
Hadoop/Hive
Microsoft Azure ML Studio
Raw data files (JSON)
Parse JSON files using Pandas into CSV
Geo-Spatial Visualization (3D map)
Power BI Dashboard for Analytics
Upload dataset to HDFS
Create tables’ schema
Clean data
Create sample/summary dataset for prediction and visualization
Upload sample dataset
Apply data transformation
Split dataset for training and scoring
Train model(s)
Evaluate model(s)

Fig. 1. Big Data Architecture for Prediction and Analysis

Once the output files have been downloaded on local machine, Excel’s 3D map and Power BI can be used to obtain the Geo-Spatial visualization of reported traffic events and traffic jams. The sampled 100,000 rows from jams file (traffic information captured by user’s device), which were randomly pulled from the whole dataset are further used for prediction in Microsoft Azure Machine Learning Studio. Traffic jams prediction can be divided into further major steps of uploading the sample dataset, applying data transformation required for accurate modeling, splitting dataset to train a machine learning model and evaluate prediction accuracy. This process will be explained in detail further in this paper.

4.2 Data Cleaning

Parsed files were uploaded and stored in HDFS (Hadoop Distributed File Systems) and then loaded into tables for analysis with Hive using *Beeline* Client. Since alerts and jams files have different information (one has information reported by app users, such as jams, road closure, hazards, car accidents; and the other has information tracked from users’ devices, such as location, speed, time deviation from original route) each was separately cleaned and then exported for further analysis.

Data cleaning was conducted using different techniques such as regular expressions, conditional statements, substrings, tables joining, date and time formatting and time conversion from UTC to PST time zone (Pacific time zone).

After cleaning and removing irrelevant fields, the attributes and metadata of alerts are the following:

Table 2. Alerts attributes

location_x	X-coordinate of location
location_y	Y-coordinate of location
street	Street name
city	City name (LA County has up to 88 cities [12])
country	Country (US);
road_type	Road type (Ex: Street, Primary street, Freeway and etc.)
report_description	Small text describing the traffic event written by user
type	Type of reported traffic event (road_closed, jam, accident, hazard)
pub_date	UTC Time of the publication of traffic report
date_pst	Pacific Time of the publication of traffic report
month	Month number of the publication (1-12)
day	Day of the publication (1-31)
hour	Hour of the publication (0-23)
min	Minute of the publication (0-59)
sec	Second of the publication (0-59)
weekday	Day of the week of the publication (Monday - Sunday)

And the attributes and metadata filtered for jams are the following, which is generated passively from device’s GPS (**pub_date**, **date_pst**, **month**, **day**, **min**, **sec**, **weekday** **location_x** **and**

location_y have the same metadata as mentioned above):

Table 3. Jams attributes

level	jam level, where 1 – almost no jam and 5 – standstill jam
speed	driver's captured speed in mph
length	length of the traffic ahead in the route of user in meters
delay	time deviation from the original time in seconds

In addition, a summary table was created to portray basic information about traffic in a smaller aggregated table. Summary table can give insights about amount of jams by time, days, and level of the traffic jam.

5. Analysis and Visualization

After data cleaning and preparation for further analysis, files were extracted into Microsoft Excel and Power BI. We used different interactive visuals in order to show traffic events (including jams) clearly on the map as well as time dependent patterns and different charts.
The geo-map, (Fig. 2 - 3) was made in add-in Excel tool 3D-map, which can be used for animated map with a timeline. This map shows a sampled day (Friday, Jan 5, 2018) from a full dataset giving an insight into traffic events reported by users and traffic jams captured from the user's device. We used the heat map to show the amount of traffic jams and clustered columns to show the amount of reported accidents (red bar) and reported road closure (yellow bar). By using the time filed, we were able to build up a dynamic geo-map changing over time, showing timeline flow of traffic on a map. Originally, this visualization consists of two 52-second videos, showing the full day span (Friday, Jan 5, 2018) of traffic patterns in LA County and traffic incidents reported by users.

