RiskCast: Social Sensing based Traffic Risk Forecasting via Inductive Multi-View Learning

Yang Zhang, Hongxiao Wang, Daniel Zhang, Yiwen Lu, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
{yzhang42, hwang21, yzhang40, ylu9, dwang5}@nd.edu

***Abstract*—Road traffic accidents are a major challenge in urban transportation systems. An effective countermeasure to address this problem is to accurately forecast the traffic risks in a city before accidents actually happen. Current traffic accident prediction solutions largely rely on accurate data collected from infrastructure-based sensors, which is not always available due to various resource constraints or privacy and legal concerns. In this paper, we address this limitation by exploring social sensing, a new sensing paradigm that uses humans as sensors to report the states of the physical world. In particular, we consider two types of publicly available social sensing data sources: social media data (e.g., traffic posts on Twitter) and open city data (e.g., traffic data from the city web portal). In this paper, we develop the RiskCast, an inductive multi-view learning approach to accurately forecast the traffic risk by exploiting the social sensing data under a principled co-regularization framework. The evaluation results on a real world dataset from New York City show that RiskCast significantly outperforms the state-of-the-art baselines in forecasting the traffic risks in a city.**

I. Introduction

Social Sensing has emerged as a new sensing paradigm where humans (or devices on their behalf) collectively report measurements about the physical world [1]. Examples of social sensing include real-time traffic condition monitoring using mobile crowdsensing [2] and obtaining real-time situation awareness for disaster response using online social media [3]. Intelligent transportation system (ITS) is a critical application domain where sensing, communication, and control techniques are used to improve safety and efficiency of the transportation systems [4]. Current ITS applications primarily rely on various types of infrastructure-based sensors (e.g., speed sensors, CCTV cameras, loop detectors) to collect real-time traffic information [4]. However, such infrastructure-based sensors are not always available due to the resource constraints, privacy concerns, and legislation [5]. In contrast, social sensing provides an *infrastructure-free* solution [6] that is more pervasive and scalable than the traditional solutions for ITS applications by exploring the open and publicly available data from human sensors (e.g., social media data and traffic reports published by a city) [7]. In this paper, we focus on a social sensing based *traffic risk forecasting* problem, where the goal is to accurately forecast the traffic risks (i.e., the probability of the traffic accidents) at a fine-grained spatial granularity (e.g., a road intersection in city).

Recent progress has been made to address the traffic risk prediction problem in intelligent transportation systems, geographical information systems, and data mining communities [8]–[10]. However, these solutions cannot be directly adapted to solve our problem because they largely rely on accurate traffic sensor data collected from *infrastructure monitoring devices* (e.g., traffic cameras, radar detectors, GPS sensors) in the traffic systems. However, such monitoring devices and data may not always be available [5]. For example, less than 3% US cities install road traffic cameras and traffic monitoring devices are prohibited by 10 states in US [1]. In New York City (NYC), more than 85% of the fatal and serious injury crashes happen at locations where the traffic monitoring devices are not available [2].

To address the above limitation, we develop a social sensing based traffic risk forecasting scheme that does not depend on the infrastructure-based sensors and monitoring devices. In particular, we consider two types of widely available social sensing data sources: *social media data* and *open city data* (as shown in Figure 1). For social media data, we refer to the traffic related claims people have posted on online social media (e.g., real-time tweets collected from the Twitter API). For the open city data, we refer to traffic accident reports published by the city governments (e.g., motor vehicle collision reports periodically updated by the New York police department). Both types of data are generated by *human sensors* (e.g., Twitter users and police officers) but have different yet complementary characteristics [11]. In particular, the social media data is often timely but sparse in terms of accident coverage [12]. In contrast, the open city data has a good accident coverage but is less timely due to delays in the data collection, processing, and review process [13]. In this

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08...$15.00
http://doi.org/10.1145/3341161.3342912

[1] https://www.iihs.org/iihs/topics/laws/automated_enforcement/enforcementtable?topicName=speed

[2] https://www1.nyc.gov/office-of-the-mayor/news/403-17/visionzeromayor-deblasiofamiliesseniorsofficialscallexpansionlifesaving#0

paper, we develop RiskCast, a social sensing based multi-view learning scheme that explores the benefits from both types of social sensing data for the urban traffic risk forecast. To the best of our knowledge, the RiskCast is the first social sensing based solution to address the traffic risk forecasting problem in intelligent transportation systems using a multi-view learning approach. We evaluate the RiskCast scheme on a real-world traffic dataset from New York City. The results show that our scheme significantly outperforms the state-of-the-art baselines in various application scenarios.

