Inferring Road Type in Crowdsourced Map Services

Ye Ding[1], Jiangchuan Zheng[1], Haoyu Tan[1], Wuman Luo[1], and Lionel M. Ni[1,2]

[1] Department of Computer Science and Engineering
[2] Guangzhou HKUST Fok Ying Tung Research Institute
[3] The Hong Kong University of Science and Technology
{valency,jczheng,hytan,luowuman,ni}@cse.ust.hk

Abstract. In crowdsourced map services, digital maps are created and updated manually by volunteered users. Existing service providers usually provide users with a feature-rich map editor to add, drop, and modify roads. To make the map data more useful for widely-used applications such as navigation systems and travel planning services, it is important to provide not only the topology of the road network and the shapes of the roads, but also the types of each road segment (e.g., highway, regular road, secondary way, etc.). To reduce the cost of manual map editing, it is desirable to generate proper recommendations for users to choose from or conduct further modifications. There are several recent works aimed at generating road shapes from large number of historical trajectories; while to the best of our knowledge, none of the existing works have addressed the problem of inferring road types from historical trajectories. In this paper, we propose a model-based approach to infer road types from taxis trajectories. We use a combined inference method based on stacked generalization, taking into account both the topology of the road network and the historical trajectories. The experiment results show that our approach can generate quality recommendations of road types for users to choose from.

1 Introduction

In recent years, crowdsourced map services has become a powerful competitor to public and commercial map service providers such as Google Maps. Different from commercial map services in which maps are produced from remote sensing images and survey data by a small group of professionals, crowdsourced maps are maintained by tens of thousands of registered users who continuously create and update maps using sophisticated map editors. Therefore, crowdsourced map services can be better in keeping up with recent map changes than existing commercial map services. For instance, it has been reported that OpenStreetMap (OSM) [1], the world's largest crowdsourced mapping project, can provide richer and more timely-updated map data than comparable proprietary datasets [2].

Similar to other crowdsourcing applications, crowdsourced map services rely on lots of volunteered works which are error-prone and can have severe consistency problems. In fact, providing quality maps is far more challenging than

S.S. Bhowmick et al. (Eds.): DASFAA 2014, Part II, LNCS 8422, pp. 392–406, 2014.

most crowdsourcing applications such as reCAPTCHA [3]. One major reason is that map objects (e.g., roads and regions) are usually complex, which makes it difficult to make map editors both feature-rich and user-friendly. To address this issue, a recent work proposed a map updating system called CrowdAtlas[4] to probe map changes via a large number of historical taxi trajectories. CrowdAtlas reduces the cost of drawing roads by generating the shapes of new/changed roads from trajectories automatically. The generated shapes of roads can be used as recommendations in a map editor. A contributor can then directly use the generated roads or slightly adjust them based on his/her own knowledge and experience.

Existing works only focus on generating the shapes of roads automatically. However, the metadata of roads is also important to many map-based applications such as navigation systems and travel planning services. Typical metadata of roads includes width, speed limit, direction restriction and access limit. These metadata can be effectively reflected by the type of road segments [1], which often includes: motorway, primary / secondary way, residential road, etc. For example, the speed limit is often higher of a motorway than a secondary way; a motorway or a primary way is often a two-way street, while a residential road may be a single-way street. Therefore, to contribute to a quality crowdsourced map service, users need to provide not only the shapes of the roads, but also the types of the roads. Consequently, to further reduce the cost of updating crowdsourced maps for users, it is necessary to automate the process of labeling road types.

There are many challenges of inferring the types of road segments. The types may be directly inferred from the topology of the road network, e.g., road segments with the same direction may have the same type. However, it is often not accurate, as in our experiments. Hence, in this paper, we combine the inference based on the topology of the road network, with the real trajectories of vehicles driving on the road segments. Trajectories can effectively show the types of road segments. For example, it is generally believed that vehicles drive faster on motorways than small roads, thus it is possible to infer the type of a road segment as motorway if the average driving speed of it is much faster than other road segments. However, the method of combining two inference methods, as well as the weights between them are very difficult to conduct. Moreover, the types between each other may be ambiguous, and there are no exact definitions to draw the borders of different types, which makes it difficult to find an accurate inference result.

Since it is private and difficult to obtain private vehicle data, we use the trajectories of taxis in this paper. There are many challenges using the trajectories of taxis. For example, the taxi data is very sparse due to the inaccuracy of the GPS device, and we have to filter the inaccurate data through preprocessing. Moreover, since taxis are only a part of all the vehicles in the city, the trajectories of taxis are biased, thus 1) not every road segment has been traversed; and 2) the density of taxis cannot directly show the traffic of the road segment [5].

[1] A road consists of a number of road segments which could be of different types. Therefore, the meaning of *road type* is essentially road segment type.

