Traffic Crash Severity Prediction with Deep Learning

Xining Chen
School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100000, China
19120794@bjtu.edu.cn

Abstract. This paper is based on deep learning to explore the accuracy forecasting model of traffic crash severity. A prediction model in terms of convolutional neural networks is proposed in the paper. The first is to collect the basic information of all traffic crash data of the entire urban road network in Chicago from 2016 to 2020. Then using the PivotTable to observe the distribution of each variable, and filtering them. The required prediction model is based on Python and it mainly includes three parts, the input layer, the hidden layer, and the output layer. Besides, using SGD optimizer to accelerate the training of the network model is a need. Aiming to avoid problems such as over-fitting, the dropout technique is adopted, and its parameter value taking 0.1 is better. Then comparing and analyzing with three traditional machine learning models, the decision tree classification prediction model, the logistic regression prediction model and the support vector machine prediction model, it is obvious to show the better performance of prediction model based on the convolutional neural network method. After optimization, its highest accuracy is 89.2%.

1. Introduction
According to data provided by the World Health Organization, the amount of deaths in the world is far more than 1 million per year due to the road traffic crashes. Under the background, this paper takes the traffic accident data of the road network in Chicago, USA as the research foundation, and combines the deep learning to explore the impact of various factors on the severity of traffic accidents and make as much as possible accurate predictions[1]. Therefore, it is possible to propose effective accident prevention measures based on the results, and start with the factors that have a relatively large impact, and handle each link. In addition, the scientific prediction of the severity of traffic accidents can provide relative technical services for the traffic safety management department.

2. Data Processing and Analysis
Due to different national conditions, the definitions of traffic crashes in different countries around the world are not the same. The United States defines road traffic accidents as follows: According to the National Security Council, road traffic crashes are unforeseen, harmful, or dangerous events that occur on the road.

2.1. Classification of Road Traffic Crash Severity
According to the traffic crash data of Chicago for the whole year of 2018, the following results can be obtained through data perspective.
Table 1. Most severe injury percentage table

| Most Severe Injury (%)          |     |
|--------------------------------|-----|
| Fatal                          | 0.09% |
| Incapacitating injury          | 1.78% |
| No indication of injury        | 86.15% |
| Nonincapacitating injury       | 7.79% |
| Reported, not evident          | 4.18% |
| Total                          | 100% |

This is a pivot table about the severity of a traffic crash. As can be seen from the above table, the proportion of “No indication of injury” is the heaviest, up to 86.15%, which indicates that most of the road traffic crashes that occurred in Chicago in 2018 were property losses and no injured. “Nonincapacitating injury” and “Reported, not evident” account for a relatively low percentage of all traffic accidents, so they can usually be classified as one type and slightly injured, which helps to improve the accuracy of the predictive model and prevent over-fitting. “Fatal” is very low, only 0.009%. This data shows that in road traffic accidents, there are very few people who die. In order to facilitate the later optimization process, it should be classified into “Incapacitating injury”, and they are classified as serious injury.

Therefore, in this paper, the severity of road traffic accidents is divided into three standards, no indication of injury, nonincapacitating injury and incapacitating injury.

2.2. Data Processing and Analysis

The data selected in this paper is from the traffic accident database in Chicago, USA. Since there are many feature variables involved, it needs to delete unessential feature variables according to some principles. The principle of judgment is: whether the weight of each influencing factor under a certain characteristic variable is uniform; whether the unknown quantity in a certain characteristic variable accounts for a large proportion; according to experience and other information, whether a certain characteristic variable itself is related to traffic severity with a greater impact[2]. Then, using the pivot table in Excel to analyze each characteristic variable and filter it, 14 characteristic variables that have the greatest impact on the severity of road traffic accidents were obtained.

The initial processing above only filters the feature variables, but some problems in the original data are not solved, so the next step is to assign values to the original data. Before that, using Excel to delete unknown data is a need. The final useful data volume is 57472. Then we have to perform different assignments according to the characteristics of each feature variable.

Table 2. Characteristic variable assignment table

| Assignment                          |     |
|-------------------------------------|-----|
| Posted speed limit                  | 1, 2, 3 |
| Traffic control device              | 1, 2, 3, 4 |
| Weather condition                   | 1, 2, 3, 4 |
| Lighting condition                  | 1, 2, 3, 4, 5 |
| First crash type                    | 1, 2, 3, 4, 5, 6 |
| Trafficway type                     | 1, 2, 3, 4, 5 |
| Roadway surface condition           | 1, 2, 3 |
| Vehicle year                        | 1, 2, 3 |
For example, the characteristic variable of weather conditions is assigned values in turn according to the proportion of each category. The first category is the largest, and clear is assigned a value of 1; the second category is rain, which is assigned a value of 2; the third category is snow, which is assigned a value of 3; the fourth category is others, which is assigned a value of 4.

### 3. Model Construction

The use of deep learning technology is becoming more and more extensive, so there are many deep learning predictive model frameworks that can be used as references[3]. This model construction is based on a relatively mature convolutional neural network. The environment where the entire model is built is Python.

- **Read the data:**
  - It is necessary to store the preprocessed data in the "csv" file format, which is conducive to data reading in the Python environment.

