Design Engineering

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

Large-Scale Traffic Accident Data Classification Method Based on XGBoost

Jie Liu
Department of Computer and Information Engineering, Guangxi Vocational Normal University, Nanning, Guangxi, China

Abstract:

Analyzing the subjective and objective of accidents causes through information from large-scale traffic accident databecomes more and more important nowadays. It is a significant reference for traffic safety prevention by researching classify large-scale traffic accident data to predict traffic accidents. In this paper, we have taken California traffic accident big data which is from US-ACCIDENT data set as our test subject, then introduces the XGBoost algorithm to classify and predict the severity of traffic accidents. The experiment results show the XGBoost algorithm looks significantly better by comparing it with several common classification algorithms. Not only in data fitting degree but also classification prediction accuracy, XGBoost works well than the other similar algorithm models. Our methods will improve the accuracy of traffic accident severity prediction to a certain extent, and provides further analysis and early warning for traffic accidents, provide an important reference for decision-making by government apparatus.

Keywords: *XGBoost, Traffic accident, Classification, Integrated learning.*

I. INTRODUCTION

According to statistics, there are tens of thousands of deaths direct or indirectly caused by traffic accidents in CHINA each year. The data shows that only in 2003, Death Toll exceeds 100,000. Although some measures have been taken by the government, in 2019, it dropped to 62,000, but it remained above 50,000 each year. Occur traffic accidents has certain accidental reasons, not only because the subjective cause like human driving hobbies but also related to various objective factors such as traffic environment and human element. How to analyze historical related traffic accident data and discover the causes of traffic accidents, in what methods we can adopt relevant policies and measures to lessen the number of traffic accidents? The prevention strategy provided by the transport department to reduce effectively the severity of traffic accidents is still one of the current hot issues in the transport field.

Most previous research into traffic accident prediction has focused on neural network related algorithms. According to the annual number of traffic accidents in Shandong province

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

from 2006 to 2016, Wang Xiaofan [2] used GM grayscale operator combined with BP neural network fitting method to provide a prediction. To find out the timing dependence in accident data, Zhang Zhihao [3] constructed a LSTM recursive neural network model to predict the traffic safety level, while Zhu Boya [4] established a spatio-temporal influence prediction model to study traffic accidents by using nonlinear regression and BP neural network. However, there is some evidence to suggest the neural network related methods has strict requirements on data, great difficulty in parameter adjustment and high requirements on training hardware. So it is not suitable for processing some data set, which feature data may abnormal, or missing values with high noise level, and it's hard to deal with for neural network algorithms. The other scenario that's worse iswhen neural network algorithmsface small datasets will cause overfit in training.

Some researchers study traffic accidents with specific regions and landforms, such as the literature [5-7], which is mainly adopts integrated learning model to study occur traffic accidents and environmental reasons of relevance.Zhang has build an AdaBoost classifier to forecast the traffic accidents [7], while He Ke and others by using the method of principal component analysis and random forest, to predict severity of tunnel traffic accidents.First according PCA to analyze 16 kinds of dependent variable, then adjust the parameters of the random forest by OOB error, Finally, MAE and RMSE are used to evaluate the model error [8].

Although there have been many relevant studies on traffic accident prediction, research articles consist of the following two parts. A fairly small data set with inadequate training samples , which can only describe a particular geographical location or specific scene; too much manual design is added in model training, which may easily cause overfit and affect the generalization ability of the learning model. In this paper, we have based on the traffic accident data of about 300,000 traffic Accidents in California, the most serious among the open source data set: US-ACCIDENTS [1], the XGBoost [9] integrated learning algorithm was introduced to conduct prediction research on traffic Accidents and verify the effectiveness and feasibility of the algorithm.

II. MATERIALS AND METHODOLOGY

2.1 XGBoost Algorithm Framework

2.1.1 Algorithm Features

XGBoost[9] is a called "extreme gradient boosting tree" supervised learning, which is composed of GDBT[14] algorithm to improve learning effect. XGBoost is designed base on GDBT algorithm, it did many improvement to this algorithm. There are several similarities between XGBoost and other tree model, such as decision tree[10], gradient boosting tree[12], random forests[16]. Recent research has shown that XGBoost had perform a better job compare with the other tree model, especially in training effect, prevent overfitting. What is more surprising is the parallel computing of the XGBoost, the training time can be reduced objectively. In general, XGBoost mainly has the following characteristics: (1) Regularization control model is introduced to prevent overfit, and reduce the model complexity. By the original