(a) Accident Reported on Social Media Data

Time & Date	Borough	Latitude	Longitude	Vehicles
9:50, 07/18/2017	MANHATTAN	40.761147°	-73.97952°	PASSENGER VEHICLE & TAXI

(b) Accident Recorded in Open City Data

Figure 1. Example of Social Sensing Data for Traffic Risk Forecasting

II. Related Work on Traffic Risk Prediction

Previous efforts have made good progress to address the traffic risk prediction related problems in intelligent transportation systems, geographical information systems, and data mining communities [8]–[10]. For example, Lin *et al.* developed a frequent pattern tree based approach to predict the traffic risk using traffic data collected from interstate highways [8]. Sun *et al.* proposed a dynamic Bayesian network based model to predict car crashes using the traffic speed data collected from freeway traffic sensors [9]. Shi *et al.* developed a random forest and Bayesian inference based framework for real-time traffic safety prediction using the data collected from traffic loop detectors deployed on urban expressways [10]. These approaches cannot be directly adapted to solve our traffic risk forecasting problem because they rely on a large amount of accurate traffic sensor data collected from infrastructure monitoring devices, which are not always available due to resource and legal constraints [5]. In contrast, we develop a novel multi-view co-regularization learning scheme to address the traffic risk forecasting problem by taking advantage of social sensing, which collects traffic information from human sensors.

III. Problem Definition

In this section, we formulate the traffic risk forecasting problem in intelligent transportation systems. We first define the terms that will be used in the problem statement.

Definition 1: **Sensing Cell** (SC): We divide the sensing area (e.g., New York City) into disjoint sensing cells where each cell represents a subarea of interest. In particular, we define C to be the number of cells in the sensing area and SC_c to be the c^{th} sensing cell in the sensing area $(c = 1, 2, \cdots, C)$.

Definition 2: **Social Media Data** (SD): We define the Social Media Data (SD) to be the self-reports about traffic accidents from social media users (e.g., tweets shown in Figure 1(a)).

Definition 3: **Open City Data** (OD): We define the open city data (OD) to be the publicly accessible traffic accident reports published by cities (e.g., accidents reports published by NYC Police department shown in Figure 1(b)).

Definition 4: **Forecasting Window**: A Forecasting window is a period of time in the upcoming future where we predict the traffic risk in a city based on the social sensing data collected before the forecasting window. In particular, we define T to be the total number of forecasting windows in the traffic risk forecasting application and t to be the t^{th} forecasting window.

Definition 5: **Traffic Accident Rate** (Y): In this paper, we use the Traffic Accident Rate (Y) to indicate the *traffic risk level* of a location in a city at a given time. In particular, we define Y_t^c and $\widehat{Y}_t^c$ to be *real* and *estimated* traffic accident rate of cell SC_c at forecasting window t, respectively.

Using the above definitions, we can formally define our traffic risk forecasting problem. The goal is to correctly forecast the traffic accident rate of each sensing cell at each forecasting window based on the collected social sensing data SD and OD. Formally, our problem is defined as:

$$\arg\min_{\widehat{Y}_t^c}\left(\frac{1}{C}\cdot\sum_{c=1}^{C}\frac{1}{T}\cdot\sum_{t=1}^{T}abs(Y_t^c-\widehat{Y}_t^c)\mid SD, OD, SC, T\right) \tag{1}$$

where $abs()$ is function to generate the absolute value of a given number.