From the geo-map of jams (Fig. 2), we were able to see condensed traffic on highways - 101, 405, 10; Downtown LA - west area (major concentration of business centers); Santa Monica – area close to pier (tourist place); Beverly Hills – along major streets, such as Santa Monica Blvd. Also, the most condensed traffic hours appeared from 3 pm to 6 pm, although morning hours (7 am - 9 am) are heavy as well, with less intension, this can be also seen on Power BI bar chart (Figure 4). Interestingly, heavy hours around the airport appeared from 5 am to 8 am and from 7 pm to 10 pm. Another interesting insight is huge traffic in Topanga on Topanga Canyon Blvd and Tuna Canyon Blvd.

From the map of traffic events reported by users (Fig. 3) we can conclude that users tend to report less traffic jams than their devices can capture. However, most condensed traffic seems to be reported in the same pattern.

The pie chart (Fig. 5) shows percentage portion of traffic jams by days of week. It can be clearly seen that the most congested days are Monday and Friday, while Sunday is least one.

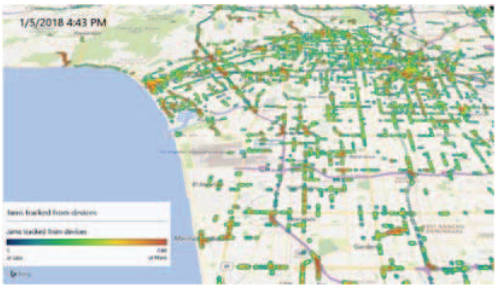

Fig. 2. Jams tracked from users' devices

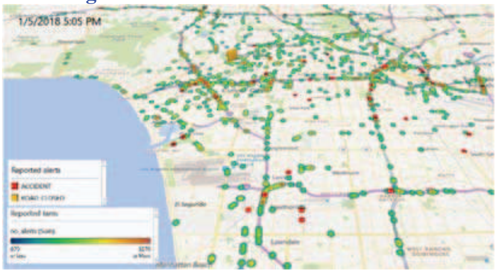

Fig. 3. Jams and other traffic incidents reported by users

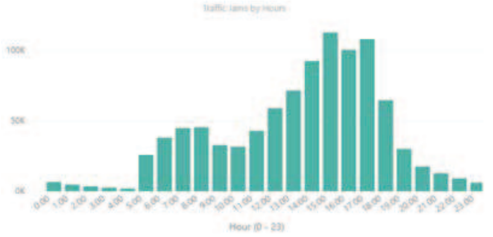

Fig. 4. Traffic Jams by Hours

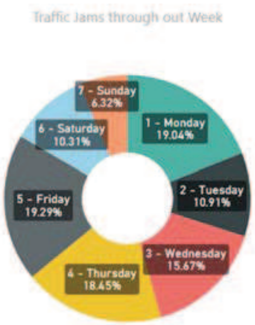

Fig. 5. Traffic Jams by Days of Week

6. Prediction with Machine Learning

6.1 Machine Learning Flow

As mention before, traffic jams dataset, which was passively captured from users' devices GPS, has more than 16 million rows of data. This data is huge for training machine learning model without appropriate high computational speed. Microsoft Azure ML Studio is adopted for predictive analysis and the sampled dataset is uploaded to build a machine learning model, which is a GUI-based integrated environment for constructing Machine Learning workflow [13]. Sampled dataset of size 10 MB (100,000 rows) was randomly pulled using HiveQL from HDFS in a csv file format and then uploaded to for prediction modeling.

The workflow of machine learning process is pictured on **Fig. 6**.