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

loss function, XGBoost algorithm creatively introduces the regularization function to evaluate the complexity of the learning model, the connotation of the algorithm can be described as follow formula:

$$Obj(\theta) = \mathcal{L}(\theta) + \Omega(\theta) \quad (1)$$

$Obj(\theta)$represents the objective we aim to learn,the optimal split subtree is found by calculating this function,$\mathcal{L}(\theta)$called the loss, while $\Omega(\theta)$represents the regularization. The learning goal is to minimize the objective, which requires to optimize the loss and the regularization to reach a balance, which effectively restrains the overfitting problem of the learning model.

(2) XGBoost algorithm is suitable for sparse data, since it can do effective treatment for lack of diverse data. During the data collection, missing is the unavoidable exist problem. Missing in the data set may cause distortion in the learning for some algorithm, such as neural network. There is mounting evidence that neural network has a high requirement of the consistency of the data set. On processing diverse sparse data sets, XGBoost shows an even greater advantage, because there is an internal automatic data filling in the algorithm, then for increases the robustness of the algorithm.

(3) CART tree supports parallel construction to reduce training time. For integration tree algorithm, tree construction is a time-consuming part. XGBoost algorithm can build regression tree in parallel according to the multicore characteristics of hardware platform, which improves the efficiency of algorithm.

(4) XGBoost adopts the depth-first tactics when tree splitting, which is helpful for the algorithm to jump out of the problem of local best solution. XGBoost also improves calculate performance, the best split point search can be described by Fig 1 as follows.

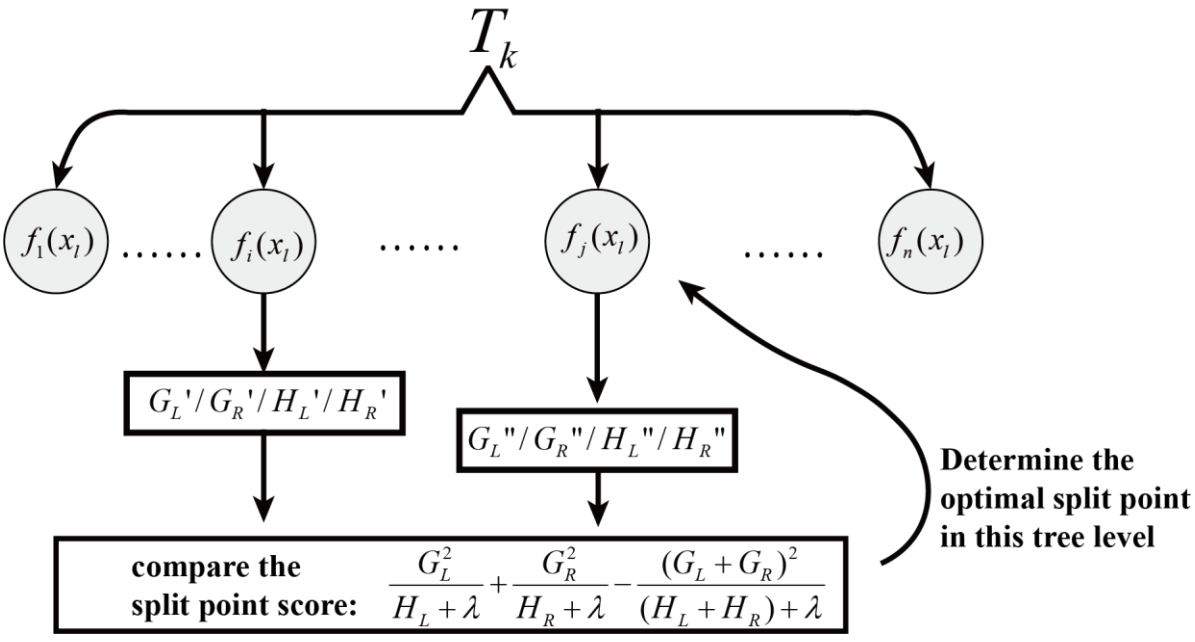

Fig 1: Searching the best score of split point

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

2.1.2 Specific Steps of Algorithm of XGBoost

XGBoost specific training as shown in Fig 2, according to the different stages, this can be divided into the following four steps:

Step 1: Given training sample$I < x, y >$, $, y$ stands for labels. Loss can be chose according to the nature of machine learning tasks. For example, regression tasks can use formula (2) as loss of square error, while classification task can pick up formula (3): a cross entropy loss.

