Postprint: A Combined Model-Based Method for Traffic Accident Severity Prediction

Authors: Shi Xuehuai, Qi Yong, Zhang Weibin, Li Qianmu

Date: 2018-07-23T00:00:00+00:00

Abstract

Due to the limitations of individual single classification models in predicting road traffic accident severity, we endeavor to establish a combined model to enhance prediction performance. By leveraging convolutional neural networks to extract feature information from spatio-temporal dimensions and employing a stacking approach to combine CNN with XGBoost, we ultimately generate a classification model for road traffic accident severity (multi-layer boosting algorithm). Experimental results demonstrate that this model achieves a prediction accuracy of 91.51% on the test set, with the combined model yielding superior classification results compared to single classification models. Based on the classification results of the combined model, we conduct importance ranking of traffic accident features and perform feature correlation analysis, providing a reference basis for management measures aimed at reducing road traffic accidents and mitigating accident severity levels.

Full Text

Preamble

Traffic Accident Severity Prediction Method Based on Ensemble Model

Shi Xuehuaia, Qi Yonga†, Zhang Weibinb, Li Qianmua

(a. School of Computer Science & Engineering; b. School of Electronic Engineering & Photoelectric Technology, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract: Due to the limitations of individual classification models in predicting road traffic accident severity, this study proposes an ensemble approach to improve prediction performance. By combining convolutional neural networks (CNNs) to extract feature information from spatial dimensions and stacking

CNN with XGBoost, we develop a classification model for traffic accident severity (multi-level boosting algorithm). Experimental results demonstrate that the proposed model achieves a prediction accuracy of 91.51% on the test set, outperforming single classification models. Based on the classification results, we rank the importance of accident features and conduct correlation analysis, providing reference for management measures to reduce traffic accidents and mitigate their severity.

Keywords: traffic safety; traffic accident severity; XGBoost; convolutional neural network; inducement analysis

0 Introduction

Traffic accidents have become a critical societal problem causing severe harm worldwide. In China alone, road traffic accidents result in massive casualties and economic losses each year. In 2016, over 150,000 traffic accidents with casualties occurred, causing more than 100,000 deaths and injuries and direct economic losses exceeding 5 billion yuan [?]. Scientific prediction of traffic accident severity, combined with inducement analysis, is essential for understanding the patterns and characteristics of high-severity accidents, thereby enabling effective traffic safety countermeasures to minimize risks.

Numerous studies have investigated traffic accident severity prediction \cite{2~12}. Domestic research includes Ma Zhuanglin et al. [?], who employed ordered Logit and generalized Logit models to predict accident severity in highway tunnels, and Liu Haizhu [?], who identified primary influencing factors from a macro perspective and established a cumulative Logistic regression model with three severity levels. International studies include Fogue et al. [?], who developed an intelligent classification system using features such as vehicle speed, type, and airbag status to guide emergency services, and Sameen et al. [?], who applied recurrent neural networks for deep learning-based severity classification.

Given that accident data exhibits sequential characteristics with spatiotemporal correlations, O'Donnell et al. [?] utilized ordered Logit and probit models to analyze relationships between severity and driver/vehicle attributes, while Sameen et al. [?] employed recurrent neural networks with multiple LSTM units and fully connected layers with softmax for classification. However, existing models primarily predict severity based on accident outcome information without capturing sufficient road segment data. For instance, Ma et al.' s tunnel models [?] lacked adjacent segment information, Fogue' s system [?] focused on emergency guidance rather than root cause analysis, and Sameen' s RNN model [?], which used road and environmental features, achieved only 71.77% accuracy on the validation set without leveraging neighboring segment information.

To address these limitations, this study extracts adjacent road information through road IDs, employs CNNs to capture spatial dimension features, and combines CNN with XGBoost via stacking to develop a traffic accident severity classification model (multi-level boosting algorithm). This approach enables both improved prediction accuracy and inducement analysis to support accident reduction and severity mitigation strategies.

