Prediction of Road Traffic using Naive Bayes Algorithm

E. Baby Anitha[1]
[1]Professor,
Department of Computer Science and Engineering.
K.S.R. College of Engineering,
Tiruchengode, India.

R. Aravinth[2], S. Deepak[3],
R. Jotheeswari[4], G. Karthikeyan[5]
[2,3,4,5] UG Students
Department of Computer Science and Engineering.
K.S.R. College of Engineering,
Tiruchengode, India.

Abstract **-Road traffic speed prediction is a challenging problem in intelligent transportation system (ITS) and has gained increasing attentions. Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In order to give safe driving suggestions, careful analysis of roadway traffic data is critical to find out variables that are closely related to fatal accidents. In response to these challenges, in existing system using the unified probabilistic framework, called Topic-Enhanced Gaussian Process Aggregation Model (TEGPAM), consisting of three components, i.e., location disaggregation model, traffic topic model, and traffic speed Gaussian Process model, which integrate new-type data with traditional data. Roadway traffic safety is a major concern for transportation governing agencies as well as ordinary citizens. In our project we can find the traffic issue and solutions. We upload the traffic dataset and preprocessing. That dataset store to sql after preprocessing it is remove the null value in dataset we clustering using k-means algorithm clustering is the group of information we using find the solution using Naive Bayes algorithm we can find the exact solution which time date traffic occur and accident time so we improve the our safety training. Certain safety driving suggestions were made based on statistics, association rules, classification model, and clusters obtained. Clustering is the get the fetch data for apply the algorithm. Feature extraction is we get the particular data from dataset. After we get the result analysis the dataset it is the output for our project. The percentage of fatal accidents happened at different speed limit in comparison of people involved and fatal.**

Keywords: Traffic prediction, gaussian process, topic modelling, multi-source data, naive bayes .

I. 1. INTRODUCTION

ROAD traffic monitoring is of great importance for urban transportation system. Traffic control agencies and drivers could benefit from timely and accurate road traffic prediction and make prompt, or even advance decisions possible for detecting and avoiding road congestions. Existing methods mainly focus on raw speed sensing data collected from cameras or road sensors, and suffer severe data sparsity issue because the installation and maintenance of sensors are very expensive. There are a lot of vehicles driving on the roadway every day, and traffic accidents could happen at any time anywhere. Some accident involves fatality, means people die in that accident. As human being, we all want to avoid accident and stay safe. To find out how to drive safer, data mining technique could be applied on the traffic accident dataset to find out some valuable information, thus give driving suggestion. Data mining uses many different techniques and algorithms to discover the relationship in large amount of data. It is considered one of the most important tool in information technology in the previous decades. Association rule mining algorithm is a popular methodology to identify the significant relations between the data stored in large database and also plays a very important role in frequent item set mining. A classical association rule mining method is the Apriori algorithm who main task is to find frequent item sets, which is the method we use to analyse the roadway traffic data. Classification in data mining methodology aims at constructing a model (classifier) from a training data set that can be used to classify records of unknown class labels. The Naive Bayes technique is one of the very basic probability-based methods for classification that is based on the Bayes' hypothesis with the presumption of independence between each pair of variables. We used the FARS dataset for our study. The Fatal Accidents Dataset contains all fatal accidents on public roads in 2007 reported to the National Highway Transportation Safety Administration. It could possibly reduce the fatality rate. Using a road safety database enables to reduce the fatality by implementing road safety programs at local and national levels. That database scheme which describes the road accident via roadway condition, person involved and other data would be useful for case evaluation, collecting additional evidences, settlement decision and subrogation. Using web data a self-organizing map for pattern analysis was generated. It could classify information and provide warning as an audio or video. It was also identified that accident rates highest in intersections then other portion of road.

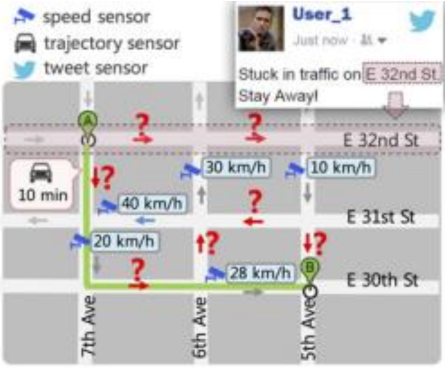

Fig. 1. Problem setting. Our goal is to predict the traffic speed of specific road links, as shown with the red question marks, given: 1) Some speed observations collected by speed sensors, as shown in blue; 2) trajectory and travel time of OD pairs. Note that speeds of passed road links are either observed or to be predicted; 3) tweets describing traffic conditions. Note that the location mentioned by a tweet may be a street covering multiple road links.

