Collective Traffic Prediction with Partially Observed Traffic History using Location-Based Social Media

Xinyue Liu
Dept. of Computer Science
Worcester Polytechnic Insitute
Worcester, MA, USA
xliu4@wpi.edu

Xiangnan Kong
Dept. of Computer Science
Worcester Polytechnic Insitute
Worcester, MA, USA
xkong@wpi.edu

Yanhua Li
Dept. of Computer Science
Worcester Polytechnic Insitute
Worcester, MA, USA
yli15@wpi.edu

ABSTRACT

Traffic prediction has become an important and active research topic in the last decade. Existing solutions mainly focus on exploiting the past and current traffic data, collected from various kinds of sensors, such as loop detectors, GPS devices, *etc.* In real-world road systems, only a small fraction of the road segments are deployed with sensors. For all the other road segments without sensors or historical traffic data, previous methods may no longer work. In this paper, we propose to use location-based social media, which captures a much larger area of the road systems than deployed sensors, to predict the traffic conditions. A simple but effective method called CTP is proposed to incorporate location-based social media semantics into the learning process. CTP also exploits complex dependencies among different regions to improve the prediction performances through collective inference. Empirical studies using traffic data and tweets collected in Los Angeles area demonstrate the effectiveness of CTP.

Keywords

Traffic Prediction; Collective Inference; Social Media; Data Mining

1. INTRODUCTION

With ever-increasing urban population and the slow development of transport infrastructure, traffic congestion has become a major issue in many cities. Excessive traffic congestion can cause serious problems for road users, such as travel delays, resource wasting, and pollution. In 2011, traffic congestion costs urban Americans 5.5 billion hours of travel delay, 2.9 billion gallons of extra fuel, for a total congestion cost of 121 billion dollars [7]. To alleviate these issues, there is a great need for building models to accurately predict traffic conditions in the near future.

Traffic prediction problem has been extensively studied [2,3,9]. Previous research mainly focus on exploiting historical traffic data, which are collected from sensors deployed on the roads [9]. It is usually assumed that all the historical traffic data are given beforehand, or all road segments to be predicted are deployed with sensors. However, in many real-world tasks, only a small fraction of the regions, such as the major highways, are deployed with sensors. While for all the other regions, historical and current traffic conditions are usually unknown, due to the lack of road sensors. *Location-based social media* (LBSM), such as Twitter and Foursquare, has become popular in the last decade. LBSM can provide abundant information about the road users in real-time, covering a wide range of geographic areas. For example, many car drivers can tweet through the dictation systems in smart phones (such as Siri) or through the modern car consoles. Passengers in the cars can also tweet using mobile devices, especially when being stuck in traffic. In LBSM, these messages are often associated with location tags, indicating the geolocations of the users. The contents of these messages may also be related to current traffic conditions, accidents or future events. In the left part of Fig. 1, we show an example of a user writing a tweet "Traffic jam on Storrow Drive, Boston, Massachusetts" tagged with his/her geolocation. By mining the *semantic* and *spatial* information from LBSM, we can effectively infer the future traffic conditions on a wide range of regions, including the road segments without sensors.

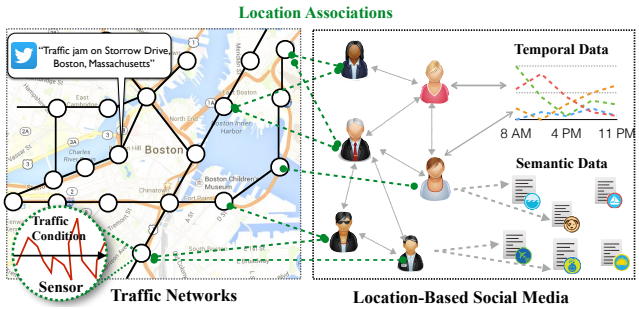
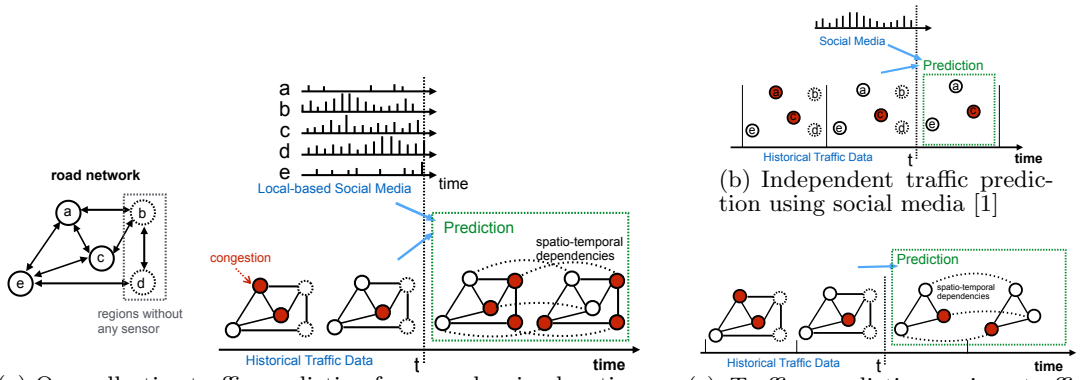


Figure 1: An illustration of using location-based social media to predict traffic conditions.

