Real-time Traffic Accident Severity Prediction using Data Mining Technologies

Xiao-Ling Xia
Department of Computer Science and Engineering
Donghua University
Shanghai, China
sherlysha@dhu.edu.cn

Bing Nan
Department of Computer Science and Engineering
Donghua University
Shanghai, China
2151541@mail.dhu.edu.cn

Cui Xu
Department of Computer Science and Engineering
Donghua University
Shanghai, China

Abstract. With the urban transport development in the recent years, frequent traffic accidents and other problems need to be improved. Understanding the causes of traffic accidents and making an early alarm model for the driver will be crucial to solve traffic accident problems in some way. In this paper, we focus on the factors that can be collected in real-time and process the factors using data mining technologies. Finally, we evaluate the performance of different classifiers. The results show that our feature processing is effective in improving the classification accuracy and we can use the model to predict the severity of traffic accident furthermore prevent traffic accidents.

Keywords-traffic accident; data mining; feature processing; real-time severity prediction

I. INTRODUCTION

Over the past decades, the rapid development of cities has generated a series of new problems and has aggravated the severity of existing problems. The problems related to traffic jam or accident should be resolved by the governments efficiently and quickly. According to the study of the World Health Organization[WHO 2015] ,there was 1.25 million deaths occurred due to road accidents in 180 countries in 2013 alone.

A huge amount of traffic accident data have been collected with the advancement in sensor technologies[1]. Using data mining methods such as classification, we can find out relations of traffic factors which would lead to accident[2]. Data mining is the exploration and analysis of large dataset in order to discover knowledge and patterns. Data mining is typically conceptualized as a three part process: preprocessing, learning, and post-processing[3]. The analysis to the massive data of traffic accidents using data mining technologies can give very helpful insights from the hidden patterns in the data. There are many unimpressive elements that will produce a traffic accident, just like weather and light condition, traffic flow, road condition and driver behavior.

The real-time traffic volume data and vehicle navigation system based on GPS will enable the driver to check the traffic information and select a less congested route to avoid traffic jams[4]. Our problem is, can we predict the severity of traffic accident for the driver rapidly according to the factors that we can collect in real-time. As we know, many factors will cause a traffic accident, like junction type, speeding, weather and road condition[5][6][7]. These factors can be obtained in real-time and make a traffic accident alarm model and we can develop an application like vehicle navigation system to take a warning to the driver. So, the driver will pay more attention to their current traffic environment if the application predicts that serious injury accident is much possible.

In our paper, we use the traffic accident dataset and pay attention to the factors that can be collected in real-time. We preprocessed the dataset using data mining technologies and used different classifiers to model training. The dataset after preprocessing will provide more accurate classification results. The main contributions of this paper can be concluded as follows: We focus on the factors that can be obtained in real-time and first use the method to deal with the time in the traffic accident time.

II. RELATED WORK

Many previous researches on traffic accident data tried to understand the main factors that would cause accident. Chen[4] used big and heterogeneous data on traffics and human mobility in Japan. The study is to understand how human mobility will affect traffic accident risk using Stack denoise Autoencoder to learn hierarchical feature representation of human mobility.

Gang [8] focused on the rough set theory to calculate and analyze the influence of factors on traffic accident morphology. They only used the information included in traffic accident statistics data. Deb[3] analyzed the incomplete traffic accident data and proposed a new Decision tree and sampling based missing value imputation algorithm for missing value imputation. Najada[9] used H2O and WEKA mining tools to evaluate five classifiers on traffic accident data set. The used classifiers are: Naive Bayes, C4.5, Random Forest,

978-1-5386-1618-5/19/$31.00 ©2019 IEEE
DOI 10.1109/ICNISC.2017.00059

AdaboostM1 and Bagging. And in their experiments, Naïve Bayes gave the optimum results.

III. Dataset

The traffic accident data that we want to analysis consists 158877 examples[10]. It displays the locations and attributes of collisions that occur within Seattle. All collisions are provided by the Seattle Police Department and Washington State Department of Transportation and are recorded by Traffic Records for the last 10 years. Figure 1 shows the number of traffic accidents group by month from 2010 to 2015. We can see that the number in the first half year is usually lower than those in the second half year according to the broken lines. Hence, there must be some hidden patterns in the traffic accident. The data is described by the features like address type, weather condition, incident time, collision type and so on.

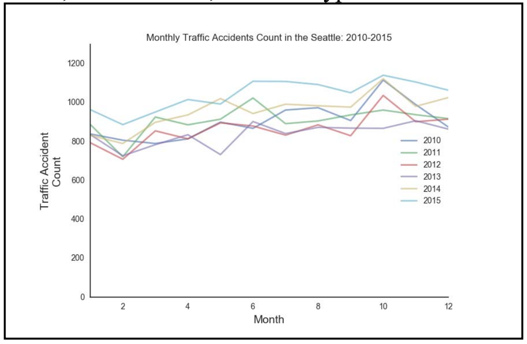

Figure 1. Monthly traffic accidents count in the Seattle 2010-2015

IV. Experiment

A. Data Preprocessing

1) Data cleaning

Our traffic accident dataset consists of 158877 records in Seattle, USA from January 2004 to November 2016 and every record has 43 features.