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{2}$$

$$-y_i \cdot \ln\hat{y}_i - (1 - y_i) \cdot \ln(1 - \hat{y}_i) \tag{3}$$

Step 2: Select one of the candidate eigenvector x_i, and this feature is selected as an imaginative split point to define the split conditions. All the training samples can be divided into the left subtree I_L and right I_R, then calculate the matching subtrees respectively by G_L, G_R, H_L, H_R,which can be shown as follows:

$$\begin{cases} G_L = \sum_{x_i \in I_L} g_i \\ G_R = \sum_{x_i \in I_R} g_i \end{cases}, \begin{cases} H_L = \sum_{x_i \in I_L} h_i \\ H_R = \sum_{x_i \in I_R} h_i \end{cases} \tag{4}$$

The g_i is Taylor first derivative of the exhibition of loss, while $h_i$is the second derivative of The Taylor expansion. The t iteration is calculated as follows:

$$\begin{cases} g_i = \partial_{\hat{y}_i^{(t-1)}} l((y_i, \hat{y}_i^{(t-1)}) \\ h_i = \partial^2_{\hat{y}_i^{(t-1)}} l((y_i, \hat{y}_i^{(t-1)}) \end{cases} \tag{5}$$

Then the gain fraction with this feature as the splitting condition is calculated as follows:

$$\text{score} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \tag{6}$$

In this step, the eigenvector that makes the gain fraction maximum is calculated from all the eigenvectors, and the point is taken as the splitting node of the tree level.

Step 3: Decide whether the maximum depth of the split tree is reached. If so, update the weight w^*(leaf node) with Formula (7). jrepresent direction of the subtree to which the node falls $j \in (L, R)$. Then go to construct next tree; otherwise, add the layers by 1 and take the remaining feature vectors as new candidate subsets and build the next subtree.

$$w^* = -\frac{G_j}{H_j + \lambda} \tag{7}$$

Step 4: Finally, All the trees have been built in this process. After specified number of update iterations has been reached, latest weight will be reached. Training will be finished and enter the effect testing stage.

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

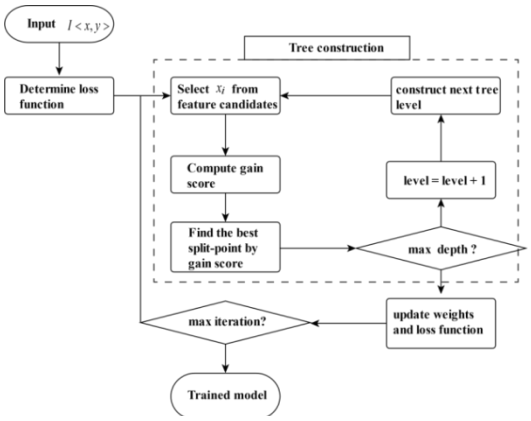

Fig 2: XGBoosttraining process

III. DATAANALYSIS

3.1 Data Description

To corroborate the algorithm's classification capacity on large-scale data, the national traffic accident statistics of the United States from February 2016 to March 2019 were adopted [1]. There is about 300,000 data related accident records of California, were screened out.The data is divided into two subsets 80% for the training set, 20% for testing set, and take accident severity as learning label.

In the data set of US-accidents [1], each traffic accident recorded four category attributes related to the event, region, climate and POI, the specific classification is described as follows:

(1) event attributes: This group of attributes includes the geographical location of the traffic accident (measured by GPS latitude and longitude data), incident type (accident, vehicle damage, traffic maintenance engineering major holiday, influence, for example.) , accident severity (United States Department of Transportation according to the influence of accident, divided accident severity into 1 ~ 4 grades, represent the mild, moderate, serious, very serious.), start time and end time of accident.

(2) Regional attribute: describes the administrative division of the geographical location of traffic accidents, and details the region-related attributes such as street signs and road directions of counties, townships, towns or urban areas.

(3) Climatic properties: describe the climatic conditions of the traffic accident. Including temperature, humidity, wind direction, wind speed, atmospheric pressure level and other relevant climatic information, which are obtained through the climatic environment sensor in accident by local traffic bureau.

(4) POI attribute: describe the traffic signs or similar sign types in the vicinity of the accident. Use GIS to record the traffic signs and traffic characteristics around the accident, such as whether there is an intersection, whether there is a deceleration sign, and so on.