1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), first proposed by Fukushima et al. in 1980 [?] and later optimized by LeCun in 1998 [?], have become fundamental to deep learning. The architecture is illustrated in [Figure 1: see original paper]. CNNs primarily consist of convolutional layers, rectified linear unit (ReLU) layers, pooling layers, fully connected layers, and loss function layers. Compared to traditional neural networks, CNNs offer several advantages: (a) they capture regional correlations in unstructured data; (b) they encapsulate feature extraction, simplifying training; (c) they simultaneously learn and generate feature extraction and classification; and (d) weight sharing reduces training parameters, simplifying network structure and enhancing adaptability [?]. However, CNNs also have limitations: (a) their characteristics make generalization difficult; and (b) extensive computations in convolutional and pooling layers require significant hardware support [?]. CNNs are now widely applied in image recognition, video analysis, natural language processing, drug discovery, and game playing.

1.2 XGBoost

XGBoost, proposed by Chen et al. in 2016 [?], is based on Gradient Boosting Decision Trees (GBDT, also known as MART—multiple additive regression tree), an iterative decision tree algorithm where predictions from all trees are summed for the final result [?]. XGBoost optimizes GBDT by providing cache-aware prefetching, distributed external memory computing, and AllReduce fault tolerance tools, enhancing computational speed and addressing the limitation to million-scale datasets. The structure is shown in [Figure 2: see original paper], with detailed descriptions available in [?, ?].

2 Model Establishment

2.1 Strong Association Rule Feature Selection Method

According to the GA1082 general accident collection standard, the Shenzhen accident dataset collects 69 features per accident. We categorize these features based on traffic safety engineering principles [?], as detailed in .

** Feature classification of Shenzhen accident data (2014-2016)**

Using the strong association rule feature selection algorithm, we prioritize short rules with the target label as the consequent to reduce feature space dimensionality. The algorithm pseudocode is presented in Algorithm 1. Based on the selection results, we obtain feature sets for the target road segment and its three adjacent segments, combined with other features to predict severity. The selected features are shown in .

Algorithm 1 Strong Association Rule Feature Selection Algorithm

```
Input: data set X, feature set F, association rule set ϒ, target label y,
       iteration number T, support threshold  , confidence threshold  ,
       final feature set M
Initialize: M ← {}
for each rule   in ϒ do
  if y in consequent( ) and lift(y, ) > 1 then
    Φ ← sort(ϒ) by support and confidence
  end if
end for
while |M| < T do
  get first rule from Φ
  add features from first rule to M without repetition
end while
Output: M
```

** Feature set selected by strong association rule method**

2.2 Accident Severity Prediction Method

In deep learning, features are transmitted across layers. Inspired by this, we combine CNNs with XGBoost using stacking: CNNs extract spatial information from the target and three adjacent road segments, which is then integrated with other features and fed into XGBoost. We term this the multi-level boosting algorithm, with pseudocode in Algorithm 2 and flowchart in [Figure 3: see original paper].

Algorithm 2 Multi-Level Boosting Algorithm

```
Algorithm configurations: parameters in CNN, parameters in XGBoost
Input: data set X (include spatial features Xs, other features Xo),
       target label y, cross-validation parameter k, performance threshold
Initialize: level l ← 0
while performance improvement >   do
  fit CNNs(Xs) to extract 3-dim spatial features
  concatenate spatial features with Xo
  fit XGBoost(X) to predict severity
  get 3-dim prediction vector as enhanced features
  k-fold estimation(X, y, k)
  if performance not improved then
    break
  end if
  l ← l + 1
```

```
14: end while
15: Output: ŷ = argmax(ŷXGBoost)
```

The number of levels is automatically determined using k-fold cross-validation, inspired by deep forest [?]. If prediction performance shows no significant improvement, layer expansion stops.

3 Experimental Analysis

The experimental data was provided by the Road Traffic Safety Research Information Sharing Platform' s Shenzhen Traffic Accident Data Analysis Competition, containing over 200,000 accident records from 2014-2016, including both casualty and minor accidents. Using road and environmental information, we predict accident severity and evaluate results via 5-fold cross-validation.

