Context-Aware Online Spatiotemporal Traffic Prediction

Jie Xu[1], Dingxiong Deng[2], Ugur Demiryurek[2],Cyrus Shahabi[2], Mihaela van der Schaar[1]
Department of Electrical Engineering, University of California, Los Angeles[1]
Department of Computer Science, University of Southern California, CA[2]
{jiexu, mihaela}@ucla.edu[1], {dingxiod, demiryur, shahabi}@usc.edu[2]

***Abstract*—With the availability of traffic sensors data, various techniques have been proposed to make congestion prediction by utilizing those datasets. One key challenge in predicting traffic congestion is how much to rely on the historical data v.s. the real-time data. To better utilize both the historical and real-time data, in this paper we propose a novel online framework that could learn the current situation from the real-time data and predict the future using the most effective predictor in this situation from a set of predictors that are trained using historical data. In particular, the proposed framework uses a set of base predictors (e.g. a Support Vector Machine or a Bayes classifier) and learns in real-time the most effective one to use in different contexts (e.g. time, location, weather condition). As real-time traffic data arrives, the context space is adaptively partitioned in order to efficiently estimate the effectiveness of each predictor in different contexts. We obtain and prove both short-term and long-term performance guarantees (bounds) for our online algorithm. Our experiments with real-world data in real-life conditions show that the proposed approach significantly outperforms existing solutions.**

I. Introduction

Traffic congestion is caused when the traffic demand approaches or exceeds the available capacity of the traffic system. Fortunately, due to thorough sensor instrumentations of road networks, a large volume of real-time and historical traffic data at very high spatial and temporal resolutions has become available. One challenge in predicting traffic is how much to rely on the historical data vs. the real-time data. Previous studies [1] [2] showed that depending on the situation one dataset may be more useful than the other but there is no holistic approach on when to switch from one dataset to the other for a more effective prediction. This becomes even more challenging when considering different causes for congestions. Our main thesis in this paper is that we try to learn the current situation from the real-time data and then predict the future by matching the current situation to the most similar situation we have seen in the past (using the historical data). We achieve this in two phases. First, in an off-line phase, we categorize the historical data into classes of similar "situations", for each of which we train one or more predictors (e.g., an SVM and a Bayes-Classifier). Next, in an on-line phase, suppose we would like to predict speed at a location X. We learn which situation the location X is similar with and choose the most effective predictors. To identify the most similar situation and select the most effective predictor, we utilize the context information of the traffic incidents such as location, time, weather condition, number of lanes, area type (e.g., business district, residential), etc. The context can be meta-data but can also be subsets or functions of features of the data. The context space is adaptively partitioned online based on the dimensions and the domain of each feature in order to efficiently estimate the "reward" of each predictor in different contexts where the reward is calculated based on how accurate each predictor has been in predicting, say, speed value, given the actual speed values we have observed in the recent past via the real-time data.

Our approach has three important byproducts. First, since the reward is continuously being updated/aggregated, we are utilizing what we learn in real-time to adapt to both short-term and long-term changes. Second, our approach is agnostic to the congestion cause. Finally, since location and time are two features of our context space, our approach is inherently spatiotemporal and takes into consideration the sensor readings that are spatially and temporally closest to the target location.

The majority of traffic prediction techniques focused on predicting traffic in typical conditions (e.g., morning rush hours) [1], [3]–[5], and more recently in the presence of accidents (e.g., [3], [6]. Existing techniques are only applicable to predict one of the scenarios. Moreover, the model used for prediction is learned offline and thus cannot adapt (and learn from) dynamically changing traffic conditions. In our model, the system has access to many predictors (or classifiers). When there are many classifiers, another approach is to use ensemble learning [7] [8] [9] [10], including weighted majority based algorithms [11] [12] [13]. However, most of them provide algorithms which are asymptotically converging to an optimal or locally-optimal solution without providing any rates of convergence. On the contrary, we prove regret bounds that hold uniformly over time; the proving technique is adapted from contextual multi-armed bandits (MAB) framework [14]. Since the learner can observe the rewards of all classifiers, the considered problem is more related to prediction with expert advice (PWEA) [15]. However, we focus on how contextual specialization of classifiers can be discovered over time to create a strong classifier from many weak classifiers while existing works ignore the context information.

