Detecting Taxi Speeding from Sparse and Low-Sampled Trajectory Data

Xibo Zhou[1,2,3,4(✉)], Qiong Luo[1], Dian Zhang[4], and Lionel M. Ni[5]

[1] Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{xzhouaa,luo}@cse.ust.hk
[2] Guangzhou HKUST Fok Ying Tung Research Institute,
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[3] Guangdong Key Laboratory of Popular High Performance Computers,
Shenzhen, China
[4] Shenzhen Key Laboratory of Service Computing and Applications, Shenzhen, China
zhangd@szu.edu.cn
[5] University of Macau, Macau, China
ni@umac.mo

Abstract. Taxis are a major means of public transportation in large cities, and speeding is a common problem among motor vehicles, including taxis. Unless caught by sensors or patrol officers, many speeding incidents go unnoticed, which pose potential threat to road safety. In this paper, we propose to detect speeding behaviors of individual taxis from taxi trajectory data. Such detection results are useful for driver risk analysis and road safety management. However, the taxi trajectory data are geographically sparse and the sample rate is low. Furthermore, existing methods mainly deal with the estimation of collective road speeds whereas we focus on the speeds of individual vehicles. As such, we propose to use a two-fold collective matrix factorization (CMF)-based model to estimate the individual vehicle speed. We have evaluated our method on real-world datasets, and the results show the effectiveness of our method in detecting taxi speeding behaviors.

Keywords: Speeding · Collective matrix factorization · Trajectory

1 Introduction

In many large cities, with the popularity of private cars and taxis traveling around, the incidence of traffic accidents has been rapidly increasing, which often causes damage to personal properties and public facilities, and even leads to traffic congestions. One of the most common inducements of these accidents is speeding. Taxi speeding is the most common violation among taxi drivers [4], which reduces the quality of road safety in modern cities.

In order to solve the major problem of speeding, authorities nowadays have paid a lot of attention and effort by distributing sensors (such as loops and

Y. Cai et al. (Eds.): APWeb-WAIM 2018, LNCS 10988, pp. 214–222, 2018.
https://doi.org/10.1007/978-3-319-96893-3_16

cameras) along roadways to monitor the real-time driving speeds. Due to the nontrivial cost, most of these sensors are limited in covering freeways and arterial roads. Unfortunately, collector roads, referring to secondary main roads that connect arterial roads in cities, are often sparsely covered by these sensors. As a result, the driving speed information collected in this way is not complete. On the other hand, the ubiquitous taxi trajectory data provides alternative opportunities to estimate the driving speeds and detect speeding. Most of these trajectory data contains the instant speed information recorded by speed meters embedded on taxis. Unfortunately, due to the low sampling rate, these information are usually too sparse to cover the entire travel paths. Several approaches [2,3,5–8] have been trying to predict traffic conditions in terms of road speeds by utilizing taxi trajectory data. However, these approaches are not able to monitor the driving speeds of individual vehicles, thus cannot be applied to detect taxi speeding behaviors.

In this paper, we propose a prediction system to detect individual taxi speeding behaviors by utilizing taxi trajectory data. Different from previous works, we try to estimate the driving speed for each individual taxi along the road it traveled. We first propose a two-fold collective matrix factorization (CMF)-based model to predict the individual driving speed, capturing the spatial and temporal patterns of traffic conditions, and predict individual speeding based on the estimation result. We evaluate our system on real-world taxi trajectory data, and the results show that our system is effective to detect speeding. Moreover, we conduct an empirical study on the occurrence of taxi speeding.

2 Overview

2.1 Preliminary

Definition 1 (Road Segment). *A road segment e is a directed polyline between two road intersections v_i and v_j, and there is no other road intersection on e. We denote $v_i \in e$ and $v_j \in e$.*

Definition 2 (Road Network). *A road network is a weighted directed graph $G = (V, E)$, where V is a set of road intersections (or vertices), and E is a set of road segments (or edges). The weight of a road segment is represented by its properties.*

Definition 3 (Tracing Record). *A tracing record r of a taxi is denoted as a tuple $r = \langle id, t, p, v \rangle$, where $r(id)$ is the taxi id, $r(t)$ is the record time, $r(p)$ is the location point of the taxi at $r(t)$ represented by its latitude and longitude, and $r(v)$ is the instant speed of the taxi at $r(t)$.*