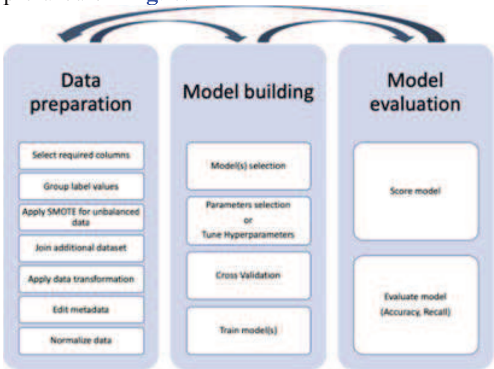

Fig. 6. Model flowchart

6.2 Data preparation

After uploading sampled dataset and then, computing and filtering out unnecessary columns, we have selected a label column – **level**, which shows jam level from 1 (almost no jams) to 5 (stand still jam). This field will be used for classification model building. However, the dataset is critically unbalanced, with the following percentages of a dataset: level 1 – 4.2%; level 2 – 30%; level 3 – 51%; level 4 – 15% and level 5 – 0,44%. In order to balance data, we grouped five categories into three groups: *1 (low)* – jam level of 1 and 2, *2 (medium)* – jam level of 3, *3 (heaviest)* – jam level of 4 and 5. In this case we assume that difference in initial *levels 1 and 2* (low levels of jam) is insignificant, as well as for *levels 4 and 5* (high levels of jams).

Although, grouping the categories help to balance data, this is not enough, as data is still biased to the medium level, which can affect model's prediction accuracy for other levels, particularly for highest level. To overcome such imbalance, we used Synthetic Minority Oversampling Technique (SMOTE), which helps statistically increase the number of under-sampled records in a dataset [14]. After application of SMOTE to the highest level of jams (level 3 out of 3), further transformations of dataset were applied:

- Additional dataset joined "national holidays", with the dates for national holidays in US for 2018.
- New fields were created "is_holiday" (1 – holiday, 0 - non-holiday), "is_weekend" (1 – weekend, 0 – not a weekend) "is_rush" (1 – if time between 7 - 9 am and 3 – 6 pm).
- Transforming cyclical fields, such as hours, minutes, seconds, weekday number and etc., to the appropriate representation. This can be done by converting features from Polar coordinate system to Cartesian, applying trigonometric functions:

$$x = r\ \sin\varphi\ \ and\ y = \ \cos\varphi,$$
$$where\ \varphi = \ k\frac{2\pi}{n},$$
$$k - value\ of\ the\ field,$$
$$n - number\ of\ possible\ values$$

Example: hour feature (0-23) transforms to: SIN(hour(2*PI()/24)) as sin_hour, COS(hour*(2*PI()/24)) as cos_hour.*

The reason for such transformation is the importance not to lose cyclical behavior of these fields, such as neighborhood of 0th to 23rd hour.

- Normalizing data using MinMax method, in order to rescale numerical data to one range.

6.3 Prediction model

In this study we aim to predict the appearance of three different levels of traffic jam (1-3) and clearly multi-class classification model is a good fit in this case. Azure ML has some of multiclass classification modes, and we chose Multiclass Decision Jungle.

Prior to model training we chose to split dataset into 70% of training set and 30% of model performance testing set. After several iterations of model training/testing and by calculating the weight of the columns, we excluded columns that have no value for traffic jams prediction. We used Cross Validation and Tune Model Hyperparameters, which helps determine the optimum parameters for selected model.

There are several metrics to validate performance of the multiclass classification model as follows: *Classification Accuracy* (overall and average) - percentage of total records classified correctly; *Precision/Sensitivity* (micro and macro) – ratio of correctly identified records as positive out of total records identified as positive; *Recall* (micro and macro) – ratio of correctly identified records as positive out of total actual positives [15]. Since our data has imbalance problem the best metric for model validation in our case is micro-averaged Recall method, where separate true positives and false negatives are summed up for different sets and then applied for recall calculation [16].

Our model for predicting the class of traffic jam has Micro-average Recall of 0.608515 and Confusion matrix shows that model is best at predicting 3rd class of label, the heaviest traffic jam level, with the accuracy of 70.7%. Although model fails to predict the lowest class of traffic jam (accuracy 38.4%), classifying 48.3% into false 2nd level (medium), a critical false classifying into the highest level of jam is low, only 13.3%. This means that model is good at predicting the heaviest traffic jams and probability of misclassifying any actual heavy congestions into low level is insignificant as well as probability of misclassifying any no-jam (low level) records into heavy congestion. (Tables 4-5).