II. Problem Formulation

A. Problem setting

We consider a set of locations $\mathcal{L}$ where traffic sensors are deployed. These locations can be either on the highways or arterial streets. In the following, we will fix a location l_o and focus on predicting the traffic at this location when incidents occur in its vicinity $\mathcal{V}(l_o)$. We consider an infinite horizon discrete time system $t = 1, 2, ...$ in which traffic incidents (e.g. accidents, road construction, road closure and etc.) occur over time at some location $l_e^t \in \mathcal{V}(l_o)$. Note that here t is not the absolute time but only represents the relative sequence order of traffic incidents. The incident causes immediate effects on the traffic at location l_e^t, which is summarized in $x^t \in \mathcal{X}$. The goal is to predict the traffic impact at location l_o in the near future and/or in the long term.

The system maintains a set of K (weak) base predictors $f \in \mathcal{F}$ that can take input of x^t and output the traffic prediction $f(x^t) \in \mathcal{Y}$ for location l_o. The prediction space $\mathcal{Y}$ depends on the specific objectives of the system. These base predictors are constructed using historical data before the system operates and hence, they may not perform well online since the transportation environment may change. Our idea is to build a strong ensemble predictor using these possibly weak base predictors by exploiting the incident context information, such as incident time, location and other attributes. This context information can better characterize the "situation" of an incident than other features of the traffic data such as traffic speed. We use $\theta^t \in \Theta$ to denote the context information associated with the t-th incident where Θ is a D-dimensional space and D is the number of types of context used. Without loss of generality, we normalize the context space Θ to be $[0, 1]^D$. For each incident, based on the context information, the system selects the prediction of one of the base predictors as the final traffic prediction, denoted by $y^t \in \mathcal{Y}$. Note again that since t is only the sequence order of the incidents, y^t represents the *future* traffic of the t-th incident. Eventually, the real traffic pattern at location l_o is revealed for the t-th incident, denoted by $\hat{y}^t \in \mathcal{Y}$. By comparing the prediction y^t and the realization $\hat{y}^t$, a reward r^t is obtained according to any general reward $r^t = R(y^t, \hat{y}^t)$. The key idea is that even though the base predictors do not work well in all contexts, they may work well in certain contexts. By exploiting the multi-predictor diversity, a strong ensemble predictor can be constructed.

B. Performance metric for our algorithm

The goal of the system is to maximize the prediction rewards by using the best predictor for each traffic incident. Since we do not have the complete knowledge of the performance of all base predictors for all contexts in the online environment, we will develop online learning algorithms that learn to select the best predictors for different contests

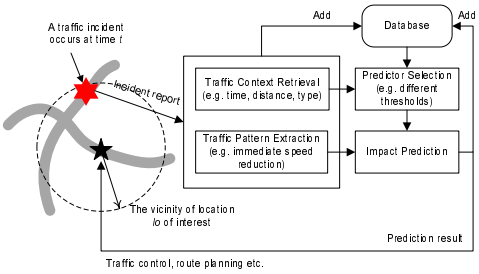

Figure 1. Traffic impact prediction system block.

over time. The benchmark when evaluating the performance of learning algorithms is the optimal solution in which the system follows the prediction of the best predictors in $\mathcal{F}$, i.e. the predictor with the highest expected reward for context $\theta(t)$, at time t. Let $\pi_f(\theta) = E\{R(f(x), \hat{y})|\theta\}$ be the expected reward (e.g. accuracy) of a predictor f conditional on the context information θ. Given context θ, the optimal strong predictor followed by the complete knowledge benchmark is $f^*(\theta) := \arg\max_{f \in \mathcal{F}} \pi_f(\theta), \forall \theta \in \Theta$.