In this paper, we propose to use LBSM data to facilitate the traffic prediction process with partially observed traffic history. The problem of traffic prediction has not been studied in this context so far. Unlike prior works on traffic prediction [9] and social media sensing [1, 10], we assume that some geographical regions in the road system are not deployed with sensors, where the historical traffic data is not observed. The major research challenges of this paper are summarized as follows: **(a) Lack of Historical Traffic Data in Partial Regions:** One fundamental problem of

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983662



(a) Our collective traffic prediction framework using location-based social media

(b) Independent traffic prediction using social media [1]

(c) Traffic prediction using traffic data only [8]

Figure 2: Comparison of the different settings for traffic prediction. Each node of the road network represents a region, and the edges represent the spatial connections between different regions.

traffic prediction problem lies in the fact that many regions of the road network are not deployed with sensors. Most existing traffic prediction methods, as shown in Figure 2(c), mainly rely on historical traffic data to make predictions on future traffic conditions. For regions without sensors, these methods cannot be applied to predict future traffic. **(b) Sparsity of LBSM Information at Fine Granularities:** One problem of traffic prediction using LBSM is the sparsity of information with the required granularity of spatiotemporal resolution. Existing traffic prediction methods using social media mainly target on large spatial granularity, such as states of a country [10], or large temporal granularity, such as days or hours [1]. This is because usually only a small amount of LBSM content is generated in a small region during a small time window, as shown in Tab. 1. However, for traffic prediction tasks, we often need to target on small temporal granularity (60 minutes or less), as well as on small spatial granularity (a few square miles). At such granularity, we usually can only collect a small number of LBSM content. The features extracted from such less content are highly sparse, which may result in weak performance.

To alleviate the aforementioned issues, we introduce CTP (Collective Traffic Prediction) framework, as illustrated in Fig. 2(a), to predict the future traffic condition on a set of regions collectively by exploiting their spatial and temporal relationships. Unlike the conventional traffic prediction methods, CTP uses both LBSM data and partially observed traffic data in prediction. Fig. 1 shows the idea of using LBSM data to help predicting traffic conditions for different road segments and time. Furthermore, CTP also exploits three types of dependencies among spatiotemporal regions: (1) *intra-region temporal dependency*; (2) *inter-region temporal dependency*. (3) *inter-region spatial dependency*. By exploiting these dependencies, CTP can effectively predict traffic conditions for a set of inter-related regions collectively.

2. PROBLEM FORMULATION

In this section, we first introduce the notations that will be used throughout this paper. then we define the problem.

2.1 LBSM-augmented Traffic Network

Suppose we are given a LBSM-augmented traffic network, which can be represent as $G(\mathcal{R}, \mathcal{E}, \mathcal{X}, \mathcal{V})$. $\mathcal{R}$ and $\mathcal{E}$ are the set of regions and the set of edges in the network respectively. We define $\mathcal{R} = \{r_1, \ldots, r_n\}$, each $r_i \in \mathcal{R}$ represents a geographical region on the map. $\mathcal{E} \subset \mathcal{R} \times \mathcal{R}$ denotes the set of spatial connections among these regions through the traffic network. To tackle the problem of the lack of traffic history for partial regions, we extract information from LBSM data to improve our predictions. For each regions $r_i \in \mathcal{R}$, we have a set of feature vectors $\mathcal{X}_i = \{\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(m)}\}$, in which the superscripts with parentheses are called *temporal indices* and the subscripts are called *spatial indices*. Specifically, $\boldsymbol{x}_i^{(j)} \in \mathbb{R}^d$ denotes the LBSM feature vector collected in the region r_i in the time window j, where $j \in \{1, \cdots, m\}$. The details of how the LBSM feature vectors are extracted will be discussed in Sec. 4.2. Let $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ represent the set of LBSM-augmented features on all regions. For each region $r_i \in \mathcal{R}$, we also have a set of target variables $\mathcal{V}_i = \{v_i^{(1)}, \ldots, v_i^{(m)}\} \in \mathbb{R}^m$ indicating the traffic conditions, where $v_i^{(j)} \in \mathbb{R}$ denotes the average speed of all vehicles traveling in the region r_i within the time window j. $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_n\}$ represents sets of targets in all regions.

Table 1: Average # of tweets in each region under different spatiotemporal resolutions in our dataset.