Table 1. Specifications of Trajectory Data

Data Type	**Description**
Taxi ID	Taxi registration plate number.
Timestamp	Timestamp of the sample point.
Latitude / Longitude	GPS location of the sample point.
Speed	Current speed of the taxi.
Angle	Current driving direction of the taxi.
Status	Indicator of whether the taxi is occupied or vacant.

In this case, we also use the topology of the road network to cover the shortage of using only trajectories. At last, a taxi has its own characteristics which may be different from other vehicles, and we have to consider them in the inference. In this paper, as many works do [6], we consider taxi drivers are experienced drivers, who often choose the fastest routes rather than shortest routes.

The main contributions of this paper are listed as follows:

- First, we propose a novel problem, as inferring the types of road segments;
- Second, we conduct a combined inference method considering both the topology of the road network, and the trajectories of taxis. The results show that our method is much better than baseline methods;
- Third, our method is flexible and scalable, where the models in our method can be replaced by other suitable models when handling different types of data;
- At last, we introduce large-scale real-life trajectories, and a real road network in a large city of China in our experiments.

The rest of this paper is organized as follows. Section 2 describes the dataset we use and formally defines the problem of road type inference. Section 3 presents the methodology in details. Section 4 presents the evaluations results. Section 5 outlines the related works and Section 6 concludes the paper.

2 Data Description and Problem Definition

2.1 Data Description

Our trajectory data is collected in Shenzhen, China in September, 2009 [7]. The data contains the trajectories of around 15,000 taxis for 26 days, and the sampling rate is around 20 seconds. A trajectory is represented as a series of sample points. The details are shown in Table 1.

Our road network data is provided by the government. There are 27 different types of road segments in our data, and they can be generally classified into 7 categories, based on the meanings of types defined by the government. The types

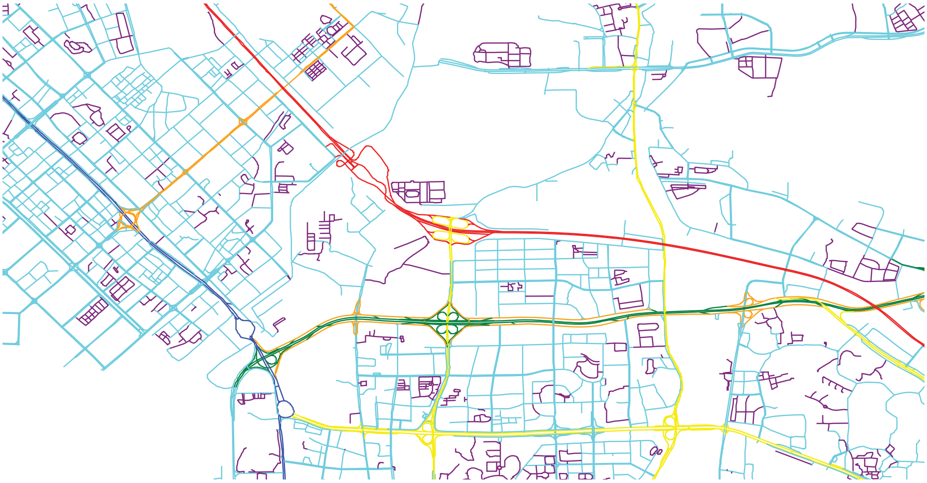


Fig. 1. The types of road segments from our data, the notations of colors are shown in Table 2

Table 2. Types of Road Segments

Type	Type ID	Color	Description
National Expressway	1-100	Red	National limited-access expressway.
City Expressway	100-200	Green	City expressway, often with relief roads.
Regular Highway	200-300	Blue	Regular highway.
Large Avenue	300-400	Orange	Direction-separated large avenue.
Primary Way	400-500	Yellow	The road that connects regions of the city.
Secondary Way	600-700	Aqua	The road that connects blocks of a region.
Regular Road	800-900	Purple	The road that constructs within a block.

are as shown in Table 2 and Figure 1. These types are used as ground truth to verify the accuracy of our model.

2.2 Problem Definition

Before introducing the problem definition, let us first introduce some terminologies used in this paper.

Definition 1 (Road segment). *A road segment τ is the carriageway between two intersections. An expressway or a large avenue may have two different road segments between two intersections, because they are different directions with limited-access.*

Definition 2 (Road network). *A road network $\{\tau_i\}_{i=1}^{n}$ consists of a set of road segments.*

Definition 3 (Taxi status). *The status of a taxi can be either occupied by a passenger or vacant for hiring. The status changes from vacant to occupied when the taxi picks up a passenger, and it changes from occupied to vacant when the taxi drops off the passenger.*

Definition 4 (Pick-up event / drop-off event). *A pick-up event is the event when the taxi picks up a passenger, and changes its status from vacant to occupied. Similarly, a drop-off event is the event when the taxi drops off the passenger, and changes its status from occupied to vacant.*

In this paper, we use a connectivity matrix $M^{n \times n}$ to represent the topology of the road network, where $m_{ij} \in M$ is the normalized angle between road segments i and j if they are connected, and 0 otherwise. Intuitively, the angle at which two neighboring road segments are connected highly determines the relation of the types of these two road segments. For example, in a common urban road network, if two road segments have an angle of 180°, they are often the same road with the same road name. When the angle becomes smaller, like 90°, they are often two different roads with different road names. Similarly, if we drive along a road, the road segments we traveled are often straight, i.e., with large angles up to 180°. When we change to another road, we will have to change the angle of our driving direction, and which will result in a smaller degree than the previous driving angle.