3.1 Prediction Results Based on Ensemble Model

The multi-level boosting algorithm tunes parameters using 5-fold cross-validation, determining the optimal number of layers as 2. Final evaluation yields a prediction accuracy of 91.51%. presents evaluation metrics where acc-on-trainSet indicates training accuracy, acc-on-cv indicates cross-validation test accuracy, and the merror metrics represent multi-class error means and standard deviations across folds. Results show: (a) all models (RF, XGBoost, ensemble) exhibit better fitting on training sets, indicating overfitting; (b) the ensemble model outperforms individual models on both training and test sets, with a test multi-class error standard deviation of 3.90%, demonstrating stable and robust performance.

** Comparison of prediction accuracy across different models**

3.2 Error Analysis and Comparison

To further validate accuracy, we compare the multi-level boosting algorithm against single XGBoost and Random Forest (RF) models. As shown in , the ensemble model achieves superior performance and stability.

3.3 Inducement Analysis

Based on ensemble model predictions and expert consultation, we identify influential features from road and environmental information sets. Feature importance ranking is presented in [Figure 4: see original paper], revealing that road width, visibility, roadside protection facility type, intersection type, road physical isolation, and traffic signal control are particularly significant.

[Figure 4: see original paper] Feature importance ranking of the model

[Figure 5: see original paper] Architecture of the proposed multi-level boosting algorithm

Figures 6 and 7 illustrate bivariate relationships between features and accident types. For categorical variables, numerical transformations are applied. In "traffic signal control," categories 1-6 (no control, police direction, signal lights, signs, markings, other) are combined into 11 types. Detailed explanations appear in .

** Explanation of feature values**

[Figure 6: see original paper] Relationships between different features and "accident type" (violin plots)

[Figure 7: see original paper] Relationships between different features and "accident type" (KDE plots)

Analysis reveals:

a) Road width: Violin plots show fatal and injury accidents occur more frequently on wider roads, with fatal accidents particularly concentrated on wide segments. The Pearson correlation coefficient between road width and accident type is 0.23 (p=0.0088), indicating significant correlation.

b) Traffic signal control: Accidents are more likely under control type 8 (signs and markings). Compared to property-damage-only accidents, this control type shows higher frequencies of injury and fatal accidents, with a Pearson correlation of 0.32 (p=0.00022).

c) Visibility: Contrary to conventional wisdom, most accidents occur at visibility level 4 (200m+), not in low visibility conditions. Injury and fatal accidents are more common at this visibility level, though the Pearson correlation of 0.038 (p=0.67) suggests weak association.

d) Roadside protection facility type: Types 1, 2, and 6 (green belts, concrete barriers, no protection) show higher accident probabilities. Green belts and no protection are particularly associated with injury and fatal accidents. The Pearson correlation of 0.087 (p=0.33) indicates limited association.

e) Road physical isolation: Isolation types 1 and 4 (center isolation, no isolation) exhibit higher accident probabilities. Center or no isolation shows greater risk than machine-non-machine isolation or combined isolation, with fatal accidents more frequent at four-way intersections. The Pearson correlation is 0.17 (p=0.053).

f) Intersection type: Most accidents occur at types 2 and 5 (four-way intersections and regular road segments). Type 2 shows higher injury accident rates. The Pearson correlation of -0.3 (p=0.00051) indicates strong association.

In summary, higher-severity accidents are more likely on wider roads, under sign-and-marking control, with visibility exceeding 200m, and with green belts or no roadside protection. These findings should inform targeted countermeasures.

4 Conclusion

This study proposes a traffic accident severity prediction method based on an ensemble model, combining CNNs for spatial feature extraction with XGBoost via stacking (multi-level boosting algorithm). Experiments on 2014-2016 Shenzhen accident data demonstrate that the ensemble model outperforms single classifiers, achieving 91.51% accuracy. Feature importance ranking and correlation analysis provide valuable references for accident reduction and severity mitigation strategies. Due to platform requirements, the data cannot be publicly released; however, experimental code is available at https://github.com/SXHSine/TrafficAccidentAnalysis.

