Is Uber really cheaper than taxi?

Yahui Ke
kyahui3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Xueyang Zhang
xzhang870@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Sikai Zhao
sikaizhao@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Hao Wang
hwang794@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Yaoxu Xiao
yxiao356@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Building and visualizing prediction model to analyze the price of Uber and Taxi is the main objective of this project. Detailed plan and information is provided and Helmeier questions are answered in this proposal.

KEYWORDS

Uber, big data analysis, prediction model, visualization

1 INTRODUCTION

With the rapid development of Mobile Internet and booming of sharing economics, Uber has become one of the top choice for customers. However, the question of whether Uber is really cheaper than taxi or not needs more analysis to answer. This project is focus on building suitable model to predict the taxi price based on history data.

2 METHODOLOGY

The historical data of New York taxi trip can be obtained through NYC-TCL records. After cleaning and processing using Spark, the locations and time of trip start, end and the distance will be taken into consideration. Time and longitude and latitude of locations can be clustered by some algorithms such as k-means to refine the data. Then a machine learning model will be built and test and predictive results will be compared with the Uber real-time data. Finally, some kinds of visualization will be performed.

3 HEIMEIER QUESTIONS

What are we trying to do?

Predict taxi fee based on history data

How is it done today?

There are some customer surveys, no big data analysis report.

What's new in your approach?

Huge mount of data, only based on New York with big data visualization and analytic techniques

Who cares?

Taxi companies, Uber, Lyft and the customers

If you're successful, what difference and impact will it make, and how do you measure them?

Change customers mindset, maybe uber not always cheaper. Analyze the Usage data of taxi, uber and lyft after to meausre.

What are the risks and payoffs?

Group GPS coordination into same drop off and drop in location. Data cleaning, processing. Fare comparasion

How much will it cost?

No money cost, just 5 persons

How long will it take?

The project takes 8 weeks for 5 persons, 10 hour/week

How will progress be measured?

Present - Nov.5th Data collecting, cleaning, refining and locations clustering;
Nov. 5th - Dec. 3rd Machine learning Model building, testing and data visualization

4 DISTRIBUTION OF TEAM MEMBER EFFORT

All team members contribute similar amount of effort. The whole project was divided into five phase evenly and each member will be in charge of one with others working together.

Xueyang Zhang: Data cleaning and processing
Hao Wang: Location clustering
Yahui Ke: Draft prediction model building
Sikai Zhao: Model and algorithm testing and evaluation
Yaoxu Xiao: Visualization

5 LITERATURE SURVEY

In the past decades, the tremendously increase of data volumes has been calling for the development of large-scale data processing and analyzing models. Apache Spark was created by AMPLab at UC Berkeley in 2009. Back to that time, the available cluster programming models were specialized for specific data or purposes. Apache Spark came out as one unified engine for diverse big data processing [1]. Nowadays, Apache Spark has become one of the most popular big data processing systems in industry and research. It can support multi-language APIs such as Java, Python and R. Besides, it shows a huge performance improvement in memory management and networking layer[2]. Spark uses Resilient Distributed Datasets(RDDs) to store data, which can be stored in memory between queries. This enables a large increase in data processing speed compared with MapReduce and DAG engines [3]. As thus, in our project, we choose to use Apache Spark to process the data.

In the clustering section, we need to translate the latitude and longitude datas to locations and there are some alternative algorithms. Wagstaff put forward a k-mean clustering[4] which combined background information. Ben-Hur present a clustering method which used support vectors[5]. Ester put forward a clustering algorithm called DBSCAN[6] from which we can get arbitrary shape clusters.

Some machine learning models will be used in the project to predict the price of taxi based on big data. A deep neural network was built by Simon[7] using six different features as input to predict the taxi fare using six different features such as time, distance. Rangapuram presents a forecasting method that parametrizes a particular linear State space models(SSM)[8] using a recurrent neural network (RNN) for time series problem. Ahmed provides a systematic review of the current time series forecasting model using neural networks[9]. Ahmed tries to define the most important theoretical contributions in the development of artificial neural network models for the forecasting of non-linear time series.