Nevertheless, such information may not always be accurate, since some road segments with the same type are also with 90° angle. Hence, the connectivity matrix is not a unique feature that can be used directly to identify the types, and it is necessary to combine the usage of the connectivity matrix with other features.

The problem, road type inference, is defined as inferring the types of road segments based on the road network and correspondent trajectories. The formal definition is shown in Definition 5.

Definition 5 (Road type inference). *Given a road network $R = \{\tau_i\}_{i=1}^{n}$, each road segment τ_i is associated with a feature vector $f_i = \langle f_i^1, f_i^2, \ldots, f_i^k \rangle$, and a connectivity variable $m_{i,j} = \Pr(\tau_i = y_{1\ldots l} | \tau_j = y_{1\ldots l})$ towards another road segment τ_j, infer the type $y_i \in Y$ for each τ_i, where $Y = \{y_i\}_{i=1}^{l}$ is the set of types.*

3 Methodology

An overview of our method is shown in Figure 2. In this paper, we firstly develop two weak predictors for the task of road type inference which exploit two different sources of information separately, and then introduce an ensemble approach that combines these two predictors to produce a strong predictor. In particular, we design a number of features to characterize each road segment, including topological features computed from the road network, and statistical features obtained from the historical trajectory data. A logistic regression model is then

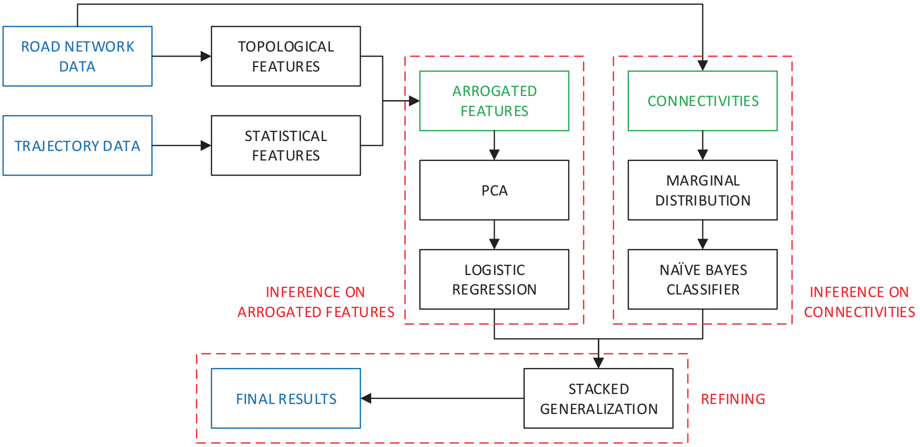


Fig. 2. An overview of our method

built upon these features to make a preliminary prediction of road segment types. However, due to the data sparsity problem, the statistical data for certain road segments may be too limited to build a reliable feature representation for those road segments. To overcome this problem, we note that there exists latent constraints on the relations of the types of two neighboring road segments depending on their connection angles. This motivates us to exploit the types of the neighboring road segments as auxiliary information for accurate road type inference. We realize this approach using a naive Bayes classifier based on the connectivity matrix built before. Finally, we combine these two predictors via stacked generalization so that their respective predictions can complement each other to achieve a more reliable prediction.

3.1 Inference on Arrogated Features

In this paper, we consider the trajectories as traveling along road segments, rather than roaming in an open area. However, the raw data of the trajectories collected from GPS devices are represented as a latitude-longitude pair with timestamp, without any road network information. Hence, we have to map the trajectories onto the road network via map matching. In this paper, we use the map matching method *ST-Matching* proposed in [4].