CHALLENGES

When integrating traditional traffic speed data (e.g., sensing data) with new-type data (e.g., Twitter data and trajectory data) to predict road traffic speed, technical challenges arise due to the characteristic of each data source: Location uncertainty of low-resolution data. Tweet data and trajectory data are called low-resolution data because we cannot directly locate them into specific road links. Most tweets have no location tags, so geographic location language is the main clue, which however is vague. For example, expression like "Stuck in traffic on E 32nd St. Stay away!" covers the whole street without precise road locations. Meanwhile, travel time of a trajectory is an aggregate measure based on the speed of multiple links, which may vary widely. Thus a strategy is required to disaggregate the data to specific road links. Language ambiguity of traffic description in tweets. The expressions depicting traffic conditions are diverse, and may denote different speed values. Fig. 2 shows an example of word frequency distribution over the degree of congestion when people use congestion-related words. Meanwhile some words not directly related to traffic may also have strong implication to link speed, such as words complaining bad weather. Thus a linguistic model is required to capture the patterns between discrete descriptive words and continuous speed values. Heterogeneity of multi-source data. Cross-domain data sources have diverse properties and contain latent relations with road traffic speed. For example, tweets possess latent topics which cluster based on speed levels, and negative correlation existed between trajectory travel time and traffic speed of involving links. Therefore a unified framework is required to model these properties and aggregate the latent relations between heterogeneous data to predict speed synthetically.

2. RELATED WORK

Traffic prediction problem can be broadly classified into short-term and long-term prediction, considering three main basic traffic measurements: traffic flow, an equivalent flow rate in vehicles; speed, mean of the observed vehicle speeds; lane occupancy, the percentage of time that the sensor is detecting vehicle presence. This paper focuses on the short-term traffic speed prediction combining multi-source heterogeneous data, which, as far as we know, has not been well explored before. This part gives a summary on short-term traffic speed prediction and the exploration on fusing multiple information sources. Short-Term Traffic Speed Prediction. The presented methods can be classified into two categories:

1) Parametric methods, assume that traffic speed follows a probability distribution based on a fixed set of parameters. Time series analysis technique is applied in traffic speed prediction based on the periodicity of traffic speed during a day or a week. Auto-Regressive Moving Average (ARMA) model and Multivariate Spatial-Temporal Auto-Regressive (MSTAR) model are adopted to include dependency among observations from neighboring locations. Auto-Regressive Integrated Moving Average (ARIMA) time series methods are reviewed for modeling and forecasting vehicular traffic flow. ARIMA and winters exponential smoothing techniques are used to forecast urban freeway flow. A single Space-Time Auto-Regressive Integrated Moving Average (STARIMA) model is proposed to describe the spatiotemporal evolution of traffic flow in an urban network, which is essentially a constrained Vector Autoregressive Moving Average (VARIMA) model with constraints that reflect the topology of a spatial network and result in a drastic reduction in the number of parameters. A Generalized Space-Time ARIMA (GSTARIMA) method is proposed to extend ARIMA in spatial and temporal dimension and is more flexible because parameters are designed to vary per special location. Kalman filter-based approaches are used and show advantages for on-line estimation of traffic flows. Markov logic network is used to simultaneously predict he congestion state. A structured time series model is proposed in multivariate form for short-term traffic prediction.

2) Non-parametric methods, make no distribution assumptions and the number of parameters scales with the number of training data. K-nearest neighbor (KNN) nonparametric regression methods find the k-nearest neighbors using Euclidean distance and calculate the weight. Neutral Networks (NNs) are trained to approximate any nonlinear function given adequate traffic sensing data and a proper network architecture. NNs have many derivatives for short-term prediction, such as back propagation neutral network with genetic algorithms and wavelet networks. Travel speed of each road segment is computed using the GPS trajectories by a context-aware matrix factorization approach. To adaptively route a fleet of cooperative vehicles under the uncertain and dynamic road congestion conditions, a GP probabilistic model is proposed to capture the spatial and temporal relationships of travel speeds over road segments and temporal contexts, especially with estimating the mean and covariance of the GP prior from the historical data. Geostatistical interpolation techniques named Kriging are proposed to capture spatial and temporal evolutions of traffic flows. Traffic Modeling with Multi-Source Heterogeneous Data. Traffic modeling problems gain further insights through fusing heterogeneous data from multiple sources, e.g., road sensors, social media and floating cars, to handle external factors such as traffic accidents, mobile sensors, and weather. Reviews the literature on the impact of weather on traffic demand, traffic safety, and traffic flow relationships. A trajectory-based community discovery method is proposed, where the trajectory similarity is modeled by several types of kernels for different information markers (e.g., semantic properties of the locations and the movement velocity). The prediction problem of rents/returns bike number is tackled using multiple features, e.g., time and meteorology, as measures of similarity functions in multi-similarity based inference model. While and introduce different information sources as new features for computing the similarity, our work assumes the latent relations between these information's, and constructs a Bayesian generative process. As crowdsourcing data from a crowd of online social platform become more available, researchers begin utilizing social content to estimate traffic conditions. Twitter data are matched to detect traffic incidents in. In traffic anomaly detection uses crowd sensing with two forms of data, human