Table 4. Confusion Matrix

		Predicted Class		
Actual class		1	2	3
	1	**38.4%**	48.3%	13.3%
	2	10.1%	**67.3%**	22.6%
	3	3.9%	25.4%	**70.7%**

Table 5. Model Metrics

	Macro - averaged (overall)	**Micro-averaged**
Accuracy	0.608515	0.73901
Precision	0.623402	0.608515
Recall	0.587839	**0.608515**

7. Conclusion

We can conclude that our traffic modeling revealed several insights into traffic situation in Los Angeles. There is denser traffic on highways 101, 405, 10. Although morning rush hours from 7 am to 9 am produce a lot of traffic, the heaviest traffic time start from 3pm and gets better after 6pm. Major areas of traffic are: Downtown Los Angeles, Santa Monica, Hollywood, and highways. There are also traffic congestions observed near LA airport from 5 am - 7 am and after business hours at 7 pm - 10 pm.

Besides, we present Big Data platform and architecture that allows storing and analyze giga-bytes of data set – possibly more data set as it is linearly scalable. Using such data and platform can also give an opportunity to predict traffic congestions. Prediction can be performed using machine learning algorithm – Multiclass Decision Jungle with the accuracy of 70% for predicting the heaviest traffic jam.

From the available data in Hadoop, which is limited to few days, we were able to provide an interactive tool for analysis, data manipulation and data prediction. Further work can be done with bigger dataset and more classification models in order to find more insights and create a data driven conclusions on LA County traffic situation by using this framework.

References

[1] M. Heiskala, J. Jokinen, and M. Tinnilä, "Crowdsensing-based transportation

services — An analysis from business model and sustainability viewpoints," *Research in Transportation Business & Management,* Vol 18, pp. 38-48, 2016.

[2] J. Barbaresso, G. Cordahi, D. Garcia et al., "USDOT's Intelligent Transportation Systems (ITS) ITS Strategic Plan 2015-2019," 2014.

[3] "Integrated Corridor Management," *Intelligent Transportation Systems - Integrated Corridor Management*, www.its.dot.gov/research_archives/icms/. Accessed April 14, 2019.

[4] J. Kestelyn, "Real-Time Data Visualization and Machine Learning for London Traffic Analysis," *Google Cloud*, 2016, cloud.google.com/blog/products/gcp/real-time-data-visualization-and-machine-learning-for-london-traffic-analysis. Accessed April 14, 2019.

[5] "Connected Citizens by Waze," *Waze*, www.waze.com/ccp. Accessed April 14, 2019.

[6] M. Schnuerle, "Louisville and Waze: Applying Mobility Data in Cities," *Harvard Civic Analytics Network Summit on Data-Smart Government*, 2017.

[7] Louisville Metro. "Thunder Jams, 2017 Traffic Delays." *CARTO*, louisvillemetro-ms.carto.com/builder/d98732d0-1f6a-4db2-9f8a-e58026bf0d39/embed. Accessed April 14, 2019.

[8] Louisville Metro. "Pothole Animation." *CARTO*, cdolabs-admin.carto.com/builder/a80f62bf-98e1-4591-8354-acfa8e51a8de/embed. Accessed April 14, 2019.

[9] E. Necula, "Analyzing Traffic Patterns on Street Segments Based on GPS Data Using R," *Transportation Research Procedia*, Vol. 10, pp. 276–285, 2015.

[10] J. Woo and Y. Xu, "Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing," *in Proc. of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), Las Vegas.* 2011.

[11] "Pandas.io.json.json_normalize." *Pandas.io.json.json_normalize - Pandas 0.24.2 Documentation*, pandas.pydata.org/pandas-docs/stable/reference/api/pandas.io.json.json_normalize.html. Accessed April 14, 2019.

[12] United States, Chief Executive Office County of Los Angeles. "Cities within the County of Los Angeles." *lacounty.gov.* Accessed April 14, 2019.

[13] Garyericson. "What Is - Azure Machine Learning Studio." *Microsoft Docs*, docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio. Accessed April 14, 2019.

[14] N.V. Chawla, et al. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.

[15] A. Tharwat, "Classification Assessment Methods." *Applied Computing and Informatics*, 2018.

[16] M. Sokolova and L. Guy, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, Vol. 45. No. 4, pp. 427–437, 2009.