ST-Matching considers both the spatial geometric / topological structures of the road network, and the temporal features of the trajectories. ST-Matching is suitable to handle low-sampled trajectories, such like the taxi trajectories in this paper. It first constructs a candidate graph based on the spatial locations of the sample points of trajectories, and then generate the matched path based on the temporal features of trajectories. If any two matched road segments are not connected, it uses path-finding methods, such like shortest path method or most frequent path method [8], to generate an intermediate path that connects

Table 3. Features of a Road Segment

Topological	Road length	Statistical	Average speed of occupied taxis
	Cumulative flutter value		Density of occupied taxis
	♯ of neighbors		Density of vacant taxis
	♯ of adjacent road segments		♯ of pick-up events

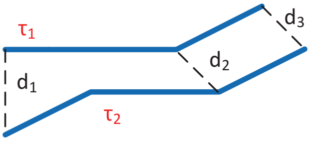


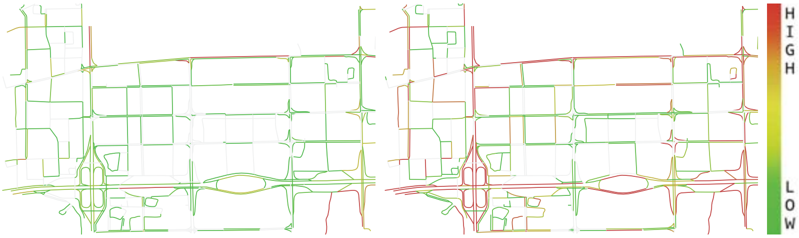
Fig. 3. The distance of two road segments. We will try to pair the road segment with fewer vertices (τ_1) towards the road segment with more vertices (τ_2). The distance of the two road segments is thus $avg(d_1, d_2, d_3)$.

them. The efficiency of ST-Matching is close to $O(nm \log m)$ using weak Fréchet distance.

In this paper, the features of road segments we use consist of two set of features: the topological features and the statistical features. The topological features are extracted from the road network, while the statistical features are extracted from the trajectory data. The details are shown in Table 3.

For the topological features in Table 3, road length and cumulative flutter value can effectively show the type of a road segment. For example, a large avenue is often limited-access, and there are few intersections within it during a long distance. Hence, according to the definition of a road segment in Definition 1, a road segment with long length is more likely to be a large avenue, or an expressway. Similarly, a road segment is more likely to be a large avenue when it is straighter, and less when it is twisted, based on our experience. Hence, we can use cumulative flutter value to show the types of road segments. For the neighbors in Table 3, we consider two road segments are neighbors when they are topologically connected. If a road segment has many neighbors, it is less likely to be a large avenue, because a large avenue, or an expressway, often has one or two neighbors as the entrance / exit of it. For the adjacent road segments in Table 3, we define adjacent as the distance between two road segments is less than a small distance (10 meters in this paper). The distance of two road segments is calculated via the average distance between each vertex of the polylines of the road segments, as shown in Figure 3. According to Definition 1, two adjacent road segments may have the same type especially when they have opposite directions.

The statistical features in Table 3 include the statistics of taxi trajectories that may reflect the types of road segments, based on our experience. These features are often time-dependent. For example, Figure 4 shows a clear difference



(a) Midnight 03:00-04:00 (b) Commuting 18:00-19:00

Fig. 4. Density of taxis on road segments in different time of a day [7]. It is obvious that the density when commuting is higher than midnight.

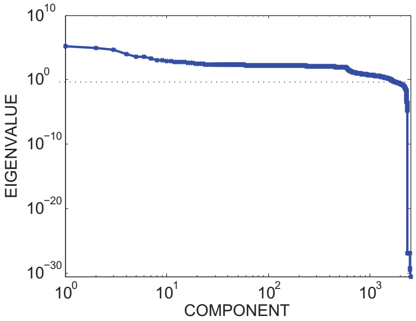
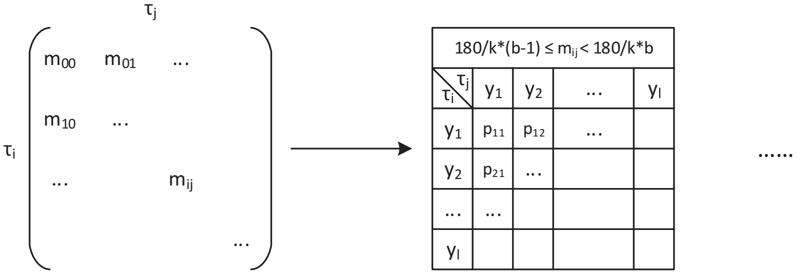


Fig. 5. Eigenvalues of the principal components in our experiments. The density of points on the plot represents the density of components with the correspondent eigenvalue.

in the density of taxis between different time slots. To account for such time-varying nature, instead of constructing a single value for each feature, we split the time domain into several time slots, and calculate the statistical features for each time slot.

To reduce the dimensionality of the feature vector, we apply principal component analysis (PCA) to find the low dimensional subspace that can well account for most variance in the data, and project each feature vector to this subspace to get its low-dimensional representation. The basis vectors of the subspace are those eigenvectors of the covariance matrix of the data set that correspond to large eigenvalues. The eigenvalues of the components of our data are shown in Figure 5. In Figure 5, it is clear that most of the components have a eigenvalue less than 1, thus we tend to select those principal components with eigenvalues larger than 1.

Based on the principal components, we can now apply a logistic regression model to infer the types of road segments. The details of the settings of our logistic regression model in the experiments are described in Section 4.