mobility and social media, and the detected anomalies are described by mining representative terms from the social media that people posted when the anomaly happened. Few methods incorporate social media text data (e.g., Twitter data) to improve traffic speed prediction. Extends spatiotemporal GP in to three dimensional topic-aware GP, where topics on road links are probabilistic modeled based on the user, space and time of tweets. Do not tackle the location uncertainty problem of tweets, because the inference of traffic status based on words of tweets only focuses on the average regional traffic flow, which is insufficient for predicting road speed.

3. RECENT METHOD

Naive Bayes is a classification algorithm based on Bayesian theorem, with the naive assumption that each pair of input variables is independent. Although this assumption is oversimplified, this algorithm has effectively been used in many complicated real-world problems especially document classification and spam filtering. Moreover, NB has proven to be very fast in learning and classifying data. In this article, the NB algorithm is applied to the accidents data set to gain insight into its performance in predicting the severity of accidents.Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

CONTRIBUTIONS

In spite of the good potential of these new-type data, to the best of our knowledge, the problem of road-level traffic speed prediction using multiple data sources has not been well explored before, especially with the aforementioned challenges. In this paper, we propose a unified statistical framework, entitled Topic Enhanced Gaussian Process Aggregation Model (TEGPAM) fusing multi-source data, which includes traditional speed sensing data, and new-type "sensing" data from social media and map services.The framework combines the location disaggregation model to decompose vague locations into specific links, the traffic topic model to handle the language ambiguity in tweets and the Gaussian Process model to capture the spatial correlation in traffic sensing data. Specifically, this paper makes the following contributions: Integration of data from multiple cross-domain sources.We implement the idea of improving traffic speed prediction by integrating speed sensing data with new-type traffic-related data, such as tweets and trajectories. Formulation of the unified TEGPAM framework. We propose a unified probabilistic framework TEGPAM that combines the disaggregation model, topic model with Gaussian Process model and is learned by variation methods and a stochastic EM algorithm. Extensive experiments to validate the performance of the proposed method. We validate our approach using real-world data collected from two large American cities. The extensive experiments show the effective- ness of TEGPAM, as well as the model efficiency and reliability.

4. DATASET AND BAYES ALGORITHM

We obtain three data sources for road traffic speed prediction: 1) Traffic speed data. INRIX database [50] provides traffic speeds for each road link at a 5-minute rate, from June 1, 2013 to March 31, 2014, across two cities: Washington D.C. and Philadelphia. 2) Trajectory data. Trajectories are generated from INRIX database at a 5-minute rate. Given a random OD pair, we synthesize a trajectory by computing the shortest path between them (i.e., using Johnson's algorithm [51]). With the length and speed information of links from INRIX, the travel time of this trajectory is obtained by adding the time of each link up and corrupting it with a Gaussian noise. 3) Twitter data. Tweets in the same time period and cities are collected via the Twitter REST search API. Traffic related tweets are preliminarily extracted by matching at least one term of a predefined vocabulary developed by domain experts, which included terms like "traffic", "accident", "stuck", "crash", etc, then further classified and filtered using an SVM classifier that was trained based on manually labeled 10,000 tweets (50 percent positive and 50 percent negative tweets).With road records containing the geo-coordinates, names and aliases, we geocode tweets to road links by matching their geo-tag and text content to the front end of those links, which corresponds to the driving out direction and is denoted as Head. Different driving directions are denoted as different road links. After geocoding, there are 5 major roads with 35 road links mentioned in the Philadelphia twitter data, and 8 major roads with 44 links in Washington D.C. respectively.

BERNOULLI NAIVE BAYES

In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks,[10] where binary term occurrence features are used rather than term frequencies. If {\displaystyle x_{i}} x_{i} is a boolean expressing the occurrence or absence of the i'th term from the vocabulary, then the likelihood of a document given a class {\displaystyle C_{k}} C_{k} is given by,{\displaystyle p(\mathbf {x} \mid C_{k})=\prod _{i=1}^{n}p_{ki}^{x_{i}}(1-p_{ki})^{(1-x_{i})}} {\displaystyle p(\mathbf {x} \mid C_{k})=\prod _{i=1}^{n}p_{ki}^{x_{i}}(1-p_{ki})^{(1-x_{i})}}

where {\displaystyle p_{ki}} p_{{ki}} is the probability of class {\displaystyle C_{k}} C_{k} generating the term {\displaystyle x_{i}} x_{i}. This event model is especially popular for classifying short texts. It has the benefit of explicitly modelling the absence of terms. Note that a naive Bayes classifier with a Bernoulli event model is not the same as a multinomial NB classifier with frequency counts truncated to one.