Definition 4 (Trajectory). *A trajectory T of a taxi with id tid is a sequence of tracing records denoted as $T = (r_1, r_2, \cdots, r_n)$, where $r_i(id) = tid$ for $i = 1, \cdots, n$. We denote $r_i \in T$ for $i = 1, \cdots, n$ and $|T| = n$.*

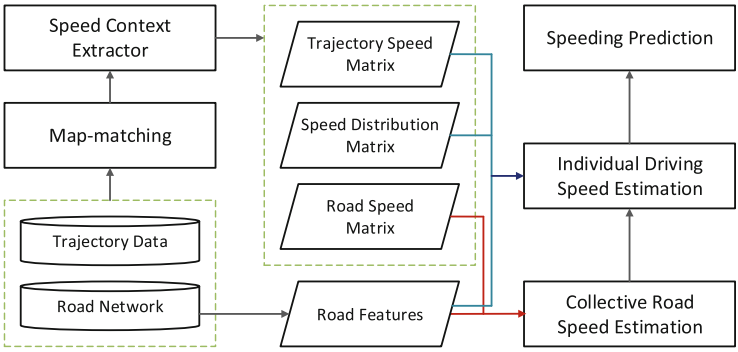


Fig. 1. The workflow of taxi speeding prediction.

Definition 5 (Path). *A path* $P = (e_1, e_2, \cdots, e_n)$ *is a sequence of road segments where* e_i *and* e_{i+1} *are connected for* $i = 1, 2, \cdots, n-1$. *Two road segments* e_i *and* e_j *are connected if there exists some intersection* v *such that* $v \in e_i$ *and* $v \in e_j$.

Definition 6 (Trajectory Speed). *Given a trajectory* T *and its map-matched path* $P = (e_1, e_2, \cdots, e_{n_e})$, *the trajectory speed of* T *is denoted as* $v(T) = (v_1(T), v_2(T), \cdots, v_{n_e}(T))$, *where* $v_i(T)$ *is the driving speed of* T *on road segment* $e_i \in P$.

Definition 7 (Speeding). *Given a trajectory* T *of a taxi and its map-matched path* $P = (e_1, e_2, \cdots, e_{n_e})$, *we say the taxi is speeding on road segment* e_i *if* $v_i(T) > v_{max}$, *where* v_{max} *is the speed limit of road segment* e_i.

In this paper, we will solve the problem of finding all the speeding behaviors from a trajectory dataset.

2.2 Framework

Figure 1 shows the process of our taxi speeding prediction system. We first map raw trajectories into connected paths constrained to the road network, and then extract the instant speed information in terms of three matrices, namely collective road speed matrix, driving speed distribution matrix, and individual trajectory speed matrix. Next, we implement a two-fold CMF-based model to estimate the driving speed of individual trajectories. The first fold of our CMF-based model is built to predict the missing values of the collective road speed matrix by utilizing the road features extracted from the road network data, and the second fold is built to predict the missing values of the individual trajectory speed matrix by utilizing the other two matrices as well as the road features. Finally, we use the completed matrices to predict speeding based on the threshold of road speed limits.

3 Methodology

3.1 Matrix Construction

In order to extract spatial features from raw trajectories, we use the Hidden Markov Model (HMM)-based map-matching algorithm [10] to convert raw location points into travel paths along the road network. In order to distinguish the traffic conditions the road network and different time slot within a day, we extract the collective road speeds from map-matched trajectories by constructing 2D matrices with the two dimensions standing for road segments and time slots. Given a trajectory dataset and the road network $G = (V, E)$, suppose we split the time in one day into a number of slots μ_t, the collective road speed v_{ij} is the average instant speed of taxis recorded on road segment $e_i \in E$ during the jth time slot. We assume that the instant speeds of all the taxis passing a certain road segment e_i during the jth time slot follows a Gaussian distribution $\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$ [7]. Thus, the speed distribution matrix can be constructed similar to the collective road speed matrix. Note that v_{ij} and $\mathcal{N}(\mu_{ij})$ is missing if no taxi is traveling on e_i during the jth time slot. Thus, the collective road speed matrix $V_c \in \mathbb{R}^{|E| \times \mu_t}$ with a large percentage of values missing due to data sparsity. In our dataset, if the time slot is set with an interval of 30 min, only 1.8% entries of V_c have values. With such a low sparsity, it is difficult to predict the missing values only using its own non-zero values.