As a preprocessing before the data mining step, we cleared the dataset and removed the missing values and removed some of the factors that look irrelevant to our purpose of mining this dataset. The removed fields are 'Object ID', 'Incident Key', 'Collision Key' and so on.

2) Feature processing

Although we have many features, this does not mean that all of them will be important to generate a good result. Decreasing the number of attributes (features) will decrease the processing time and increase the prediction accuracy [11]. According to our needs, we selected the features that can be collected in real-time, shown in TABLE Ⅰ. These features can be provided by GPS or website [12]. We generated a dataset used these features as the first dataset in our work. And some features' categories are very detailed but they are only a small part of the dataset. So we reclassified some features to simplify the model[13].

TABLE I. Features in the dataset

Feature	Type	Category number
Address type	category	3
Incident time	time	\
Junction type	category	4
Light condition	category	7
Road condition	category	8
Speeding	category	2
Weather condition	category	9

We transformed some fields by grouping or discretizing their values. 'Weather condition', 'Road condition' and 'Light condition' were simplified to take new categorical values as shown in TABLE Ⅱ.

TABLE II. Features simplified

Feature	Value
Light condition	Daylight(65.9%)
	Dark-Street Lights On(27.7%)
	Other(6.4%)
Road condition	Dry(71.5%)
	Wet(27.1%)
	Other(1.4%)
Weather condition	Clear or Partly Cloudy(63.6%)
	Overcast(15.9%)
	Raining(19.2%)
	Other(1.3%)

According to the characteristics of feature, we need transform extraction[14][15]. Most of our features are categorical instead of continuous values . These categorical values can not be used directly in our learning model . In this work, we convert these values based on one-hot encoding. This method transforms each categorical feature with M possible values into M binary features.

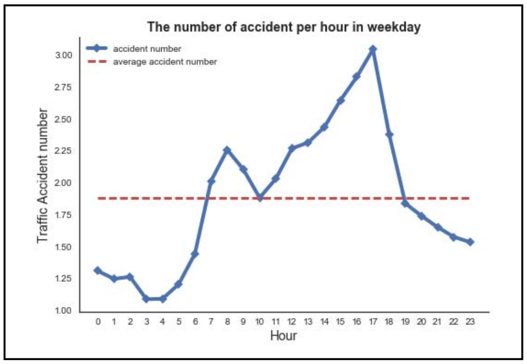

Figure 2. The number of accident per hour in weekday

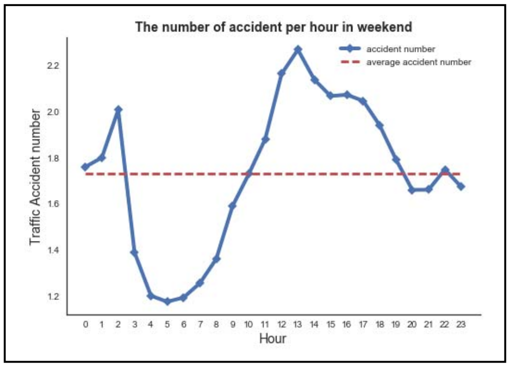

Figure 3. The number of accident per hour in weekend

3) Time period processing

We split the feature of 'Incident time' to two different features. One was simplified to just distinguish weekday and weekend. And the other one is based on the driver's daily life and travel habits. Travel conditions are similar in the same time period for different types of day. So the driving condition can be reflected by the number of traffic accidents per hour in some way. Figure2 and Figure 3 show the average number of traffic accident and the number of accident per hour. According to them, we generate a Time Flag shown in TABLE Ⅲ . Then we replace the factor of time to the feature that can respect the time segment.

TABLE III. TIME FLAG

Time	weekday	weekend	Time Flag
Hour	0,1,2,3,4,5,6, 20,21,22,23	3,4,5,6,7,8,9, 20,21,23	0
	7,8,9,10,11,12,13, 14,15,16,17,18,19	0,1,2,10,11,12,13, 14,15,16,17,18,19,22	1

In our work, a complete list of the features and their information of our traffic accident data after preprocessing that will be applied to data mining as the second dataset .

B. Prediction Metrics

In our work, we divided the dataset into two categories according the severity. The first category is: Property Damage Only Collision. The second category is Injury Collision that included Serious Injury Collision and Fatality Collision. The first class data accounted for 65.4 % and the second class data accounted for 34.6 %, which shows the imbalance of the data. Hence, we chosen F1-score and Area Under the ROC Curve (AUC) as the quality measures in our experiment. We used two different metrics because the F1-score considers both the precision and the recall of the test. And AUC will give better interpretation to the results which takes into account the imbalance of dataset.

