Transportation Research Part C 92 (2018) 298–322

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Real time location prediction with taxi-GPS data streams

Arnab Kumar Laha*, Sayan Putatunda

Indian Institute of Management Ahmedabad, India

A R T I C L E I N F O

Keywords:
Spherical data analysis
Incremental learning
Intelligent transportation systems
Multivariate multiple regression
Sliding windows
Streaming data

A B S T R A C T

The prediction of the destination location at the time of pickup is an important problem with potential for substantial impact on the efficiency of a GPS-enabled taxi service. While this problem has been explored earlier in the batch data set-up, we propose in this paper new solutions in the streaming data set-up. We examine four incremental learning methods using a damped window model namely, Multivariate multiple regression, Spherical-spherical regression, Randomized spherical K-NN regression and an Ensemble of these methods for their effectiveness in solving the destination prediction problem. The performance of these methods on several large datasets are evaluated using suitably chosen metrics and they were also compared with some other existing methods. We found that the Multivariate multiple regression method has the best performance in terms of prediction accuracy but the Spherical-spherical regression method is the best performer when we take into account the accuracy time trade-off criterion. The next pickup location problem, where we are interested in predicting the next pickup location for a taxi given the dropoff location coordinates of the previous trip as input is also considered and the aforementioned methods are examined for their suitability using real world datasets. As in the case of destination prediction problem, here also we find that the Multivariate multiple regression method gives better performance than the rest when we consider prediction accuracy but the Spherical-spherical regression method is the best performer when the accuracy-time trade-off criterion is taken into account.

1. Introduction

In recent times, across the world we have seen a spurt in the usage of GPS-based taxi (interchangeably also called cabs in this paper) services viz. Uber, Lyft, Ola, Didi, etc. For GPS-enabled taxis it is possible to continuously collect the geo-spatial location data for every trip. These recordings are often referred to as GPS traces. The GPS traces generated by a vehicle is a rich source of streaming data which can give insights on passenger demand and their mobility patterns.

Geo-spatial data streams have many applications in the passenger transportation industry. The intelligent transportation systems literature is replete with various applications viz. traffic monitoring (Herring et al., 2010), passenger finding (Veloso et al., 2011), vacant taxi finding (Phithakkitnukoon et al., 2010), hotspots identification (Chang et al., 2010), trajectory mapping (Liu et al., 2012), etc. where GPS traces have been instrumental in finding interesting insights (see Chen (2014) for more details).

Generally, some of the most important questions for a transport dispatch system as elucidated in Moreira-Matias et al. (2016b) are (a) What's the destination passengers are traveling to i.e. where the vehicle would be vacant? (b) Travel time estimation i.e. how long the vehicle would be occupied? and (c) What is the demand at a particular location in the time interval *t*? (Liu et al., 2009; Mendes-Moreira et al., 2012; Moreira-Matias et al., 2013a, 2014). The answers to the above questions give insights on the transportation system

* Corresponding author.
E-mail addresses: arnab@iima.ac.in (A.K. Laha), sayanp@iima.ac.in (S. Putatunda).

https://doi.org/10.1016/j.trc.2018.05.005
Received 12 December 2017; Received in revised form 4 May 2018; Accepted 7 May 2018
0968-090X/

properties and the passenger mobility. The knowledge of a trip destination before the passenger boards the vehicle can help the transport dispatch system in its operational planning. Also, the knowledge of the next pickup location of a vehicle can help the transport dispatch system in guiding the taxi drivers to find their next passengers in a systematic and efficient manner. These problems are well explored in the literature but mostly in the batch learning set-up (Gandhi, 2015; Veloso et al., 2011) (see Section 1.1 for more details). But analyzing the GPS trace data in streaming data context enables a transport dispatch system to take decisions in real time.

In this paper, we are interested in the real-time destination/next pickup location prediction problem. The analysis of GPS data streams of taxis or any other public transport for real time prediction opens up new research opportunities for improving the reliability of a transport dispatch system such as an introduction of real-time decision models to support "operational control" (Moreira-Matias et al., 2016a). Eberlein et al. (1999) gives more details about the various models to solve real-time transit operational control problems using real-time vehicle location information i.e. using the geospatial streaming data. It is expected that in near future we would easily be able to collect real-time mobility trace data from multiple sources such as taxis, buses, and individual smart-phones (Moreira-Matias et al., 2016b). The analysis of these data streams offer a great opportunity for development of new methodologies that have applications in the area of Intelligent Transportation Systems as mentioned above and this is our primary motivation for exploring newer methods for solving the destination/next pickup prediction problem.