TABLE I. List of some features of the traffic accident data set from US-ACCIDENT [1]

Attribute category	Attribute name
Event attribute	id, source, TMC, severity, start_time, end_time,distance, description

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

Regional attribute	Start_Lat, Start_Lng, End_Lat, End_Lng, number, street, side (left/right), city, county, state, zip-code, country
Weather attribute	time, temperature, wind_chill, humidity,pressure, visibility, wind_direction, wind_speed,precipitation, and condition
POIattribute	Junction, No_Exit,Railway,Roundabout,Station,Stop,Traffic_Calming,Traffic_Signal, Turining_loop

According to the latitude and longitude of regional attributes in TABLE I, the urban distribution map of traffic accidents in California, USA, can be roughly drawn, as shown in Figure 3:

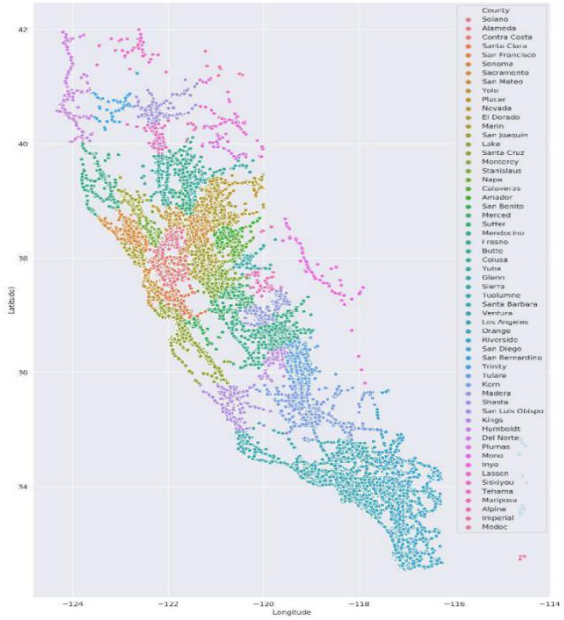

Fig 3: Map of U.S. traffic accident statistics (California)

The traffic data sets including about 47 feature columns and 30,0000 rows which presents the characteristics of the imbalance. Data set not only contains continuous variable data, but also discrete one. Some attribute contains null value or outliers. Moreover, many null values appeared in the data set, which shows heterogeneous and sparse. Therefore, in order to analyze the data and training data, it is necessary to preprocess.

3.2 Correlation Analysis

Before data preprocessing, we can see the degree of correlation between features through correlation analysis. In this we took severity of traffic accident data set (Serverity) as predicted labels, analyze the correlation between other features and target labels. The analysis results are shown in TABLE II.

TABLEII. Analysis of the Correlation Degree of Accident Degree Influence Features

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

Feature column	Correlation value
Severity	1
Start_Lng	0.136048
Junction	0.072473
Temperature(F)	0.031692
Railway	0.009647
Distance(mi)	0.008812
No_Exit	0.008713
Time_Duration(min)	0.007393
Humidity(%)	0.003138
Visibility(mi)	0.002009
Roundabout	-0.005342
Bump	-0.007303
Traffic_Calming	-0.00932
Pressure(in)	-0.010996
Give_Way	-0.013066
Amenity	-0.027269
Station	-0.030544
Stop	-0.100449
Crossing	-0.101934
Start_Lat	-0.16071
Traffic_Signal	-0.169065

We can obviously see the latitude and longitude property formed certain correlation with Severity. This is mainly because the areas of serious traffic accidents are concentrated in southeastern California. In addition, it can be found the severity of some accidents is negatively correlated with the sign attributes of traffic POI. The severity of accidents is higher in areas where stop signs, cross signs, and light signals are missing. In general, the linear correlation between single feature column and label column is low.

3.3 Data Cleaning

Some attributes have little effect on sample learning. So feature vectors can be reduced through feature column fusion. We removed the start time and end time in stead of converting them to time duration column.