C. Performance Evaluation

The classifiers used in our study are: Naïve Bayes, Random Forest, MLP and AdaBoost. We have compared the performance of the model using the first dataset and the second dataset which is processed using our methods. The TABLE Ⅳ shows that after feature processing, the dataset will provide more accurate classification results. The model can be used to predict the severity of the traffic accident as an alarm for the driver.

TABLE IV. PERFORMANCE EVALUATION

Dataset	First dataset		Second dataset	
Classifier	***F1-score***	***AUC***	***F1-score***	***AUC***
Naïve Bayes	0.67	0.74	0.81	0.88
Random Forest	0.71	0.77	0.84	0.90
MLP	0.71	0.77	0.85	0.90
AdaBoost	0.69	0.75	0.86	0.91

V. CONCLUSION

In our paper, we used the traffic accident data from Seattle to data mining. We focused on the factors that can be obtained in real-time such as weather and road condition and used some methods in the feature processing. These factors can be obtained in real-time and make a traffic accident alarm model. We reclassified some detailed features and transformed categorical features based on one-hot encoding. The feature of 'Incident time' was relabeled based on the driver's daily life and travel habits which reflected by the number of accident per hour. The classifiers used in our study are: Naïve Bayes, Random Forest, MLP and AdaBoost. The experimental results show that after feature processing, the dataset will provide more accurate classification results. The model can predict the severity of traffic accidents.

However, we still have some work to do. We will collect the human behaviors data and combine with traffic conditions data to generate a more comprehensive, effective and reliable model.

References

[1] Z.Zamani, M.Poumand and M.H.Saraee, "Application of data mining in traffic management: case of city of Isfahan," in: Proceeding of ICECT2010 Conference, KualaLumpur, Malaysia, 2010,pp.102–106

[2] Krishnaveni, S., and M. Hemalatha. "A Perspective Analysis of Traffic Accident using Data Mining Techniques," International Journal of Computer Applications 23.7(2011).

[3] R. Deb, and A Wee-Chung Liew, "Missing value imputation for the analysis of incomplete traffic accident," Information Sciences 339(2016) 274-289

[4] Q. Chen ,X. Song, H. Yamada and R. SHibasaki,"Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference," Proceeding of the Thirtieth AAAI Conference on Artifical Intelligence(AAAI-16)

[5] Park, Seong Hun, S. M. Kim, and Y. G. Ha. "Erratum to: Highway traffic accident prediction using VDS big data analysis," The Journal of Supercomputing 72.7 (2016): 2832-2832.

[6] Kim, Jeongmin, and K. R. Ryu. "Mining traffic accident data by subgroup discovery using combinatorial targets," Ieee/acs,

International Conference of Computer Systems and Applications IEEE, 2015:1-6.

[7] Fogue, Manuel, et al. “Using Data Mining and Vehicular Networks to Estimate the Severity of Traffic Accidents. Management Intelligent Systems,” Springer Berlin Heidelberg, 2012:37-46.

[8] T. Gang, H. Song, Y. Yan, M. Jafari . “Cause Analysis of Traffic Accidents Based on Degrees of Attribute Importance of Rough Set,” IEEE, Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE, Intl Conf on Autonomic and Trusted Computing and 2015 IEEE, Intl Conf on Scalable Computing and Communications and ITS Associated Workshops IEEE, 2015:1665-1669.

[9] Najada, Hamzah Al, and I. Mahgoub. “Big Vehicular Traffic Data Mining: Towards Accident and Congestion Prevention,” Iwcmc 2016 :, International Wireless Communications & Mobile Computing Conference2016.

[10] https://catalog.data.gov/dataset/seattle-collision-data-sdot

[11] H. Liu and H. Motoda,”Feature selection for knowledge discovery and data mining,” Springer Science & Business Media, 2012, vol. 454.

[12] Castro, Pablo Samuel, D. Zhang, and S. Li. Urban,”Traffic Modelling and Prediction Using Large Scale Taxi GPS Traces,” Pervasive Computing. Springer Berlin Heidelberg, 2012:57-72.

[13] Fan, Xiaoliang, et al. “Big Data Analytics and Visualization with Spatio-Temporal Correlations for Traffic Accidents,” Algorithms and Architectures for Parallel Processing. Springer International Publishing, 2015.pp.255-268,doi:10.1007/978-3-319-27122-4_18

[14] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, no. 1-2, pp. 273–324, 1997, special issueon relevance.

[15] Zhang, Fengli, and D. Wang. “An Effective Feature Selection Approach for Network Intrusion Detection,” IEEE Eighth International Conference on Networking, Architecture and Storage IEEE, 2013:307-311.