Let σ be a learning algorithm and $f^{\sigma(t)}$ be the predictor selected by σ at time t, then the regret of learning by time T is defined as $Reg(T) := \sum_{t=1}^{T} \pi_{f^*(\theta^t)}(\theta^t) - E\left[\sum_{t=1}^{T} R(f^{\sigma(t)}(x^t), \hat{y}^t)\right]$ where the expectation is taken with respect to the randomness of the prediction, true traffic pattern and predictors selected. The regret characterizes the loss incurred due to the unknown system dynamics and gives the convergence rate of the total expected reward of the learning algorithm to the value of the optimal strong predictor. Any algorithm whose regret is sublinear, i.e. $Reg(T) = O(T^\gamma)$ such that $\gamma < 1$, will converge to the optimal solution in terms of the average reward. The regret of learning also gives a measure for the rate of learning. l

III. Context-Aware Adaptive Traffic Prediction

A. Algorithm description

First we introduce several useful concepts for describing the proposed algorithm: i) **Context subspace**. A context subspace C is a subspace of the entire context space Θ, i.e. $C \subseteq \Theta$. In this paper, we will consider only context subspaces that are created by uniformly partitioning the context space on each dimension, which is enough to guarantee sublinear learning regrets. Thus, each context subspace is a D-dimensional hypercube with side length being 2^{-l} for some l. We call such a hypercube a level-l subspace. ii) **Context space partition**. A context space partition $\mathcal{P}$ is a set of non-overlapping context subspaces that cover the entire context space. Since our algorithm will adaptively partition the context space by adaptively removing subspaces from the partition and adding new subspaces into the partition, the context space partition is time-varying depending on the context arrival process of the traffic incidents. Initially, the

context space partition includes only the entire context space, i.e. $\mathcal{P}^0 = \{\Theta\}$. iii) **Active context subspace**. A context subspace C is active if it is in the current context space partition $\mathcal{P}^t$, at time t. For each active context subspace $C \in \mathcal{P}^t$, the algorithm maintains the sample mean reward estimates $\bar{r}_f^t(C)$ for each for the predictor for the context arrivals to this subspace. For each active subspace $C \in \mathcal{P}^t$, the algorithm also maintains a counter M_C^t that records the number of context arrivals to C.

The algorithm works as follows (See Algorithm 1). We will describe the algorithm in two parts. The first part (line 3 - 8) is the predictor selection and reward estimates update. When an incident occurs, the data x^t representing the immediate traffic sensor data along with the incident context information θ^t are sent to the system. The algorithm first checks to which active subspace $C^t \in \mathcal{P}^t$ in the current partition $\mathcal{P}^t$ the context θ^t belongs (line 3). Next, the algorithm activates all predictors and obtain all their predictions $f(x^t), \forall f \in \mathcal{F}$ given the input x^t (line 4). However, it selects only one of the prediction as the final prediction y^t, according to the selection as follows (line 5)

$$y^t = f^*(x^t) \quad \text{where} \quad f^* = \arg\max_f \bar{r}_f^t(C^t) \qquad (1)$$

In words, the selected predictor has the highest reward estimate for the context subspace C^t among all predictors. This is an intuitive selection based on the sample mean rewards. Next the counter M_C^t steps by 1 since we have one more sample in C. When the true traffic pattern $\hat{y}^t$ is revealed (line 6), the sample mean reward estimates for all predictors are then updated (line 7-8).

The second part of the algorithm, namely the adaptive context partitioning, is the key of our algorithm (line 9 - 11). At the end of each slot t, the algorithm decides whether to further partition the current subspace C^t, depending on whether we have seen sufficiently many incident arrivals in C^t. More specifically, if $M_C^t \geq A2^{lp}$, then C^t will be further partitioned (line 9), where l is the subspace level of C^t, $A > 0$ and $p > 0$ are two design parameters. When partitioning is needed, C^t is uniformly partitioned into 2^D smaller hypercubes (each hypercube is a level-$l+1$ subspace with side-length half of that of C^t). Then C^t is removed from the active context subspace set $\mathcal{P}$ and the new subspaces are added into $\mathcal{P}$ (line 11). In this way, $\mathcal{P}$ is still a partition whose subspaces are non-overlapping and cover the entire context space. Intuitively, the context space partitioning process can help refine the learning in smaller subspaces. In the next subsection, we will show that by carefully choosing the design parameters A and p, we can achieve sublinear learning regret in time, which implies the optimal time-average prediction performance can be achieved.