Temporal Resolution	Spatial Resolution	Ave. #Tweets
12 hours	1×1	47,113
1 hour	1×1	3,926
1 hour	2×2	1,306
1 hour	3×3	554
1 hour	4×4	389
1 hour	30×30	15

2.2 Traffic Prediction with Fully Observed Traffic History

Most existing traffic prediction methods mainly rely on historical traffic data to make predictions on future traffic conditions. Suppose $\mathcal{V}^{(t)} = \{v_1^{(t)}, \cdots, v_n^{(t)}\}$ and $\mathcal{X}^{(t)} = \{\boldsymbol{x}_1^{(t)}, \cdots, \boldsymbol{x}_n^{(t)}\}$ denote the set of all traffic conditions and the set of LBSM feature vectors for all regions in time window t respectively. Hence, the prediction task is to predict $\mathcal{V}^{(t)}$ given the traffic history $(\mathcal{V}^{(t-l_1)}, \ldots, \mathcal{V}^{(t-1)})$, where $l_1 \in \mathbb{N}^+$ is called *traffic time lag* that specifies how many time slices of historical traffic data are used in the prediction. Suppose we set $l_1 = 2$, auto regression [8] learning a

regression model that predicts each $v_g^{(t)}$ independently as

$$v_g^{(t)} = \alpha + \beta_1 v_g^{(t-1)} + \beta_2 v_g^{(t-2)} \quad (1)$$

As stated before, this model is limited since it requires fully observed historical traffic data.

2.3 Traffic Prediction with Partially Observed Traffic History

In real-world traffic prediction, the traffic history can only be observed in partial regions. Our goal is to predict future traffic conditions of different regions (under fine granularity) in the road network based upon the partially observed historical traffic data. $U \subset \{1, \cdots n\}$ is the set of spatial indices for unobserved regions, $O = \{1, \cdots n\} - U$ is the set of spatial indices for observed regions. $\mathcal{R}_U \subset \mathcal{R}$ represents the set of regions, where the traffic history data are unobserved, $\mathcal{R}_O = \mathcal{R} - \mathcal{R}_U$ represents the set of regions where the traffic history data are observed. Suppose $\mathcal{V}_O^{(t)} = \{v_i^{(t)} | i \in O\}$ to denote the set of all traffic conditions of observed regions in time window t. Hence, the prediction task is to predict $\mathcal{V}^{(t)}$ given the partially observed traffic history $\{\mathcal{V}_O^{(t-l_1)}, \ldots, \mathcal{V}_O^{(t-1)}\}$. LBSM-aided approaches also consider the historical LBSM features $\{\mathcal{X}^{(t-l_2)}, \ldots, \mathcal{X}^{(t-1)}\}$ to predict the traffic conditions $\mathcal{V}^{(t)}$, where $l_2 \in \mathbb{N}^+$ is called the *LBSM time lag*. Note that the historical LBSM contents originate in unobserved regions $\mathcal{R}_U$ are available. Previous empirical studies [1,6] typically suggest optimal values for l_1 ranging from 2 to 6 and $l_2 = 2$. In this paper, we simply use $l_1 = 2$ and $l_2 = 2$. Thus, the inference problem in LBSM-aided traffic prediction is to estimate $v_g^{(t)}$ using $v_g^{(t-2)}$, $v_g^{(t-1)}$, $\boldsymbol{x}_g^{(t-2)}$ and $\boldsymbol{x}_g^{(t-1)}$ when $g \in O$, but using $\boldsymbol{x}_g^{(t-2)}$ and $\boldsymbol{x}_g^{(t-1)}$ only when $g \in U$. Previous approaches [1] require *i.i.d.* assumptions, in which each $v_g^{(t)}$ that has observed history is estimated independently as follows

$$v_g^{(t)} = \alpha + \beta_1 v_g^{(t-1)} + \beta_2 v_g^{(t-2)} + \gamma_1 \boldsymbol{x}_g^{(t-1)} \boldsymbol{m} + \gamma_2 \boldsymbol{x}_g^{(t-2)} \boldsymbol{m} \quad (2)$$

where $\boldsymbol{m} = \mathbb{R}^d$ is the transformation vector. As to the situation of $g \in U$, no historical traffic data is available, so $v_g^{(t)}$ can only be estimated independently as follows

$$v_g^{(t)} = \alpha + \gamma_1 \boldsymbol{x}_g^{(t-1)} \boldsymbol{m} + \gamma_2 \boldsymbol{x}_g^{(t-2)} \boldsymbol{m} \quad (3)$$

The only dependency that is considered in Eq. 2 is the dependency between the prediction target $v_g^{(t)}$ and its historical traffic conditions $v_g^{(t-1)}$ and $v_g^{(t-2)}$. However, in the real-world traffic networks, there are other types of dependencies exist between regions, which cannot be ignored. We consider other types of dependencies in Sec. 3.