How to test and evaluate the clustering and machine learning models? Vafeiadis performed and evaluated five machine learning algorithms based on cross-validation and used Monte Carlo simulation[10] to determined most efficient parameter combinations; Vora tested prediction models based on large scale dataset and scalability of algorithms and accurate prediction[11] are marked as major challenges; Traditional and modern clustering algorithms are evaluated by Xu based on parameters of distance or similarity measurement and evaluation indicators[12]; “Isolability” and “Unifiability” and their normalized value[13] are defined and

proposed to be critical quality metric for graph clustering evaluation by Biswas.

As for visualization, Integrative Genomics Viewer (IGV) could efficiently interact data with different views, which is more common in biomedical studiesand our project will challenge the performance of the existing visualization tools[11]. [12] presented a zero-drift localization system, which could help to localize and reuse the map in the already mapped areas with more accuracy. This system will achieve in our NYC map with high precision in specific pick up and drop off location. An approach to match an object or a scene with obstacles and outliers from the data sets images. The approach of polygonal mesh reconstruction errors to obtain higher accuracy of this approach could apply in our visualization process. This approach is more focuses on visualization recognition than data visualization[13].

REFERENCES

[1] Vafeiadis, Thanasis, et al. "A comparison of machine learning techniques for customer churn prediction."Simulation Modelling Practice and Theory 55 (2015):1-9.

[2] Armbrust, Michael, et al. "Scaling spark in the real world: performance and usability." Proceedings of the VLDB Endowment 8.12 (2015): 1840-1843.

[3] Gopalani, Satish, and Rohan Arora. "Comparing apache spark and map reduce with performance analysis using k-means." International journal of computer applications 113.1 (2015).

[4] Ben-Hur, Asa, et al. "Support vector clustering." Journal of machine learning research 2.Dec (2001): 125-137.

[5] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

[6] Vafeiadis, Thanasis, et al. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[7] Upadhyay, Rishabh Lui, Simon. (2017). Taxi Fare Rate Classification Using Deep Networks.

[8] Rangapuram, Syama Sundar and Seeger, Matthias W and Gasthaus, Jan and Stella, Lorenzo and Wang, Yuyang and Januschowski, Tim. Deep State Space Models for Time Series Forecasting, NIPS2018$_8$004.

[9] Ahmed Tealab,Time series forecasting using artificial neural networks methodologies: A systematic review, Future Computing and Informatics Journal, Volume 3, Issue 2, 2018, Pages 334-340,ISSN 2314-7288.

[10] Vafeiadis, Thanasis, et al. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[11] Vora, Deepali, and Kamatchi Iyer. "Evaluating the Effectiveness of Machine Learning Algorithms in Predictive Modelling." International Journal of Engineering Technology 7.3.4 (2018): 197-199.

[12] Xu, Dongkuan, and Yingjie Tian. "A comprehensive survey of clustering algorithms." Annals of Data Science 2.2 (2015): 165-193.

[13] Biswas, Anupam, and Bhaskar Biswas. "Defining quality metrics for graph clustering evaluation." Expert Systems with Applications 71 (2017): 1-17.

[14] Helga Thorvaldsdóttir, et al. "Integrative Genomics Viewer (IGV):high-performance genomics data." Briefings in Bioinformatics, Volume 14, NO 2.(2013):178-192.

[15] Raúl Mur-Artal, Juan D.Tardós. "Visual-Inertial Monocular SLAM With Map Reuse." IEEE Robotics and Automation Letters, Volume 2, Issue 2 (2017):796-803.

[16] Yasutaka Furukawa, Jean Ponce. "Accurate, Dense, and Robust Multiview Stereopsis." IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 32, Issue 8 (2010)