In order to solve this problem, we build another dense matrix by extracting road features from the road network data, and use it for supplementing the speed estimation. Besides the spatial and topological information, the road network data consists of road contexts including road level, road direction, road width, road type (indicating the road is one-way or bi-directional), road length, and curvature. Each categorical feature is flattened into a vector with 0 and 1, and each numerical feature is normalized into $(0, 1)$.

3.2 Collective Road Speed Estimation

We implement a two-fold CMF-based model to estimate the driving speed of individual trajectories. The first fold of our CMF-based model is constructed to predict the missing values of the collective road speed matrix. Given the collective road speed matrix V_c, suppose the dimensionality of latent feature vectors for matrix factorization is k, V_c can be factorized into two latent feature matrices $V_c \approx W \times X = \hat{V_c}$. The dimension of W and X are $|E| \times k$ and $k \times \mu_t$. The loss function between V_c and $\hat{V_c}$ is denoted as:

$$\mathscr{J}_v = \sum_{v_{ij} \neq null} (v_{ij} - \sum_{s=1}^{k} w_{is} x_{sj})^2 \tag{1}$$

where $v_{ij} \neq null$ means that v_{ij} is not missing, w_{is} and x_{sj} represent the corresponding element in W and X, respectively.

Similarly, the road feature matrix F can be factorized into two latent feature matrices $F \approx W \times Y = \hat{F}$. Note that F and V_c shares the latent feature matrix W, and the dimension of Y is $k \times h$, where h is the number of road features. Since F is a dense matrix, roads with similar features will generate similar latent feature vectors, which will be propagated into the factorization of V_c, and thus reduce the sparsity problem of factorizing V_c.

The loss function between F and $\hat{F}$ is denoted as:

$$\mathscr{J}_f = \sum_{f_{ij} \in F} (f_{ij} - \sum_{s=1}^{k} w_{is} y_{sj})^2 \tag{2}$$

where w_{is} and y_{sj} represent the corresponding element in W and Y, respectively. The inference process is to minimize the loss functions for both of the two matrices. Thus, the objective function is denoted as:

$$O(V_c, F) = \alpha_v \times \mathscr{J}_v + \alpha_f \times \mathscr{J}_f + R(V_c, F) \tag{3}$$

where α_v is the weight of relative importance of V_c, α_f is the weight of relative importance of F, and $R(V_c, F)$ is the L2 regularization term, denoted as $R(V_c, F) = \lambda \times \sum_{i=1}^{k} \theta_i^2$, where λ is the weight of the regularization term, and θ_i is the ith value of latent factors.

We implement the Newton-Raphson method [9] to find W, X, Y that minimize the objective function $O(V_c, F)$. After that, the missing values of V_c can be predicted by calculating $W \times X$.

3.3 Individual Driving Speed Estimation

The second fold of our model is constructed for predicting the driving speed of individual taxis. Similar with the collective road speed estimation, for an individual taxi with id *tid*, we extract the instant speeds from its map-matched trajectories by constructing 2D matrices with the two dimensions standing for road segments and time slots. Since the sparsity of the speed matrix extracted from a single taxi trajectory is much lower than that of collective road speed matrix, instead of predicting the driving speed, we use standard z-score to estimate the deviation of driving speed $v(tid)$ on e_i during the jth time slot from the distribution, denoted as:

$$z(tid, i, j) = \frac{v(tid) - \mu_{ij}}{\sigma_{ij}} \tag{4}$$

Moreover, since a single taxi tends to travel around limited area within a city, we reduce the dimensionality of its speed matrix by pruning the road segments that has never been traveled along. Based on our observation, for a taxi driver, the deviation of his driving speed from the distribution tends to be more stable than the driving speed itself. Hence, even if the testing trajectory reaches the pruned road segments, the error of predicting its speed deviation is acceptable. Thus, given the trajectories of each taxi, we can construct a $n_v \times \mu_t$