We also note that the model proposed in [1] can only perform predictions under temporal resolution of 12 hours and spatial resolution of 1×1 (consider the target area as a single region). However, much finer spatiotemporal resolution is desired in real world traffic prediction application. As the spatiotemporal resolution goes finer, the amount of information can be extracted from LBSM data becomes sparser and sparser. From Tab. 1, we can see that under the resolution setting in [1], about 47,113 tweets can be collected in each time window. But when the spatial resolution increases to 30×30, only about 15 tweets can be collected for each region in each time window. It is challenging to build effective prediction models with such sparse LBSM information.

3. THE PROPOSED METHOD

To make better predictions given the sparsity of LBSM information, in Sec. 3.1-3.3, we explicitly consider three types of dependencies in the traffic networks, which address challenge (b) in Sec. 1. In Sec. 3.4, we describe how CTP simultaneously predict traffic conditions of multiple regions, which tackles challenge (a) in Sec. 1.

3.1 Intra-Region Temporal Dependency

The first type of dependency we consider is called *intra-region temporal dependency*, in which we discover the dependencies between the traffic conditions of the same region across different time slices. In traffic prediction, historical data are always considered as the primary factor since the traffic conditions across the timeline are not independent for any given location. For example, in the road networks, the probability of traffic congestion in region r_i in time window t should be high if we know that r_i was congested in $t-1$, and r_i is unlikely to be congested in t if we know r_i was not congested in $t-1$. So given a region r_g in a road network, by considering the *intra-region temporal dependency* alone, we will have the following prediction model

$$v_g^{(t)} = \alpha + \sum_{k=1}^{l_1} \beta_k v_g^{(t-k)} \quad (4)$$

where $v_g^{(t-k)}$ is the traffic history feature of $v_g^{(t)}$.

3.2 Inter-Region Temporal Dependency

Another type of dependency we consider is called *inter-region temporal dependency*, in which we discover the dependencies between the traffic conditions of the spatial related regions across different time windows. In other words, the traffic condition of any given region in time window t is correlated with the traffic conditions of its neighbors at the previous time windows $\{t-l_1, \ldots, t-1\}$. For example, the probability of traffic congestion in region r_i in time window t should be high if we know that most of its neighboring regions were congested in the previous time window $t-1$. And r_i is unlikely to be congested in time window t if we know most of r_i's neighboring regions are not congested in time window $t-1$.

We define overall traffic condition of neighboring regions of r_i in time window $t-l$ as follows

$$\mathcal{N}(v_i^{(t)}, l) = \frac{\sum_{(r_i, r_j) \in \mathcal{E}, r_j \in \mathcal{R}_O} v_j^{(t-l)}}{|\{r_j | (r_i, r_j) \in \mathcal{E}, r_j \in \mathcal{R}_O\}|} \quad (5)$$

where $\mathcal{E}$ is the set of edges we discussed in Sec. 2, and $|\{\cdot\}|$ denotes the number of elements in the set.

Hence, by considering *inter-region temporal dependency* alone, we have

$$v_g^{(t)} = \alpha + \sum_{k=1}^{l_1} \beta_k \mathcal{N}(v_g^{(t)}, k) \quad (6)$$

where $\mathcal{N}(v_g^{(t)}, k)$ is the inter-region temporal feature of $v_g^{(t)}$.

3.3 Inter-Region Spatial Dependency

The third type of dependency we consider is called *inter-region spatial dependency*, in which we discover the dependencies between the traffic conditions of spatially related regions within the same time window. For example, in the

traffic networks, the probability of traffic congestion in region r_i within the time window t should be high if we know that most of its neighboring regions are congested in the same time window t , and r_i is unlikely to be congested within t if we know most of its neighboring regions are not congested in t. Formally, we define the overall traffic condition of neighboring regions of r_i in the same time window t as follows

$$\tilde{\mathcal{N}}(v_i^{(t)}) = \frac{\sum_{(r_i,r_j)\in\mathcal{E}} v_j^{(t)}}{|\{r_j|(r_i,r_j)\in\mathcal{E}\}|} \quad (7)$$