IV. Solution

In this section, we present RiskCast to address the traffic risk forecasting problem formulated in the previous section.

A. Sensing Feature Extraction & View Construction (SFEVC)

In this subsection, we describe the SFEVC component to extract the traffic accident features from unstructured social media data and open city data.

In particular, for social media data SD, we extract the location l_s of each social media post s in SD by analyzing the content of social sensing data using location-specific regular expressions [14]. Then, the extraction of the accident time t_s from social media post s can be achieved by checking the timestamp of the data sample [15] (e.g., the "created_at" field of a tweet). For open city data OD, we can query the open city database [3] to obtain the accident location l_o and time t_o for each traffic accident report o in OD. Finally, we convert the extracted time t_s and location l_s from social media posts as the *social media data view* and the extracted time t_o and location l_o from the open city traffic reports as the *open city data view* as follows:

$$\begin{cases} X_{SD}^c = \{(l_s, t_s) | l_s \in SC_c, \forall s \in SD\}, \\ X_{OD}^c = \{(l_o, t_o) | l_o \in SC_c, \forall o \in OD\}, \end{cases} \quad \forall SC_c \in SC \tag{2}$$

[3]https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

where X_{SD}^c and X_{OD}^c are the *social media data view* and *open city data view* at sensing cell SC_c, respectively, which will serve as the inputs to the IMVCL component discussed in the next subsection.

B. Inductive Multi-View Co-Regularized Learning (IMVCL)

In this subsection, we describe the Inductive Multi-view Co-Regularized Learning (IMVCL) component that forecasts the traffic risk of each sensing cell by exploring the two social sensing views generated by the SFEVC component under a principled inductive multi-view co-regularized learning framework.

First, we formulate the traffic risk forecasting problem using data from the sensing views generated by the SFEVC component as a combined linear mapping problem as follows:

$$\widehat{Y}^c = \sum_{v=1}^{V} \sigma_v f^v(X_v^c) = \sum_{v=1}^{V} \sigma_v X_v^c W_v^c + B^c \tag{3}$$

where $\widehat{Y}^c$ is the estimated traffic accident rate in sensing cell SC_c. V is the number of sensing views in IMVCL component (i.e., $V = |\{X_{SD}, X_{OD}\}| = 2$ in this paper). X_v^c is the set of social sensing data from the v^{th} sensing view at sensing cell SC_c, where $X_v^c \in \{X_{SD}^c, X_{OD}^c\}$ (defined in Equation 2). σ_v is the weight of the v^{th} sensing view, which is usually set to be a small value (e.g., $\frac{1}{V}$) for all views if no prior knowledge is given. f^v is the prediction function for the v^{th} sensing view that takes the sensing data in sensing view X_v^c and outputs the estimated traffic accident rate $\widehat{Y}^c$ for sensing cell SC_c, and f^v is a linear mapping function (i.e., $f^v(X_v^c) = X_v^c W_v^c$). W_v^c is the mapping matrix and B^c is the coefficient matrix.

The key to solve the above problem is to obtain the optimal values of W_v^c and B^c that minimize the difference between the predicted $\widehat{Y}^c$ and the real value Y^c for all sensing cells. To learn the optimal values of W_v^c and B^c, we develop a Co-Regularized learning based framework. In our framework, we define the objective function of the multi-view learning scheme as follows:

$$\begin{aligned}\arg\min_{W_v^c, B^c} \sum_{c=1}^{C} ||Y^c - \sum_{v=1}^{V} \sigma_v X_v^c W_v^c - B^c||_2^2 + \omega \sum_{c=1}^{C} \sum_{v=1}^{V} ||W_v^c||_2^2 \\ + \theta \sum_{c=1}^{C} \sum_{\widehat{v},\overline{v}=1,\widehat{v}\neq\overline{v}}^{V} ||X_{\widehat{v}}^c W_{\widehat{v}}^c - X_{\overline{v}}^c W_{\overline{v}}^c||_2^2 \\ + \epsilon \sum_{v=1}^{V} \sum_{n=1}^{N} \sum_{c\in S_n} ||W_v^c - \frac{1}{|S_n|} \sum_{\widehat{c}\in S_n} W_v^{\widehat{c}}||_2^2\end{aligned} \tag{4}$$