180/k*(b-1) ≤ mij < 180/k*b				
τj / τi	y1	y2	...	yl
y1	p11	p12	...	
y2	p21	...		
...	...			
yl				

Fig. 6. The conversion from connectivity matrix to marginal distribution. Each multinomial distribution $180/k*(b-1) \leq m_{ij} < 180/k*b$ indicates bucket b with the range of $(180/k*(b-1))^{\circ}$ to $(180/k*b)^{\circ}$, where k is the number of buckets. p_{ij} in each bucket means the probability of $P(\tau_i = y_i | \tau_j = y_j)$.

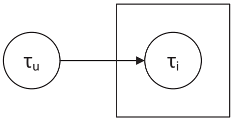


Fig. 7. The model of naive Bayes classifier in this paper. τ_u is the road segment we want to infer, and τ_i is a set of observed road segments, which are the neighbors of τ_u.

3.2 Inference on Connectivities

As mentioned before, to overcome the sparsity problem, we exploit the connectivity relationships between road segments as auxiliary information to help infer the types of road segments. Intuitively, it is observed that the type of a particular road segment has a strong indication of the possible types that its neighboring road segments can take, depending on the connection angles. Inspired by this observation, for a pair of road segments connected with each other, we model the type of one road segment as a multinomial distribution conditioned on the type of the other road segment as well as the connection angle. Equivalently, for each possible connection angle, we define for the target type a set of multinomial distributions, one for each source type. The relevant parameters can be framed as a matrix shown in Figure 6, where each row of the matrix specifies the target type distribution conditioned on a particular source type, and each matrix corresponds to a connection angle.

The graphical representation of our model is given in Figure 7, which turns out to be a naive Bayes classifier. The parent node specifies the type of the source road segment, and the child nodes are the types of neighboring road segments, one for each neighbor.

Our task then is to learn these matrix automatically from the data using maximum likelihood approach. After learning, the inference of the type of a particular road segment given the types of its neighboring road segments can be done using Bayes rule, as shown in Formula 1.

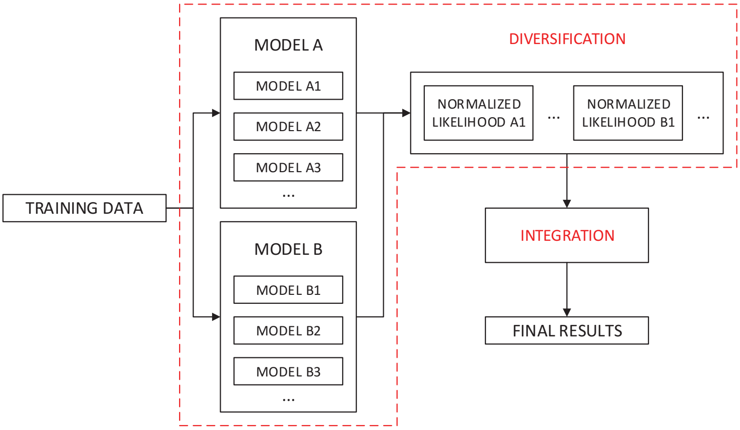


Fig. 8. The stacked generalization model used in this paper

$$\begin{aligned}
&P(\tau_u = y_u|\tau_1 = y_1, \tau_2 = y_2, \dots) \\
=&\frac{P(\tau_1 = y_1, \tau_2 = y_2, \dots|\tau_u = y_u) * P(\tau_u = y_u)}{\sum_{i=1}^{l} P(\tau_1 = y_1, \tau_2 = y_2, \dots|\tau_u = y_i) * P(\tau_u = y_i)} \\
=&\frac{P(\tau_1 = y_1|\tau_u = y_u) * P(\tau_2 = y_2|\tau_u = y_u) * \dots * P(\tau_u = y_u)}{\sum_{i=1}^{l} P(\tau_1 = y_1|\tau_u = y_i) * P(\tau_2 = y_2|\tau_u = y_i) * \dots * P(\tau_u = y_i)}
\end{aligned} \tag{1}$$

3.3 Refining

We use stacked generalization [9,10] based on logistic regression in this paper to perform refining, as shown in Figure 8. There are two phases of stacked generalization:

- Level 0: diversification through the use of different models. In this paper, we use logistic regression and naive Bayes classifier;
- Level 1: integration through meta-learning. In this paper, we use logistic regression.

Stacked generalization is one example of hybrid model combination, and it can effectively improve the accuracy of cross-validated models. Nevertheless, it is also possible to use other model combination techniques, such like random forests [11].