5. IMPLEMENTATION

5.1. SYSTEM MODEL

In this module, we develop a System with a disaggregation model for location uncertainty in tweet and trajectory data, a traffic topic model for tweet language ambiguity and a GP model for capturing the spatial correlation of speed sensing data. In this module, first we develop the system construction entitles required for the proposed model. The system provides the new user for the registration and then login authorization. The authorized users can post their tweets. The users are provided with the option of the posting comments too. The module is designed with the features of Online Social Network modeled base, with the functionalities which are correlated to the proposed model.

5.2. TRAFFIC RELATED TWEETS

Our goal is to predict traffic speed of some links at a certain time stamp using the past and current observations from multiple data sources, including traffic sensing data, Tweets and trajectories. Tweets in the same time period and cities are collected via the Twitter REST search API. Traffic related tweets are preliminarily extracted by matching at least one term of a predefined vocabulary developed by domain experts, which included terms like “traffic”, “accident”, “stuck”, “crash”,“delay”, “disaster”, “tragedy”, “problem”, “misfortune”, “difficulty”, “mishap” etc., then further classified and filtered.

5.3. DISAGGREGATION OF MODEL

To handle the challenge of location uncertainty of new-type data, this section presents a disaggregation strategy to map the low-resolution data, which are tweets and trajectories, into specific road links. Since only 1 percent of tweets have geo-coordinates, most location information’s are extracted from tweet text by mapping road names or alias.The time traveling through a trajectory is a sum of time cost on each link. Recall that vt;s is the traffic speed at time t and ls 2 L is the road length of links. By denoting vt ¼ fvt;s; s 2 Sg as speeds over all links at time t, and indicative function of links passed by trajectory p, we define a function to disaggregate travel time of a trajectory into the speed of specific links.

5.4. TRAFFIC TOPIC MODEL

To address the challenge of language ambiguity and capture the traffic description in tweets, a traffic topic model is proposed. With road records containing the geo-coordinates, names and aliases, we geocode tweets to road links by matching their geo-tag and text content to the front end of those links, which corresponds to the driving out direction and is denoted as Head. Different driving directions are denoted as different road links.

6. CONCLUSIONS

This paper proposes a novel probabilistic framework to predict road traffic speed with multiple cross-domain data. Existing works are mainly based on speed sensing data, which suffers data sparsity and low coverage. In our work, we handle the challenges arising from the multi-source data of road traffic, including location uncertainty, language ambiguity and data heterogeneity, applying Bayes' theorem with strong independence assumptions between the features. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. By using this algorithm we can exactly predict the accident details from the data set. By using this algorithm we can exactly predict the accident details from the data set.For Future work, we plan to implement kernel-based and distributive GP, so the traffic prediction framework can be applied into a real time large traffic network.

REFERENCES:

[1] Lu Lin, Jianxin Li , Feng Chen, Jieping Ye, Senior Member, “Road Traffic Speed Prediction: AProbabilistic Model Fusing Multi-Source Data.”vol. 30, no. 7, 2018.

[2] X. Yu and P. D. Prevedouros, “Performance and challenges in utilizingnon-intrusive sensors for traffic data collection,” AdvancesRemote Sens., vol. 2, pp. 45–50, 2013

[3] W. Min and L. Wynter, “Real-time road traffic prediction withspatio-temporal correlations,” Transp. Res., vol. 19, pp. 606–616,2011.

[4] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. W. C. V. Lint, “A genetic algorithm-based method for improving quality of travel time prediction intervals,” Transp. Res., vol. 19, pp. 1364–1376, 2011.

[5] M. Xinyu and H. Jianming, “Urban traffic network modeling andshort-term traffic flow forecasting based on GSTARIMA model,”in Proc. Int. IEEE Conf. Intell. Transp. Syst., 2010,pp. 19–22.

[6] J. Guo and B. M. Williams, “Real-time short-term traffic speed level forecasting and uncertainty quantification using layered kalman filters,” Transp. Res. Rec., vol. 2175, pp. 28–37, 2010.

[7] B. Ghosh, B. Basu, and M. O’Mahony, “Multivariate short-termtraffic flow forecasting using time-series analysis,” IEEE Trans.Intell. Transp. Syst., vol. 10, no. 2, pp. 246–254, Jun. 2009.

[8] Y. Kamarianakis and P. Prastacos, “Space-time modeling of trafficflow,” Computers & Geosciences, vol. 31, no. 2, pp. 119–133, 2005.

[9] M. Kamarianakis and P. Prastacos, “Forecasting traffic flow conditions in an Urban network: Comparison of multivariate and univariate approaches,” Transp. Res. Rec., vol. 1857, pp. 74–84, 2004.