Note that $\tilde{\mathcal{N}}(\cdot)$ averages all neighboring traffic conditions predicted by our iterative algorithm, even some regions are unobserved. This is because that CTP makes initial predictions of the traffic conditions in unobserved regions using LBSM features only, then update these predictions iteratively by considering the overall traffic condition of neighboring regions within the same time window. More details of our proposed framework will be discussed in Sec. 3.4. Hence, by considering *inter-region spatial dependency* alone, we have

$$v_g^{(t)} = \alpha + \beta\tilde{\mathcal{N}}(v_g^{(t)}) \quad (8)$$

where $\tilde{\mathcal{N}}(v_g^{(t)})$ is the inter-region spatial feature of $v_g^{(t)}$. For collective traffic prediction, we aim at inferring the traffic conditions of correlated regions simultaneously. Thus, when three types of dependencies are considered together with the LBSM feature vectors, for $g \in O$ we have

$$v_g^{(t)} = \alpha + \sum_{k=1}^{l_1}\beta_k v_g^{(t-k)} + \sum_{k=1}^{l_1}\gamma_k\mathcal{N}(v_g^{(t)},k) + \eta\tilde{\mathcal{N}}(v_g^{(t)}) + \sum_{k=1}^{l_2}\epsilon_k \boldsymbol{x}_g^{(t-k)}\boldsymbol{m}_{gk} \quad (9)$$

As to the case that $g \in U$, we have

$$v_g^{(t)} = \alpha+\sum_{k=1}^{l_1}\gamma_k\mathcal{N}(v_g^{(t)},k)+\eta\tilde{\mathcal{N}}(v_g^{(t)})+\sum_{k=1}^{l_2}\epsilon_k\boldsymbol{x}_g^{(t-k)}\boldsymbol{m}_{gk} \quad (10)$$

Other than considering the two more types of dependencies, our method is also different from [1] by learning separate transformation vectors $\boldsymbol{m}_k$ for each $\boldsymbol{x}_g^{(t-k)}$.

3.4 Iterative Framework

With the dependencies described above, we now present the CTP algorithm, which is inspired by ICA (Iterative Classification Algorithm) [5]. CTP algorithm is also summarized in Tab. 2. It contains the following key steps:

Training: The traffic history features, inter-region temporal features and inter-region spatial features are extracted from the traffic dataset and appended to the LBSM features to form the extended training set. Note that the inter-region spatial features are extracted using $\mathcal{N}(\cdot)$ instead of $\tilde{\mathcal{N}}(\cdot)$. This is because part of regions in traffic network is unobserved, so the traffic history in these regions is still unknown. $\mathcal{N}(\cdot)$ retrieves the overall neighboring traffic condition from observed regions. After the extended training set is built, we can apply the base learner on the data to obtain a local model f.

Input:
$\{\mathcal{X}^{(1)},\ldots,\mathcal{X}^{(t-1)}\}$: set of attribute vectors
$\{\mathcal{V}_O^{(1)},\ldots,\mathcal{V}_O^{(t-1)}\}$: set of traffic conditions of observed regions
$\mathcal{R}_O$: set of observed regions
$\mathcal{R}_U$: set of unobserved regions
$\mathcal{E}$: set of edges connecting all neighboring regions
Max_It: maximum # of iteration
A: a base learner for the local model

Training:
- Learn the local model:
1. Extended training set $\mathcal{D} = \left\{(\bar{\boldsymbol{x}}_i^{(j)}, v_i^{(j)})\right\}$ for all $2 \le j \le t-1$ and $i \in \{k|r_k \in \mathcal{R}_O\}$ where $\bar{\boldsymbol{x}}_i^{(j)} = \left(\boldsymbol{x}_i^{(j-2)}, \boldsymbol{x}_i^{(j-1)}, v_i^{(j-2)}, v_i^{(j-1)}, \mathcal{N}(v_i^{(j)},2), \mathcal{N}(v_i^{(j)},1), \mathcal{N}(v_i^{(j)},0)\right)$
2. Let $f = A(\mathcal{D})$ be the local model

Bootstrap:
- Estimate the label sets:
1. Produce an estimated value $\hat{v}_i^{(t)}$ for $v_i^{(t)}$ as follow:
for $r_i \in \mathcal{R}_O$:
$\hat{v}_i^{(t)} = f\left((\boldsymbol{x}_i^{(t-2)}, \boldsymbol{x}_i^{(t-1)}, v_i^{(t-2)}, v_i^{(t-1)}, \mathcal{N}(v_i^{(t)},2), \mathcal{N}(v_i^{(t)},1), 0)\right)$
for $r_i \in \mathcal{R}_U$:
$\hat{v}_i^{(t)} = f\left((\boldsymbol{x}_i^{(t-2)}, \boldsymbol{x}_i^{(t-1)}, 0, 0, \mathcal{N}(v_i^{(t)},2), \mathcal{N}(v_i^{(t)},1), 0)\right)$