4 Evaluation

4.1 Metrics

In this paper, we evaluate our model via both the *accuracy* and the *expected reciprocal rank* [12]. For accuracy, it takes the maximum value of the likelihood

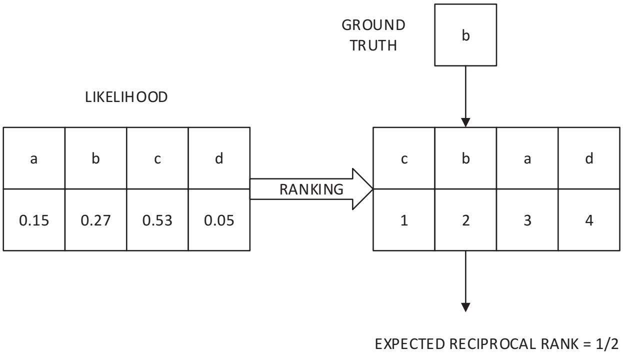


Fig. 9. An example of expected reciprocal rank used in this paper

of each road segment as the type of the road segment. For expected reciprocal rank, it evaluates each prediction as the derivative of the rank of its maximum likelihood, as an example shown in Figure 9.

In Figure 9, there are four likelihoods for the four types (a, b, c, and d) of a road segment. We rank the four likelihoods and then we get the ranks of c, b, a, and d are 1, 2, 3, and 4, respectively. Since the ground truth of the road segment is type b, we can find that the rank of type b is 2. Thus the expected reciprocal rank of the road segment is the derivative of 2, which is 1/2.

Expected reciprocal rank can be considered as a fairer metric comparing with accuracy. For example, if the likelihoods of the types of a road segment are $\langle a : 0.49, b : 0.48, c : 0.03\rangle$, it is actually difficult to determine whether the type of the road segment is a or b, but it is clear that the type is not c. However, using accuracy cannot give the bonus of such observation, since no matter the ground truth is a or b, the accuracy is similar (0.49 and 0.48). If we evaluate the likelihoods using expected reciprocal rank, it will give us a comprehensive distribution of the accuracy (either 1 or 1/2). Hence, the evaluation of expected reciprocal rank is fairer, which widens the gap between different likelihoods. Nevertheless, in our experiments, we will conduct both metrics.

Besides the two metrics introduced before, in this paper, we use random guess as the baseline method, and we assume the probability of guessing any type of a road segment is $1/l$. For the accuracy, the expectation of random guess is $1/l$. For the expected reciprocal rank metric, the expectation is $(1 * 1/l + 1/2 * 1/l + 1/3 * 1/l + \cdots + 1/l * 1/l)/(l * 1/l) = \sum_{i=1}^{l} 1/i/l$. In this paper, we have 7 different types of road segments as introduced in Table 2, thus the expectations are around 0.1429 and 0.3704, respectively.

4.2 Experiments

In this paper, we use four topological features and four statistical features as shown in Table 3. Since the statistical features are time-dependent, we split the statistical features by hours, as 24 features per day for a total of 26 days. Hence there are total 2,500 features.

There are total 44,793 road segments in this paper. In order to clearly show the accuracy of our model, we use 10-fold cross-validation to conduct the experiments, and each fold contains around 4,479 randomly selected road segments. In our experiments, we take one fold as test data, and the rest as training data.

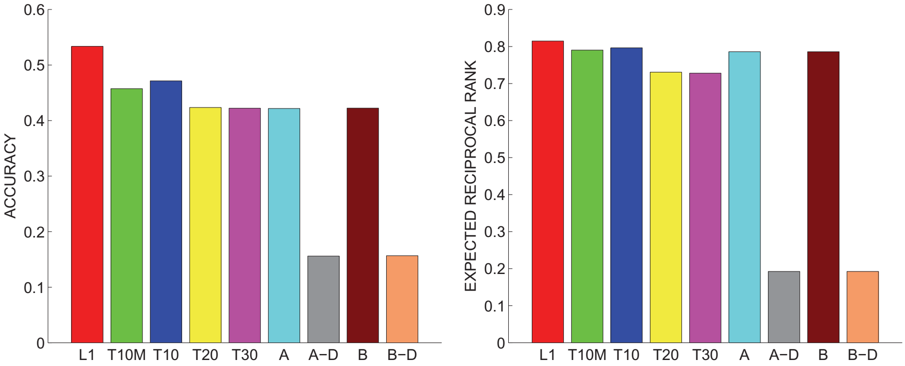
In order to clearly show the differences between different settings of different models, in this paper, we have adopted the following eight settings of models in level 0 diversification of stacked generalization:

- L0-LR-T10M: the multinomial logistic regression using the features through principal component analysis. This model uses the principal components with the top ten eigenvalues, but not the top one.
- L0-LR-T10/20/30: similar as L0-LR-T10M, but uses the principal components with the top 10/20/30 eigenvalues, including the top one.
- L0-BAYES-A: the naive Bayes classifier using the connectivity as introduced in Section 3.2 where $0 \leq m_{ij} < 180$. This model sets the equal initial probabilities of $P(\tau_u = y_u) = 1/l$ for each type.
- L0-BAYES-A-D: similar as L0-BAYES-A, but sets the initial probabilities of $P(\tau_u = y_u)$ being the statistical distribution of the types of training data. That is, if there are k road segments with type u among a total of n road segments, it sets $P(\tau_u = y_u) = k/n$.
- L0-BAYES-B: similar as L0-BAYES-A, but uses $0 \leq m_{ij} < 90$. If $m_{ij} > 90$, it sets $m_{ij} \leftarrow 180 - m_{ij}$. This metric is based on the assumption that two road segments tend to have the same type if they have a smaller acute angle.
- L0-BAYES-B-D: similar as L0-BAYES-B, but sets the initial probabilities of $P(\tau_u = y_u)$ being the statistical distribution of the types of training data as L0-BAYES-A-D.

In this paper, we use multinomial logistic regression model in level 1 integration of stacked generalization. The comparison of the evaluations of both level 0 and level 1 models is shown in Figure 10.

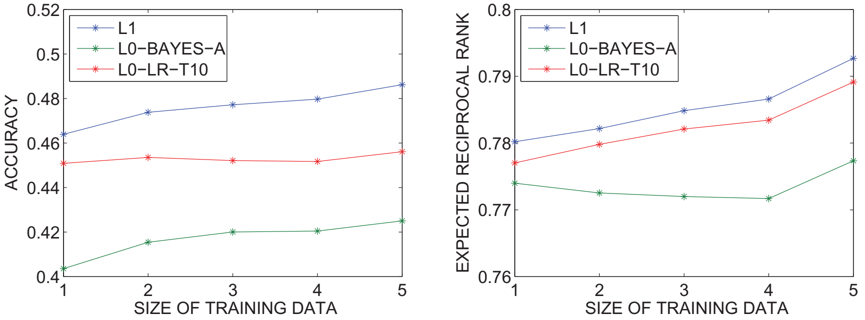
In Figure 10, it is clear that our method is always better than single level 0 model. Moreover, the average expected reciprocal rank of the final prediction reaches 0.81859, which is a very accurate result, and it is far better than baseline method. Some models in our results are not performing better than the baseline method, such like L0-BAYES-A and L0-BAYES-B. It shows that using the statistical distribution of the types of training data is a good choice, and it can indeed increase the accuracy.

In order to show the scalability of our model, we conduct 5 experiments based on different sizes of the training data, as shown in Figure 11. In Figure 11, it is clear that the performance of our model is not dropping dramatically when shrinking the size of the training data, comparing with the scalability of logistic regression and naive Bayes classifier. Thus, our model can be considered scalable upon the size of the training data.



(a) Average Accuracy (b) Average Expected Reciprocal Rank

Fig. 10. The comparison of level 1 prediction and all level 0 predictions, evaluated via different metrics. From left to right: L1, L0-LR-T10M, L0-LR-T10/20/30, L0-BAYES-A, L0-BAYES-A-D, L0-BAYES-B, L0-BAYES-B-D.



(a) Average Accuracy (b) Average Expected Reciprocal Rank

Fig. 11. The comparison of the experiments using different sizes of training data, based on L0-REG-T10 and L0-BAYES-A as level 0 models

5 Related Works

The information of a road network is an essential requirement to enable further analysis of urban computing [13]. The inference of a road network generally consists of two categories: inference from aerial imagery, and inference from trajectories. Both inference method aims to discovery the missing road segments in the data. The inference from aerial imagery is often based on pattern recognition methods [14], and it is often very difficult to find those road segments in the shadow of skyscrapers or forests. Hence, such methods often require a very high resolution color orthoimagery [15].

The inference from trajectories is an effective, efficient, and inexpensive method comparing with the inference from aerial imagery. Many works in this category are often based on the clustering of trajectories, such like k-means [16,17,18], kernel density estimation [19,20,21], trace merging [22], and some other methods, like TC1 [23]. These methods often first identify the trajectories in different clusters, and then apply fitting to the trajectories in the clusters. Some works also split the entire map into small grids to increase performance [20]. As mentioned in Section 1, most of these works only focus on identify the missing road segments that are not existed in the current road network, but few of them focus on the inference of the properties of road segments, such like the types introduced in this paper.

6 Conclusion

In this paper, we propose a novel problem, as identify the type of a road segment. To solve the problem, we introduce a combined model based on stacked generalization, using both the topology of the road network, and the knowledge learned from taxi trajectories. For level 0 diversification, we use 1) a multinomial logistic regression model on a set of arrogated features consists of both the topological features from the road network, and the statistical features from the taxi trajectories; and 2) a naive Bayes classifier based on the connectives of road segments. The experimental results show that our method is much better than the baseline method.