where Y^c is the *true* traffic accident rate. $||W_v^c||_2^2$ is the l2-regularizer of the mapping matrix W_v^c to control the sparsity of each learned mapping matrix W_v^c to avoid the over-fitting of our forecasting model. $\sum_{c=1}^{C} \sum_{\widehat{v},\overline{v}=1,\widehat{v}\neq\overline{v}}^{V} ||X_{\widehat{v}}^c W_{\widehat{v}}^c - X_{\overline{v}}^c W_{\overline{v}}^c||_2^2$ is the co-regularizer to enforce the agreement on the prediction results made by different sensing views at the same sensing cell. $\sum_{v=1}^{V} \sum_{n=1}^{N} \sum_{c\in S_n} ||W_v^c - \frac{1}{|S_n|} \sum_{\widehat{c}\in S_n} W_v^{\widehat{c}}||_2^2$ is the clustered mean-constrained regularization term to encode the spatial correlations into our objective function, where N is the number of sensing cell clusters in the sensing area and S_n is the set of sensing cells in the cluster n.

The above objective function can be solved using gradient descent techniques [16] to obtain the solution of the mapping matrix W and coefficient matrix B. After we obtain the optimal solutions of W and B, we can apply them to forecast the traffic accident rate for each cell using the prediction function in Equation 3.

V. Evaluation On Real World Data

A. Dataset

In our evaluation, we use *Get Old Tweets* [4] to collect a dataset from Twitter about traffic accidents over the time period from Jan. 1st, 2016 to Jun. 30th, 2018 in New York City as our *social media data*. In addtion, we use a public traffic accident report dataset provided by the New York City Police Department (NYPD) [5] at the same time-frame as the *ground-truth data* to evaluate all compared schemes. We also generate the *open city data* from the NYPD traffic accident report dataset. Different from the ground truth data, we postpone the available time of the traffic accident reports for a month.

B. Baseline and Metrics

We choose several representative traffic risk forecasting baselines that are applicable to the social sensing data paradigm we studied in this paper. In particular, each baseline consists two parts: i) the *data sources* it uses for traffic risk forecasting; ii) the *forecasting algorithm* it adopts.

Data Sources

- **Social-based (S)**: Social-based schemes predicts the traffic risk of a sensing cell based on the social media data.
- **Open-based (O)**: Open-based schemes forecasts the traffic risk of a sensing cell based on the open city data.

Forecasting Algorithm

- **Linear Regression (LR)** trains a linear regression model to minimize the difference between the *true* and *estimated* traffic accident rate [17].
- **Ridge Regression (Ridge)** tries to learn the weight of the forecasting model by adding a Ridge regularizer to enforce the robustness of the learned model [18].
- **MultipleLayer Perception (MLP)** is a well-known deep neural network framework that models the non-linearity in traffic accident data to predict traffic accident rate [19].

The combinations of the data sources and forecasting algorithms discussed above comprise the baselines.

In our evaluation, we define the *Mean Absolute Error (MAE)* to evaluate the performance of all compared scheme: $MAE = \frac{1}{C} \cdot \sum_{l=1}^{C} \frac{1}{T} \cdot \sum_{t=1}^{T} abs(Y_t^c - \widehat{Y}_t^c)$, where C is the number of the sensing cells, and T is number of the forecasting windows. Y_t^c and $\widehat{Y}_t^c$ are the *true* and *estimated* traffic accident rate for cell l at forecasting window t.