- **Data division:**
  - After reading the data, it is divided according to the standard that the test set accounts for 20% and the training set accounts for 80%.

- **Decide the input layer structure:**
  - In this paper, a fully connected layer is used as the input layer. Since there are 14 feature variables, the number of nodes in the input layer is 14.

- **Decide the hidden layer structure:**
  - The initial choice is to set the hidden layer to five layers. The number of neuron nodes in the first, second, third, fourth and fifth layers are 100, 75, 50, 25 and 10, respectively. The activation function of each layer uses the ReLU function.

- **Decide the output layer structure:**
  - The output layer also uses a fully connected layer. Since a model with an output classification result of 3 needs to be built, the number of nodes in the output layer is set to 3. It uses the Softmax activation function.

- **Decide the optimizer:**
  - The current model selects the optimizer "Adam" with a learning rate of 0.1.

- **Output results:**
  - Through the above steps, this neural network prediction model can be built.

### 4. Result Analysis and Optimization

#### 4.1. Result Analysis

For the output results, this paper analyzes the accuracy, sensitivity, recall rate, confusion matrix, and receiver operating characteristic curve.

- **Accuracy:**
  - The accuracy of deep learning prediction model is 86%; the accuracy of SVM prediction model is 83.9%; the accuracy of logistic regression prediction model is 83.2%; the accuracy of the decision tree is 76.5%.

- **Sensitivity:**
The sensitivity of deep learning prediction model is 76.4%; the sensitivity of SVM prediction model is 82.3%; the sensitivity of logistic regression prediction model is 83.7%; the sensitivity of the decision tree is 85.5%.

- **Recall rate.**
The recall rate of deep learning prediction model is 85.7%; the recall rate of SVM prediction model is 84%; the recall rate of logistic regression prediction model is 83.2%; the recall rate of the decision tree is 76.4%.

- **Confusion matrix.**
  For the same data set, in the confusion matrix of the deep learning prediction model, the amount of data on the diagonal is the most, and the amount of data on the off-diagonal is less, indicating that the prediction result of the model is good. The performance of the remaining three traditional forecasting models is fairly good.

- **Receiver operating characteristic curve.**
The most important of the curves is the micro-average receiver operating characteristic curve. The curve performance of the deep learning prediction model is the best, followed by the SVM prediction model and the logistic regression prediction model, and the worst performance is the decision tree prediction model.

4.2. **Optimization**

Although there are many indicators for judging the pros and cons of the model, the optimization evaluation indicators here are based on accuracy[4].

- **Increase data amount.**
  In the previous article, only the relevant data of Chicago in 2018 was processed. Now, in order to increase the amount of data, the relevant data of 2016, 2017, 2019 and 2020 have been processed. After running the same code, the accuracy rate increased from 86% to 88.9%.

- **Adjust number of hidden layers and neurons.**
  First, the number of hidden layers is reduced, and then gradually increased, and the corresponding number of neurons is adjusted appropriately, combined with the increased amount of data, after running the code many times, the highest accuracy rate is 88.5%.

- **Add dropout layers.**
  The main function of the Dropout layer is to avoid over-fitting of the prediction model, so in general, after the dropout layer is added, the prediction result of this prediction model will be worse. After adding the dropout layer and running the code, the highest accuracy rate is 88.2%.

- **Adjust optimizer.**
  The three most commonly used optimizers are SGD, Adam and RMSprop. Calling them in turn and running the code, the highest accuracy rates are 89.2%, 88.2%, and 87.9%. Therefore, the optimal optimizer is SGD.

- **Adjust relative parameters.**
  The adjustment of these parameters has little effect on the results.

After considering all the above optimization factors, the best accuracy of this deep learning prediction model is 89.2%.

5. **Conclusion**

This article uses the data on the severity of road traffic accidents obtained from Chicago, the United States, and after relevant processing, is imported into the constructed deep learning prediction model to run, and the output results are analyzed with multiple indicators. Whether it is before or after optimization, comparing the three traditional prediction models, the deep learning prediction model constructed in this article has better performance. However, the model constructed in this article is relatively simple, and neither the number of selected feature variables nor the number of output types is very large, so its prediction results only have a certain reference value. There is still much room for digging to predict the severity of road traffic accidents based on deep learning.
Acknowledgments
Thanks to the National Natural Science Foundation of China's project "Real-time query and visualization of massive multi-source smart traffic data streams based on cloud computing"; thanks to the Chicago road traffic accident database; thanks to teachers and students for their help.

References
[1] Xuehuai S., et al. (2019) Traffic Accident Severity Prediction Method Based on Combined Model. Application Research of Computers, 08: 1-9.
[2] Chen, X. (2018) Short-term forecasting of traffic flow and speed: a deep learning approach. CICTP 2018 Intelligence, Connection, and Mobility, 21: 86-96.
[3] LeCun, Y., Y. Bengio and G. Hinton. (2015) Deep learning. Nature, 521: 436-444.
[4] Yang, H., T. Dillon and Y.P. Chen. (2017) Optimized Structure of the Traffic Flow Forecasting Model with a Deep Learning Approach. Neural Computation, 10: 2371-2381.
[5] Schmidhuber, J. (2015) Deep learning in neural networks: An overview. Neural Networks, 61: 85-117.