Iterative Inference:
- Repeat until convergence or #iteration$> Max_It$
1. Construct the extended testing instance:
for $r_i \in \mathcal{R}_O$:
$\bar{\boldsymbol{x}}_i^{(t)} = \left(\boldsymbol{x}_i^{(t-2)}, \boldsymbol{x}_i^{(t-1)}, v_i^{(t-2)}, v_i^{(t-1)}, \mathcal{N}(v_i^{(t)},2), \mathcal{N}(v_i^{(t)},1), \tilde{\mathcal{N}}(v_i^{(t)})\right)$
for $r_i \in \mathcal{R}_U$
$\bar{\boldsymbol{x}}_i^{(t)} = \left(\boldsymbol{x}_i^{(t-2)}, \boldsymbol{x}_i^{(t-1)}, 0, 0, \mathcal{N}(v_i^{(t)},2), \mathcal{N}(v_i^{(t)},1), \tilde{\mathcal{N}}(v_i^{(t)})\right)$
2. Update the estimated values $\hat{v}_i^{(t)}$ for $v_i^{(t)}$ on each testing instance:
$\hat{v}_i^{(t)} = f(\bar{\boldsymbol{x}}_i^{(t)})$

Output:
$\hat{\mathcal{V}}^{(t)} = \{\hat{v}_i^{(t)}|r_i \in \mathcal{R}_O \text{ and } r_i \in \mathcal{R}_U\}$

Table 2: The CTP algorithm

Bootstrap: Two problems are raised before CTP infers the future traffic conditions. The first one is that how to initialize the inter-region spatial features while all traffic conditions in the future time window t are still unknown. CTP uses *bootstrap* to tackles this problem, in which it initializes all the inter-region spatial features as zeros [4]. The second problem is that how to initialize the traffic history features for unobserved regions, *i.e.* $v_i^{(t-1)}$ where $i \in U$. CTP also initializes these unknown traffic history features as zeros (*i.e.* the universal average value of traffic condition in the de-trended dataset). Then the local model f can be applied to make the initial predictions on all regions in the future time window t.

Iterative Inference: In this step, we first update the inter-region spatial features based upon the initial predictions, then we apply f on the extended feature vectors with updated inter-region spatial features, which updates the predictions. Again, we can use these updated predictions to further update the inter-region spatial features. This iterative procedure continues until the predictions converge or a maximum number of iteration has been reached.

4. EXPERIMENTS

4.1 Data Collection

We collected the traffic data set from traffic detectors in the Greater Los Angeles area between October 19, 2014 and November 28, 2014 using the California Performance Measurement System(PeMS[1]). This collection results in a total number of 31,102,272 entries of traffic records. We also collected tweets from the same area during the same time range

[1]http://pems.dot.ca.gov

using the Twitter streaming API[2] with a geolocation filter defining the bounding box of the Greater Los Angeles area. For each tweet, we store all meta-data and post content. This collection results in a total number of 2,648,446 tweets.

4.2 Data Preparation

In the PeMS traffic dataset, each data entry has the format $(t, lat, long, v)$, which keeps track of the average speed v of all vehicles passing by the detector located at coordinates $(lat, long)$ within the 1-hour time window t. Different from [1], which was designed for a lower spatial resolution by treating the target area as a single region, our experiments aim at a much finer spatial granularity. We evenly divide the target area into $r \times r$ grids, where r is the spatial resolution parameter. Then we transform the dataset into the format $(t, r_g, v_g^{(t)})$, which keeps track of the average speed $v_g^{(t)}$ of all vehicles traveling in region (grid) r_g in time window t. To exclude the periodic fluctuations in the traffic data, we follow the approach in [1] to de-trend each $v_g^{(t)}$.

As to the Twitter dataset, for each time window t and region g, we put the contents of all tweets originate in t and from g together and cast them into the space of stemmed words (stop-words removed), which generates a non-negative sparse vector $\boldsymbol{x}_g^{(t)}$. These vectors are used as the LBSM semantic features. Stacking all such vectors in time window t together, we obtain a semantic feature matrix $\mathbf{F}^{(t)} \in \mathbb{R}^{n \times d}$, where n is the number of regions in the graph, and d is the number of stemmed words.

4.3 Compared Methods

We compare the proposed CTP with following methods:

- **Tweets Semantics Only (TwSeO)** [1]: TwSeO uses the semantic feature matrices $\mathbf{F}^{(t-1)}$ and $\mathbf{F}^{(t-2)}$ to predict the average speed of each region in time window t. The model used by TwSeO is shown in Eq. 3. The original method proposed in [1] requires both tweet semantic data and historical traffic data. However, in our experiments, the historical data is not available in some regions. Thus, we compare CTP with TwSeO method which can be considered as a degenerated version of the method proposed in [1].

- **Traffic Data Only (TDO)** [8]: Due to the lack of historical traffic data, TDO always take the universal average speed of the entire traffic network, which is 0 in our de-trended traffic dataset, as the prediction for the unobserved regions. As to the observed regions, TDO uses the auto-regression model [8] shown in Eq. 1.