matrix Z of its travel speed deviations, where n_v is the number of road segments traversed by the taxi. Suppose the dimensionality of latent feature vectors for matrix factorization is k', Z can be factorized into two latent feature matrices $Z \approx W' \times X' = \hat{Z}'$. Similar with the collective road speed estimation, we implement the Newton-Raphson method to find W', X' that minimize the objective function. After predicting the missing values of the speed deviation matrix, we can estimate the driving speed of taxi *tid* passing on road segment e_i during the jth time slot as follows:

$$v(tid, i, j) = z(tid, i, j) \times \sigma_{ij} + \mu_{ij} \tag{5}$$

Finally, given the speed limit v_{max} of each road segment e_i, it is straight forward to predict the speeding behavior of a taxi, by utilizing its driving speed matrix.

4 Evaluation

4.1 Experiment Setup

The experiments are conducted on a Linux server with a CPU of Intel Core i5-4590 and 8 GB memory. The operating system is Ubuntu 14.04, and the code is written in Python 2.7.6. We use a dataset collected from a large city in China. The dataset contains 90 million taxi tracking records of 12,000 taxis for 30 days. The road network consists of 74,184 intersections and 54,723 road segments. There are five levels of road segments in our road network data, and the speed limits of each level are 30 km/h, 50 km/h, 60 km/h, 80 km/h, 90 km/h, respectively.

We use three metrics to evaluate the performance of our proposed model, namely the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and the normalized Root Mean Square Error (NRMSE) [1]. A smaller MAE, RMSE and NRMSE indicates that the predicted values are closer to the ground truth, which means a better performance.

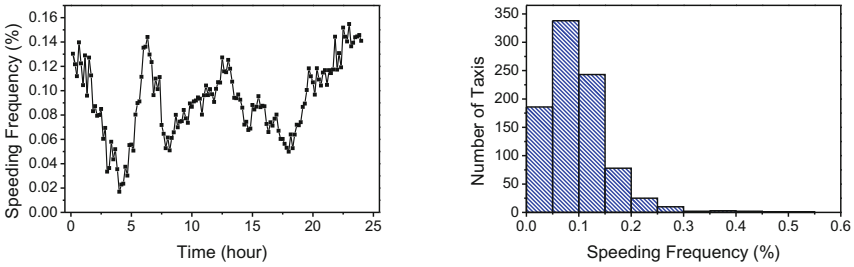
In order to evaluate the performance of our individual speed estimation model, we compare our framework with the following baselines: (1) the average value of the instant speeds of the two adjacent trajectory records (AAS); (2) the average speed of the sub-path between each two consecutive trajectory records (APS); (3) the collective road speed of the nearest road segment that is available during the corresponding time slot (NRS); (4) the collective road speed of the corresponding road segment and time slot predicted by our first-fold CMF-based model (CRS); and (5) the driving speed predicted by a straight forward implementation of our first-fold CMF-based model on the individual taxi speed matrix (SCMF).

4.2 Experiment Results

In our experiments, we build the collective road speed matrix and the individual driving speed matrix using the instant speed information contained in the

Table 1. Performance of speed estimation model

Method	RMSE(km/h)	NRMSE(%)	MAE(km/h)
AAS	26.279	55.465	15.793
APS	18.474	34.940	10.951
NRS	12.677	18.718	7.248
CRS	8.469	10.080	5.185
SCMF	23.584	48.720	12.604
Two-fold CMF	5.988	6.301	4.068



(a) Time (b) Taxi

Fig. 2. Frequency of speeding occurrence among different time slots and taxis.

trajectory dataset. After building the matrices, we use 10-fold cross validation to evaluate the performance of our model. The default values of each parameter described in Sect. 3 are: $k = 5$, $k' = 15$, $\alpha_v = 0.5$, $\alpha_f = 0.8$, $\alpha_z = 0.8$, $\alpha_{f'} = 0.3$, $\lambda = 0.5$, $\lambda^{'} = 0.2$.