Future research will leverage spatial accident data to more accurately capture adjacent segment information and incorporate driver attributes and behavioral data for enhanced prediction and inducement analysis.

References

[1] Statistics of National Road Traffic Accident Data in 2016 [N/OL]. [2017-12-14]. http://www.peichang.cn/detail/id18666.html.

[2] Ma Zhuanglin, Zhang Yiyi, Yang Yang, et al. Research on models for predicting severity of traffic accident in highway tunnel [J]. China Safety Science Journal, 2015, 25(5): 75-79.

[3] Sun Chongjing, Xin Feifei. An overview of the conflict indicators between vehicles and pedestrians [J]. Journal of Transport Information and Safety, 2016, 34(2): 9-16.

[4] Ma Zhuanglin, Shao Chunfu, Dong Chunjiao, et al. Temporal-spatial analysis model of traffic accident severity based on cumulative logistic model [J]. China Safety Science Journal, 2011, 21(9): 94-99.

[5] Hou Shuzhan, Sun Xiaoduan, He Yulong, et al. Relationships between crash severity and traffic flow characteristics on freeways [J]. China Safety Science Journal, 2011, 21(9): 106-112.

[6] Liu Haizhu. The analysis of influencing factors for crash severity and the establishment of predication model [D]. Changchun: Jilin University, 2014.

[7] Ma Zhu, Chen Yuren, Zhang Lanfang. Influence factors of accident severity for urban road [J]. Journal of Chongqing Jiaotong University: Natural Sciences, 2014, 33(1): 111-114.

[8] Fogue M, Garrido P, Martinez F J, et al. Using data mining and vehicular networks to estimate the severity of traffic accidents [M]// Management Intelligent Systems. Berlin: Springer, 2012: 37-46.

[9] Sameen M I, Pradhan B. Severity prediction of traffic accidents with recurrent neural networks [J]. Applied Sciences, 2017, 7(6).

[10] Khera D, Singh W. A review on injury severity in traffic system using various data mining techniques [J]. International Journal of Computer Applications, 2014, 100(3): 17-22.

[11] Gupta M, Solanki V K, Singh V K. A novel framework to use association rule mining for classification of traffic accident severity [J]. Ingeniería Solidaria, 2017, 13(21): 37-44.

[12] Baru A. Injury severity levels and associated factors among road traffic accident victims referred to emergency departments of selected public hospitals in Addis Ababa, Ethiopia: the study based on haddon matrix [J].

[13] O' Donnell C J, Connor D H. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice [J]. Accident Analysis & Prevention, 1996, 28(6): 739-753.

[14] Xiao Guiping, Zhu Xiaoning. Road safety engineering [M]. Beijing: China Railway Publishing House, 2011.

[15] Fukushima K, Miyake S. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition [M]// Competition and cooperation in Neural Nets. Berlin: Springer, 1982: 267-285.

[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[17] Convolutional Neural networks (LeNet): DeepLearning 0.1 documentation. DeepLearning 0.1. LISA Lab. [2013-08-31].

[18] Olshausen B A. Emergence of simple-cell receptive field properties by learning a sparse code for natural images [J]. Nature, 1996, 381(6583): 607-609.

[19] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[20] Sabes P N, Jordan M I. Advances in neural information processing systems [C]// Advances in Neural Information Processing Systems. 1995.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(2): 2012.

[22] Chen Tianqi, Guestrin C. XGBoost: a scalable tree boosting system [C]// Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 785-794.

[23] Chen Tianqi, He Tong, Benesty M, et al. XGBoost: extreme gradient boosting [EB/OL]. (2017). https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf.

[24] Friedman J H. Greedy function approximation: a gradient boosting machine [J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[25] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297-336.

[26] Zhou Zhihua, Feng Ji. Deep forest: towards an alternative to deep neural networks [J]. arXiv preprint arXiv:1702.08835, 2017.

Note: Figure translations are in progress. See original paper for figures.