- **Traffic Data Only with Complete Traffic History(TDO-floor)** [8]: TDO-floor predicts future traffic condition by using the auto regression model proposed in [8] with fully observed historical traffic data. TDO-floor serves as a lower-bound baseline since it assumes all historical traffic data in the road network is observed.

The maximum number of iteration in the inference procedure of CTP are all set as 20.

[2]https://dev.twitter.com/streaming/overview

4.4 Experimental Settings

We partition the data into two parts, with the beginning $(u-1)/u$ $(u = 3, \ldots, 7)$ as the training set and the remaining as the test set. Moreover, k-fold cross-validation is used to randomly sample $1/k$ regions as unobserved, *i.e.* regions without historical traffic data. Note that all models are required to make predictions on all regions in test set, including the unobserved ones. Various spatial resolutions ($r \times r$, where $r = 5, 10, \ldots, 30$) and different fractions of unobserved regions($1/k$, where $k = 2, 3, 4, 5$) are tested respectively.

4.5 Evaluation Metrics

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used to evaluate the performance of compared methods, the definitions are as follows

$$\text{RMSE} = \sqrt{\frac{1}{m-l}\sum_{t=l+1}^{m}\sum_{g=1}^{n}(\hat{v}_g^{(t)} - v_g^{(t)})^2}$$

$$\text{MAE} = \frac{1}{m-l}\sum_{t=l+1}^{m}\sum_{g=1}^{n}|\hat{v}_g^{(t)} - v_g^{(t)}|$$

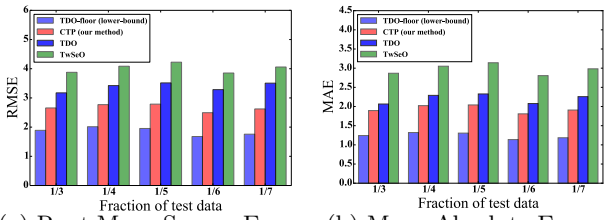
, where $l = \max(l_1, l_2)$ and $\hat{v}_g^{(t)}$ is the estimated value of $v_g^{(t)}$.

4.6 Results

We first study the effectiveness of our proposed CTP method on traffic prediction. Fig. 3 shows the comparison of CTP and other methods in terms of RMSE and MAE, where we set $r = 5$ and $k = 2$. The results under other spatial resolutions and ratio of unobserved regions are similar and will be discussed in Sec. 4.7. We can observe that the TDO-floor method outperforms other three methods in terms of both metrics for all training/test ratios. This is because TDO-floor uses the complete traffic history data while other compared methods only have traffic history of 50% regions. Among all three methods using only partially observed traffic history, our proposed CTP method performs the best in terms of both metrics. This improvement can be explained by the fact that CTP leverages the semantic features extracted from LBSM data as well as exploits the spatiotemporal dependencies between regions. Due to the novel experimental setting used in this work, the method proposed in [1] no longer works as it relies on complete traffic history. A degenerated version of their method, TwSeO, which only uses LBSM semantic features is compared here. We can observe that TwSeO is outperformed by all compared methods, which indicates that using tweets semantics alone can not achieve satisfying performance at fine spatiotemporal granularity. This observation shows the limitation of [1] under the assumption that historical traffic data of some regions is not available.

4.7 Parameter Study

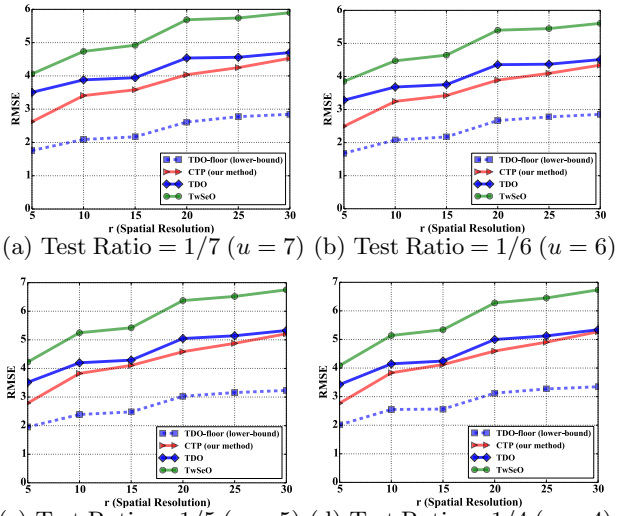
We now study the effect of the spatial resolution parameter in our experimental setting. Fig. 4 compares the performance of all four methods at different spatial resolutions. Each figure in Fig. 4 shows the results under different ratios of test data and we set $k = 2$ for all of them. Specifically, we set the spatial resolution parameter r to $\{5, 10, \ldots, 30\}$ and show the RMSE of each method (MAE scores are similar and we omit them due to the page limitation). The higher spatial



(a) Root Mean Square Error (b) Mean Absolute Error

Figure 3: Comparison of different methods

resolution (larger r) produces smaller region, which results in sparser LBSM data in each region. We observe that CTP achieves the best performance consistently under different spatial resolutions compared to other methods using only partially observed traffic history. Even under large r, where the LBSM information is extremely sparse (about 15 tweets per region in a time window), CTP can reduce the prediction errors compared to TDO. Meanwhile, TwSeO is consistently outperformed by TDO and the gap between them becomes larger when r increases. This observation demonstrates the effectiveness of exploiting the spatiotemporal dependencies.



