On learning from taxi-GPS traces

João Mendes-Moreira[1] and Luís Moreira-Matias[2]

[1] LIAAD-INESC TEC, Faculty of Engineering, University of Porto
Tel.: +351-22-5082142
jmoreira@fe.up.pt
[2] Nec Lab, Heidelberg
luis.matias@neclab.eu

1 The challenge

Electronic taxi dispatch systems are in wide use today. These systems have replaced the traditional VHF-radio dispatch by installing mobile data terminals in taxis, which typically provide GPS localization information and taximeter state. In the last couple of years, the broadcast-based radio messages for service dispatching were replaced by unicast-based messages between the taxi central and the selected vehicle.

In most cases, taxi drivers operating through an electronic dispatch system do not indicate the final destination of the ride. In some cases, particularly when the demand for taxis is higher than the taxi availability, the closest taxi to a particular location is exactly the taxi that will end its current ride at that location. While in broadcast-based radio dispatching this was not a problem, in unicast-based electronic dispatching it becomes a problem, given that most drivers do not indicate the final destination of their current ride. To improve the efficiency of electronic taxi dispatching systems it becomes important to be able to predict the final destination of busy taxis. The spatial trajectory of a busy taxi could provide some hints on where it is going. Similarly, given the taxi id, it might be possible to guess its final destination based on the regularity of pre-hired services. In a significant number of taxi rides (approximately 25%), the taxi has been called through the taxi call-center, and thus the passenger's telephone id can be used to narrow the destination prediction based on the historical ride data of such telephone id.

In this challenge, the goal was to build a predictive framework able to infer the final destination of each taxi ride based on their (initial) partial trajectories. This challenge was divisible in two different outputs: (a) the destination coordinates (WGS84) and (b) the total trip's travel time (counting from the service's starting point, in seconds). The entries were evaluated using the Mean Haversine Distance and the Root Mean Squared Logarithmic Error, regarding the destination and the travel time prediction problems, respectively.

2 The dataset

As training set it was provided an accurate dataset describing a complete year (from 01/07/2013 to 30/06/2014) of the (busy) trajectories performed by all the

442 taxis running in the city of Porto (Fig.1), in Portugal (i.e. 3Gb of data stored in one single CSV file). These taxis operate through a taxi dispatch central, using mobile data terminals installed in the vehicles. We categorize each ride into three categories: A) taxi central based, B) stand-based or C) non-taxi central based. For the first, it was provide an anonymized id, when such information is available from the telephone call. The last two categories refer to services that were demanded directly to the taxi drivers on a B) taxi stand or on a C) random street.

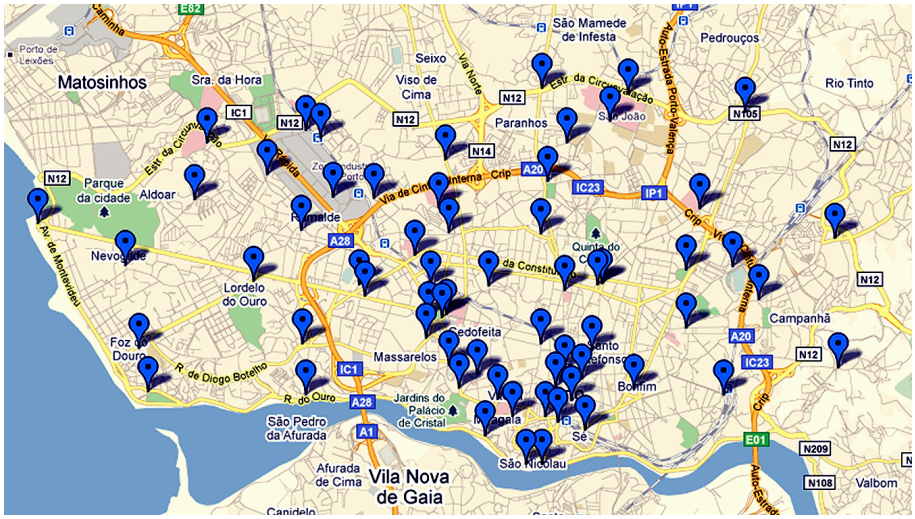

Fig. 1. Porto city map with taxi-stands signalized

Each data sample corresponds to one completed trip. It contains a total of nine features: trip_id, call_type, otigin_call, origin_stand, taxi_id, timestamp, day_type, missing_data, polyline.

Five datasets were provided for validation.

A total of two sample submission files were provided. One regarding the destination prediction problem while the second one focusing on the travel time prediction problem. Both always using the same output value for all samples. For the first problem, it used the location of Porto's main Avenue, in downtown (i.e. Avenida dos Aliados). For the second one, the averaged travel time of all trips in the training set was used.

3 The contest and the papers

The contest was done through Kaggle [3]. More than 700 teams have participated. The three papers presented are the winners of each of the two subproblems (a)

[3] https://www.kaggle.com/

and (b) a third paper whose team was 7th and 3rd in subproblems (a) and (b) respectively. Alexandre de Brébisson, Étienne Simon, Alex Auvolat, Pascal Vincent and Yoshua Bengio, all working partial/full time at MILA lab, University of Montréal, Canada won subproblem (a) using multi-layer perceptrons, bidirectional recurrent neural networks and models inspired from memory networks. Thomas Hoch from Software Competence Center Hagenberg GmbH, Austria won subproblem (b) using ensemble learning combined with a spatial clustering approach. Hoang Thanh Lam, Ernesto Diaz-Aviles, Alessandra Pascale, Yiannis Gkoufas, and Bei Chen from IBM research, Ireland present a solution based on trip matching and ensemble learning.

Acknowledgements

We thank all members of GeoLink, for the data and the given support to mount this challenge. We also thank the program committee of this challenge as well as the organizers of the ECML-PKDD 2015 conference.