In this paper we discuss four new methods which to the best of our knowledge have never been applied before for solving the destination/next-pickup location prediction problem in a streaming data context. One of these methods has its origin in the literature on directional data analysis, another one is an adaptation of a popular machine learning algorithm for analysis of streaming spherical data, the third one is an adaptation of a multivariate statistics method which has been previously implemented only with static (a.k.a. batch) data, and finally we build an ensemble using the above-mentioned three methods. We give details of these methods in Section 3. An extensive performance study of these methods is carried out with five real world datasets and recommendation for use is also given. Our contribution is not only in development of new approaches for the real-time destination/next-pickup location prediction problem but also in development of a framework for selection of suitable methods for different scenarios.

The paper is structured as follows. In Section 1.1, we give a brief review of literature. This is followed by a background of Streaming data and challenges associated with it in Section 1.2. Section 2 describes the Destination Prediction problem statement. Section 3 discusses the methodology and Section 4 describes the datasets used in this paper. Section 5 discusses the various evaluation metrics used in the paper and in Section 6 we give an overview of the concept drift phenomenon. Section 7 discusses the results of the various experiments conducted. Section 8 presents the Next Pickup Prediction problem. Finally, Section 9 concludes the paper.

1.1. Related work

Analysis of GPS trace data is a challenging problem with lots of applications in the transportation domain. Most of the currently available methods use batch learning methods (Gambs and Killijian, 2012; Krumm, 2008; Krumm and Horvitz, 2006; Li et al., 2012; Simmons et al., 2006). It's only recently, that some papers have taken into account the streaming nature of GPS trace data where online and incremental learning methods for solving some specific problems have been proposed (Lam et al., 2015; Moreira-Matias et al., 2013b, 2016b; Sun et al., 2012).

Predicting destination from partial trajectories is a problem that has tremendous potential of real world applications. Several solutions of this problem from different perspectives has been proposed in the literature using Origin-Destination (OD) matrix (Barceló et al., 2010; Jin et al., 2008; Moreira-Matias et al., 2016b; Park et al., 2014; Yue et al., 2009; Zhang et al., 2011), Bayesian inference (Hazelton, 2008; Parry and Hazelton, 2012), Support Vector Machine (SVM) (Li et al., 2011), Clustering (Chang et al., 2010; Li et al., 2012), Mobility markov chains (Gambs and Killijian, 2012), Optimization (Miao et al., 2015), ensemble learning (Lam et al., 2015) and many more. In this section, we have given a brief literature review of some of the works done in this field.

Krumm and Horvitz (2006) implemented a method termed *predestination* that predicts where the driver is going on the go. A Bayesian inference model is built that uses driving behavior data along with GPS data. Other papers that have used Bayesian inference for solving similar problems are Marmasse and Schmandt (2002) and Liao et al. (2004). Gambs and Killijian (2012) discusses the next place prediction problem using the information of coordinates of visited places applying the n-MMC (Mobility Markov Chain) algorithm. n-MMC is a modified version of the Mobility Markov Chain model where they keep track of n-previous locations visited.

Simmons et al. (2006) used a hidden markov model (HMM) for prediction of route and intended destination. Markov models have also been used in other studies for making short term route predictions (Krumm, 2008). Xue et al. (2013) proposed the sub-trajectory synthesis method for destination prediction. Chen et al. (2011) worked on extracting route pattern from user's personal trajectory data using a probabilistic model which they termed as "Continuous Route Pattern Mining (CRPM)".

In recent years, some work has been reported in the literature which has given emphasis to the streaming nature of the GPS traces data and has shifted the focus towards real time or near real time prediction. Moreira-Matias et al. (2013b) worked on predicting the passenger demand in a streaming data context and proposed an ensemble method with sliding window technique. Sun et al. (2012) worked on taxi trajectory anomaly detection in real time from GPS data streams. A *nearest neighbor technique (NNT)* was proposed by Tiesyte and Jensen (2008), where they used Euclidean distance as a distance measure for travel time prediction of vehicles. Moreira-Matias et al. (2016a) worked on eliminating bus bunching in real time using an online learning approach. Ermagun et al. (2017) used online location based search and discovery services for real time trip purpose prediction. Simoncini et al. (2018) used vehicle GPS data to classify the different types of vehicles on a road network. They used Long Short-Term Memory (LSTM) recurrent neural networks. For a more comprehensive review of literature please see chapter 2 of Putatunda (2017).

The destination prediction problem in streaming data context has not been explored much in the literature. Lam et al. (2015) worked on real-time prediction of destination and travel time estimation from given partial trajectories using an ensemble learning