(a) Test Ratio = 1/7 ($u = 7$) (b) Test Ratio = 1/6 ($u = 6$)

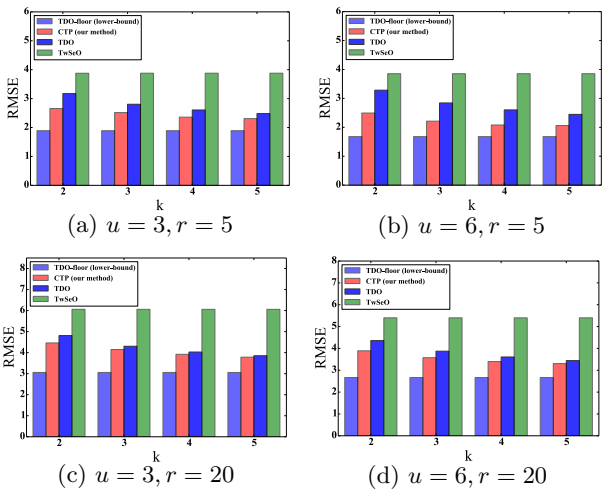
(c) Test Ratio = 1/5 ($u = 5$) (d) Test Ratio = 1/4 ($u = 4$)

Figure 4: Different spatial resolutions

We also study the effect of parameter k, which indicates the ratio of unobserved regions in our experiments. To do this, we set k equals to $\{2, 3, 4, 5\}$ respectively, which is equivalent to set 50%, 33.3%, 25%, and 20% of all regions in the target area as unobserved (with no historical traffic data). The results in Fig. 5 show that CTP achieves better performance compared to TwSeO and TDO consistently under different values of k. We note that the performances of TwSeO and TDO-floor do not change under different values of k. This is because TwSeO only use the LBSM semantic feature matrices $\mathbf{F}^{(t-1)}$ and $\mathbf{F}^{(t-2)}$ that are not affected by k, and TDO-floor uses the complete traffic history regardless of the value of k.

5. CONCLUSION

In this paper, we studied the problem of collective traffic prediction with LBSM information. Our work is different from the conventional traffic prediction methods in several aspects. The proposed CTP method extracts and



(a) $u = 3, r = 5$ (b) $u = 6, r = 5$

(c) $u = 3, r = 20$ (d) $u = 6, r = 20$

Figure 5: Different ratios of unobserved regions

exploits the semantic and spatial information in LBSM. Besides, CTP also makes use of the complex dependencies exist in the road traffic networks to tickle the sparsity problem in LBSM data. With these novelties, CTP can even make predictions in the regions without historical traffic data under much finer temporal and spatial granularity. Experimental results on traffic data and Twitter data collected from the Los Angeles area of California demonstrate that CTP improves the performance of traffic prediction.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation through grant CNS-1626236.

7. REFERENCES

[1] J. He, W. Shen, P. Divakaruni, L.Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *IJCAI*, pages 1387–1393, 2013.

[2] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *UAI*, 2005.

[3] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. Van Lint. A genetic algorithm-based method for improving quality of travel time prediction intervals. *Transportation Research Part C: Emerging Technologies*, 19(6):1364–1376, 2011.

[4] X. Kong, X. Shi, and P.S. Yu. Multi-label collective classification. In *SDM*, pages 618–629. SIAM, 2011.

[5] Q. Lu and L. Getoor. Link-based classification. In *ICML*, pages 496–503, Washington, DC, 2003.

[6] W. Min and L. Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616, 2011.

[7] D. Schrank, B. Eisele, and T.Lomax. TTI's 2012 urban mobility report. *Texas A&M Transportation Institute. The Texas A&M University System*, 2012.

[8] B.L. Smith and M.J. Demetsky. Traffic flow forecasting: comparison of modeling approaches. *Journal of transportation engineering*, 123(4):261–266, 1997.

[9] B.L. Smith, B.M. Williams, and R.K. Oswald. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4):303–321, 2002.

[10] J. Xu, A. Bhargava, R. Nowak, and X. Zhu. Socioscope: Spatio-temporal signal recovery from social media. In *Machine Learning and Knowledge Discovery in Databases*, pages 644–659. Springer, 2012.