We evaluate the overall performance of our model compared with the baseline listed above. The results are shown in Table 1. The first three baselines (AAS, APS and NRS) are calculated from the geographic information or statistics extracted from the dataset without matrix completion, and the performances of these methods are poor. If we directly use the collective road speed calculated by the first-fold CMF-based model to predict the individual trajectory speed, the performance is better but still not satisfactory. Meanwhile, since the individual speed matrix is too sparse, the performance is even worse if we directly implement the CMF-based model on it. Finally, it is observed that our proposed two-fold CMF-based model achieves the best performance compared to all the baselines, whose prediction error (in terms of normalized root mean square error) is only 6.301%, which is satisfactory for individual trajectory speed estimation.

4.3 Empirical Study

We evaluate the occurrence of taxi speeding detected by our proposed model among different time periods and taxis. Figure 2a demonstrates the frequency of

taxi speeding occurrence on different time periods within a day. As we can see, the frequency of taxi speeding occurrence is low before dawn (around 4–6 a.m.) because most of the taxi drivers are off work, and also limited during rush hours (around 8–10 a.m. and 5–8 p.m.) because the traffic is usually congested. On the other hand, the frequency of taxi speeding occurrence is relatively high in the morning before rush hours (around 7 a.m.), after lunchtime (around 1–2 p.m.), and at midnight (from 23 p.m. to 2 a.m.). There are various possible reasons for this phenomenon, such as less congested traffic conditions, requirements of quick deliveries to the work places, or less effective monitoring (at midnight). Last but not least, Fig. 2b shows the frequency of taxi speeding occurrence among different taxi drivers. According to the statistics, most of the taxi drivers do not often cross the speed limits (with the frequency below 15%). On the contrary, we observe that a small amount of taxi drivers (around 5%) have high frequency of speeding (over 20%). Therefore, our speeding prediction system is helpful in finding out these drivers with bad habits of speeding, providing guidance to authorities.

5 Conclusion

In this paper, we propose a prediction system to detect individual taxi speeding behaviors by utilizing sparse and low-sampled trajectory data. Most of the existing approaches are designed to predict the collective speed of roads or paths, considering spatial and temporal dynamics and patterns. However, they cannot estimate the driving speed of an individual vehicle from the trajectory datasets. We implement a two-fold (CMF)-based model to predict the individual driving speed, and use the completed speed matrix to predict taxi speeding. We conduct intensive experiments on real trajectory data. The results show that our proposed system achieves a satisfactory performance.

Acknowledgments. This work is supported in part by the Guangdong Pre-national project 2014GKXM054 and the Guangdong Province Key Laboratory of Popular High Performance Computers 2017B030314073.

References

1. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)?-arguments against avoiding RMSE in the literature. Geosci. Model Dev. **7**(3), 1247–1250 (2014)
2. Jenelius, E., Koutsopoulos, H.N.: Travel time estimation for urban road networks using low frequency probe vehicle data. Transp. Res. Part B: Methodol. **53**, 64–81 (2013)
3. Liu, Y., Li, Z.: A novel algorithm of low sampling rate GPS trajectories on map-matching. EURASIP J. Wirel. Commun. Netw. **2017**(1), 30 (2017)
4. Tseng, C.-M.: Operating styles, working time and daily driving distance in relation to a taxi driver's speeding offenses in Taiwan. Accid. Anal. Prev. **52**, 1–8 (2013)

5. Wang, Y., Zheng, Y., Xue, Y.: Travel time estimation of a path using sparse trajectories. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 25–34. ACM (2014)
6. Wang, Z., Li, M., Wang, L., Liu, X.: Estimation trajectory of the low-frequency floating car considering the traffic control. Math. Prob. Eng. **2013**, 11 (2013)
7. Xin, X., Lu, C., Wang, Y., Huang, H.: Forecasting collector road speeds under high percentage of missing data. In: AAAI, pp. 1917–1923 (2015)
8. Xu, J., Deng, D., Demiryurek, U., Shahabi, C., van der Schaar, M.: Mining the situation: spatiotemporal traffic prediction with big data. IEEE J. Sel. Top. Sig. Process. **9**(4), 702–715 (2015)
9. Ypma, T.J.: Historical development of the Newton-Raphson method. SIAM Rev. **37**(4), 531–551 (1995)
10. Zhou, X., Ding, Y., Tan, H., Luo, Q., Ni, L.M.: HIMM: an HMM-based interactive map-matching system. In: Candan, S., Chen, L., Pedersen, T.B., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS, vol. 10178, pp. 3–18. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55699-4_1