Algorithm 1 Context-aware Traffic Prediction (CATP)

1: Initialize $\mathcal{P}^0 = \{\Theta\}$, $\bar{r}_f(\Theta) = 0, \forall f \in \mathcal{F}$, $M_\Theta^0 = 0$.
2: **for** each traffic incident (time slot t) **do**
3: Determine $C^t \in \mathcal{P}^t$ such that $\theta^t \in C^t$.
4: Generate the predictions results for all predictors $f(x^t), \forall f$.
5: Select the final prediction $y^t = f^*(x^t)$ according to (1).
6: The true traffic pattern $\hat{y}^t$ is revealed.
7: Update the sample mean reward $\bar{r}_f(C^t), \forall f$.
8: $M_C^t = M_C^t + 1$.
9: **if** $M_C^t \geq A2^{pl}$ **then**
10: C^t is further partitioned.
11: **end if**
12: **end for**

B. Learning regret analysis

In this subsection, we analyze the regret of the proposed traffic prediction algorithm. The following technical assumption is needed

Assumption 1: For each $f \in \mathcal{F}$, there exists $L > 0$, $\alpha > 0$ such that for all $\theta, \theta' \in \Theta$, we have

$$\pi_f(\theta) - \pi_f(\theta')| \leq L\|\theta - \theta'\|^\alpha \qquad (2)$$

This states each predictor achieves similar expected rewards (accuracies) for similar contexts.

The following theorem establishes the regret bound when the context arrivals are uniformly distributed over the context space, which is the worst case arrival process for regret minimization.

Theorem 1: If the context arrival by time T is uniformly distributed over the context space, the regret is upper-bounded by $T^{\frac{D+2\alpha}{D+3\alpha}} 2^{(D+2\alpha)l}(2LD^{\alpha/2} + 2 + \log(T)) + T^{\frac{D}{D+3\alpha}} 2^{Dl} 2K \sum_{t=0}^{\infty} t^{-2})$.

We have shown that the regret upper bound is sublinear in time, implying that the average traffic prediction rewards (e.g. accuracy) achieves the optimal reward as time goes to infinity. Moreover, it also provides performance bounds for any finite time T rather than the asymptotic result. Ensuring a fast convergence rate is important for the algorithm to quickly adapt to the dynamically changing environment.

IV. Experiments

A. Experimental setup

1) Dataset: Our experiment utilizes a real-world traffic dataset, which includes both real-time and historically archived data since 2010. The dataset consists of two parts: (i) Traffic sensor data from loop-detectors. There are totally 9300 sensors located on the highways and arterial streets of Los Angeles County (covering 5400 miles cumulatively) collecting several main traffic parameters such as occupancy,

	Proposed	NB	SVM	WM	PAN
Upstream highway	0.72	0.65	0.55	0.59	0.66
Inter highway	0.70	0.61	0.61	0.55	0.57
Arterial way	0.92	0.84	0.57	0.76	0.86

Table I
PREDICTION ACCURACY COMPARISON FOR $\lambda = 0.5$

volume and speed at the rate of 1 reading per sensor per minute; (ii) Traffic incidents data. This dataset contains the traffic incident information in the same area as in the traffic sensor dataset. On average, 400 incidents occur per day and the dataset includes detailed information of each incident, including the severity and location information of the incident as well as the incident type etc.