The model proposed in this paper is highly flexible and scalable. For level 0 diversification, it is possible to use different models despite of the models proposed in this paper, such like decision tree, expectation-maximization algorithm, kernel density estimation, etc. For level 1 integration, it is also possible to use different models, like support vector machine. Moreover, since the taxi trajectories we use in this paper are sparse and bias, it is also eligible to use other measurement methods, such like PageRank values [24], rather than the connectivities in level 0 models. A comparison of different models upon different trajectory datasets will be our future work.

Since the road network is often partially available in a crowd-sourcing platform, we only adopt supervised models in this paper. However, in some cases, we may not have any information of the road network besides the topology. Hence, inferring the types of road segments based on unsupervised / semi-supervised models is also a challenging problem.

Acknowledgment. This paper was supported in part by National Basic Research Program of China (973 Program) under grant no. 2014CB340304; Hong Kong, Macao and Taiwan Science & Technology Cooperation Program of China under grant no. 2012DFH10010; Nansha S&T Project under grant no. 2013P015; and NSFC under grant no. 61300031.

References

1. Haklay, M.M., Weber, P.: Openstreetmap: User-generated street maps. IEEE Pervasive Computing 7(4), 12–18 (2008)
2. Neis, P., Zielstra, D., Zipf, A.: The street network evolution of crowdsourced maps: Openstreetmap in germany 2007-2011. Future Internet 4(1), 1–21 (2012)
3. von Ahn, L., Maurer, B., Mcmillen, C., Abraham, D., Blum, M.: Recaptcha: Human-based character recognition via web security measures, 1465–1468 (2008)
4. Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y.: Map-matching for low-sampling-rate gps trajectories. In: GIS, pp. 352–361 (2009)
5. Liu, S., Liu, Y., Ni, L.M., Fan, J., Li, M.: Towards mobility-based clustering. In: KDD, pp. 919–928 (2010)
6. Yuan, J., Zheng, Y., Xie, X., Sun, G.: Driving with knowledge from the physical world. In: KDD, pp. 316–324 (2011)
7. Ding, Y., Liu, S., Pu, J., Ni, L.M.: Hunts: A trajectory recommendation system for effective and efficient hunting of taxi passengers. In: MDM, pp. 107–116 (2013)
8. Luo, W., Tan, H., Chen, L., Ni, L.M.: Finding time period-based most frequent path in big trajectory data. In: SIGMOD Conference, pp. 713–724 (2013)
9. Wolpert, D.H.: Stacked generalization. Neural Networks 5(2), 241–259 (1992)
10. Ting, K.M., Witten, I.H.: Stacked generalizations: When does it work? In: IJCAI (2), pp. 866–873 (1997)
11. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
12. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM, pp. 621–630 (2009)
13. Zheng, Y., Liu, Y., Yuan, J., Xie, X.: Urban computing with taxicabs. In: Ubicomp, pp. 89–98 (2011)
14. Hu, J., Razdan, A., Femiani, J., Cui, M., Wonka, P.: Road network extraction and intersection detection from aerial images by tracking road footprints. IEEE T. Geoscience and Remote Sensing 45(12-2), 4144–4157 (2007)
15. Chen, C.C., Shahabi, C., Knoblock, C.A.: Utilizing road network data for automatic identification of road intersections from high resolution color orthoimagery. In: STDBM, pp. 17–24 (2004)
16. Edelkamp, S., Schrödl, S.: Route planning and map inference with global positioning traces. In: Klein, R., Six, H.-W., Wegner, L. (eds.) Computer Science in Perspective. LNCS, vol. 2598, pp. 128–151. Springer, Heidelberg (2003)
17. Agamennoni, G., Nieto, J.I., Nebot, E.M.: Robust inference of principal road paths for intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems 12(1), 298–308 (2011)
18. Schrödl, S., Wagstaff, K., Rogers, S., Langley, P., Wilson, C.: Mining gps traces for map refinement. Data Min. Knowl. Discov. 9(1), 59–87 (2004)
19. Wang, Y., Liu, X., Wei, H., Forman, G., Chen, C., Zhu, Y.: Crowdatlas: Self-updating maps for cloud and personal use. In: MobiSys, pp. 27–40 (2013)
20. Davies, J.J., Beresford, A.R., Hopper, A.: Scalable, distributed, real-time map generation. IEEE Pervasive Computing 5(4), 47–54 (2006)
21. Biagioni, J., Eriksson, J.: Map inference in the face of noise and disparity. In: SIGSPATIAL/GIS, pp. 79–88 (2012)
22. Cao, L., Krumm, J.: From gps traces to a routable road map. In: GIS, pp. 3–12 (2009)
23. Liu, X., Biagioni, J., Eriksson, J., Wang, Y., Forman, G., Zhu, Y.: Mining large-scale, sparse gps traces for map inference: Comparison of approaches. In: KDD, pp. 669–677 (2012)
24. Yang, B., Kaul, M., Jensen, C.S.: Using incomplete information for complete weight annotation of road networks - extended version. CoRR abs/1308.0484 (2013)