[4] https://github.com/Jefferson-Henrique/GetOldTweets-python

[5] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

C. Evaluation Results

In the experiments, we evaluate the performance of all schemes by selecting different set of sensing cells. We focus on the sensing cells with more than 100 accidents over the studied time period, which translates to an average of more than 1 accident per week. In particular, we select three subsets of sensing cells by gradually increasing the accident rate of the sensing cells from above 100 to above 200 over the study's time period (we refer to them as $A > 100$, $A > 150$, and $A > 200$). We set the forecasting window to be one week by considering the frequency of the accidents in the studied area. The results are presented in Table I. We observe that the RiskCast scheme outperforms all of the baselines at locations with different traffic risks. In terms of the mean absolute error (MAE), the performance gains achieved by RiskCast compared to the best-performing baseline with $A > 100$, $A > 150$, and $A > 200$ are 10.9%, 11.9%, and 4.5% respectively. This is because the RiskCast accurately forecasts the traffic risk by judiciously exploring both social media data and open city data through a principled multi-view co-regularized learning framework.

Table I
PERFORMANCE COMPARISONS (MAE) ON LOCATIONS WITH DIFFERENT ACCIDENT RATES

		Different Accident Rates		
Category	Algorithm	$A > 100$	$A > 150$	$A > 200$
Social-based	S-LR	1.176	1.388	1.876
	S-Ridge	1.132	1.327	1.748
	S-MLP	1.168	1.302	1.640
Open-based	O-LR	1.325	1.479	1.733
	O-Ridge	1.303	1.462	1.725
	O-MLP	1.423	1.601	2.007
Our Alg	**RiskCast**	**1.020**	**1.163**	**1.569**

VI. CONCLUSION

In this paper, we develop the RiskCast scheme to solve the traffic risk forecasting problem in intelligent transportation systems. The RiskCast scheme addresses the limitation of current solutions that largely depend on accurate sensing measurements from infrastructure based sensors by exploring two widely available yet complementary social sensing data sources: social media data and open city data. In particular, RiskCast makes accurate traffic risk forecasting in a city by exploiting the social sensing data under a principled multi-view co-regularized learning framework. The evaluation results on the real-world dataset from New York City demonstrate that the RiskCast scheme achieves significant performance gains compared to the state-of-the-art baselines and provides opportunities to improve the safety of the urban traffic systems.

ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation under Grant No. CNS-1831669, CBET-1637251, Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.

[2] S. Ilarri, O. Wolfson, and T. Delot, "Collaborative sensing for urban transportation." *IEEE Data Eng. Bull.*, vol. 37, no. 4, pp. 3–14, 2014.

[3] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on*. IEEE, 2012, pp. 233–244.

[4] Y. Lin, P. Wang, and M. Ma, "Intelligent transportation system (its): Concept, challenge and opportunity," in *Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2017 IEEE 3rd International Conference on*. IEEE, 2017, pp. 167–172.

[5] A. Najjar, S. Kaneko, and Y. Miyanaga, "Combining satellite imagery and open data to map road safety." in *AAAI*, 2017, pp. 4524–4530.

[6] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 966–976.

[7] Y. Zhang, H. Wang, D. Zhang, and D. Wang, "Deeprisk: A deep transfer learning approach to migratable traffic risk estimation in intelligent transportation using social sensing," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019.

[8] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 444–459, 2015.

[9] J. Sun and J. Sun, "A dynamic bayesian network model for real-time crash prediction using traffic speed conditions data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 176–186, 2015.

[10] Q. Shi and M. Abdel-Aty, "Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 380–394, 2015.

[11] D. Wang, T. Abdelzaher, and L. Kaplan, *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[12] D. Y. Zhang, R. Han, D. Wang, and C. Huang, "On robust truth discovery in sparse social media sensing," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1076–1081.

[13] C. J. Bennett and C. D. Raab, *The governance of privacy: Policy instruments in global perspective*. Routledge, 2017.

[14] Y. Zhang, Y. Lu, D. Zhang, L. Shang, and D. Wang, "Risksens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1544–1553.

[15] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *2013 IEEE 33rd International Conference on Distributed Computing Systems*. IEEE, 2013, pp. 530–539.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[18] A. Alaoui and M. W. Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," in *Advances in Neural Information Processing Systems*, 2015, pp. 775–783.

[19] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM conference on recommender systems*. ACM, 2016, pp. 191–198.