2) Evaluation methods: We will evaluate the prediction accuracy in different spatial settings, including the upstream stretch of the highway, its adjacent arterial way and the intersected highway. For each spatial setting, we choose the traffic sensors located at around 100 locations, and retrieve the accidents from its nearby highway from 2012 to 2013 (around 300 for each location. The system aims to predict whether a nearby incident has an impact of the selected location. If the traffic speed drop exceeds λ (in percentage), then the location is labeled as being affected by the nearby incident. We will show the results for different values of λ. The context information that we use in the experiments include the incident start time (peak hour or off-peak hour, weekdays or weekends), incident type (vehicle collision, hazard, etc.) and the distance between the incident location and the selected location for traffic prediction. We use the simple binary reward function for evaluation. That is, the system obtains a reward of 1 if the prediction is correct and 0 otherwise.

3) Base predictors and baseline approaches: We construct base predictors in typical settings using Support Vector Machine (SVM) and Naive Bayes (NB). The baseline approaches that we compare against include the single base predictors using SVM and NB and ensemble learning techniques based weighted majority [11]–[13]. We also compare with our prior work [1], labeled by "PAN".

B. Results

We compare the prediction accuracy of the proposed algorithm against the various baseline solutions in Table I and II for three types of spatiotemporal prediction problems. The impact threshold is set to be $\lambda = 0.5, 0.3$, respectively. It can be seen that our proposed method performs consistently better than the baseline approaches. Although the accuracy of offline benchmarks improves with a larger training dataset, they are still worse than our proposed online algorithm which can adapt to changes in transportation environment online. Among the different types of spatiotemporal prediction, the accuracy for the locations on arterial streets is higher than those on upstream and intersected highway.

	Proposed	NB	SVM	WM	PAN
Upstream highway	0.90	0.83	0.46	0.62	0.85
Inter highway	0.79	0.70	0.59	0.55	0.71
Arterial way	0.86	0.79	0.54	0.54	0.81

Table II
PERFORMANCE COMPARISON FOR $\lambda = 0.3$

V. CONCLUSIONS

In this paper, we proposed a framework for online traffic prediction. The framework utilizes the real-time data to select the most effective predictor in different contexts, thereby self-adapting to the dynamically changing traffic conditions. We systematically proved both short-term and long-term performance guarantees for our algorithm and our experiments on real-world dataset verified the efficacy of the proposed approach. As a future work, we plan to also adapt the individual base predictors using real-time data in addition to selecting the most effective one to use.

REFERENCES

[1] B. Pan, U. Demiryurek, and C. Shahabi, "Utilizing real-world transportation data for accurate traffic prediction," ser. ICDM '12, 2012, pp. 595–604.

[2] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, "Forecasting spatiotemporal impact of traffic incidents on road networks," in *ICDM*, 2013, pp. 587–596.

[3] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *KDD*, 2011, pp. 1010–1018.

[4] S. Lee and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," in *TRR98*, 1998.

[5] X. Li, Z. Li, J. Han, and J.-G. Lee, "Temporal outlier detection in vehicle traffic data," in *ICDE*, 2009, pp. 1319–1322.

[6] M. Miller and C. Gupta, "Mining traffic incidents to forecast impact," ser. UrbComp '12. New York, NY, USA: ACM, 2012, pp. 33–40.

[7] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[8] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[9] M. Sewell, "Ensemble learning," *RN*, vol. 11, no. 02, 2008.

[10] P. Bühlmann and B. Yu, "Boosting with the l 2 loss: regression and classification," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[11] J. Gao, W. Fan, and J. Han, "On appropriate assumptions to mine data streams: Analysis and practice," in *ICDM*. IEEE, 2007, pp. 143–152.

[12] W. Fan, S. J. Stolfo, and J. Zhang, "The application of adaboost for distributed, scalable and on-line learning," in *KDD*. ACM, 1999, pp. 362–366.

[13] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and computation*, vol. 108, no. 2, pp. 212–261, 1994.

[14] C. Tekin, S. Zhang, and M. van der Schaar, "Distributed online learning in social recommender systems," IEEE JSTSP, vol. 8, no. 4, pp. 638-652, 2014.

[15] V. G. Vovk, "A game of prediction with expert advice," in *Proceedings of the eighth annual conference on Computational learning theory*. ACM, 1995, pp. 51–60.