Each accident is described as text message and recorded in the MaqQuest attribute value in the data feature. The U.S. Department of Transportation considers the duration and distance of an accident when assessing the severity of an accident. Therefore, formula (8) is adopted to calculate the characteristic parameters of accident reduction. C_{st} represents the time when the accident began, while C_{et} indicates the end time. C_d is the distance affected by the accident, α and β stands for the characteristic constraint factor. The formula (8) can fuse three columns of

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

eigenvalues into one column describe as C_p.

$$C_p = \alpha(C_{st} - C_{et}) + \beta C_d \quad (8)$$

There is a part of abnormal row data in the data set. The row where the abnormal value is marked as noise, then recorded and filtered from the data set to avoid entering the data training. The other part of the feature has a certain number of null values, including continuous feature attributes and discrete feature attributes.First we extract the statistical subset through formula (10). The $C_f$represents the value to be filled. $Grid(C_f)$is that subset which centered by C_f. Position of the element falls in a circle with a radius of 30 km show as R_D, or time lay in the absolute difference less than 72 hours described as R_T. For the discrete feature, use the category with the highest frequency and fill in C_f. For the continuous feature, according to Formula (9), fill in the mean value of the subset. If it is not meet the filling conditions, it should be clean up as outliers.

$$C_f = \frac{1}{m}\sum_{i=1}^{m} C_i C_i \in Grid(C_f) \quad (9)$$

$$Grid(C_f) = \begin{cases} \{C \| C - C_f| < R_D\} \\ or \\ \{C \| C - C_f| < R_T\} \end{cases} \quad (10)$$

3.4 Other Data Attribute Processing

Part of the data (especially POI attribute data) belong to text attribute feature, which cannot be directly entered the training. Therefore, these text features need to be transformed into numerical features. One-hot encoding method is adopted to transform text feature into multicolumn numerical feature codes, so as to prepare for further classification training. To make the data adapt algorithm training, it is necessary to scale the range of data feature. Constrain the range across too large numerical range to a reasonable range, and improve the training effectiveness of the data.

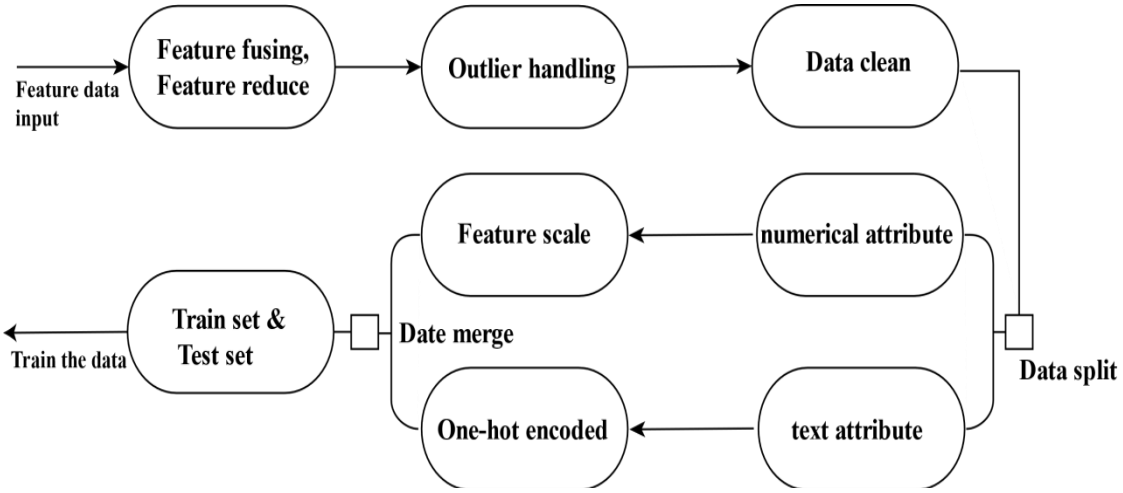

Fig 4: Feature data preprocessing flowchart

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

3.5 Comparative Analysis of Algorithm Experiments

To evaluate the performance of the algorithm, SGD, KNN, decision tree, random forests, XGBoost classification algorithm are selected for comparison in this experiment. The sample data, 80% was selected randomly as training set, and remaining 20% as test sets. The data training model was performed on a server training machine with 4 core 16GB memory and GPU for K80. A 4 fold cross validation is used on the training set, we compare the pros and cons of models through acc score and F_1score, The F_1 calculated expression is shown in Formula (11), which showed that overall performance of the classification model is evaluated at the precision and recall:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2 \times TP}{(TP+FN)+(TP+FP)} \tag{11}$$

The results of cross-validation are shown in TABLE III:

TABLE III. Results for classification algorithms in cross-validation process

Classification model	parameter characteristic	ACC scores				F1 scores			
SGD linear classifier	Max iteration：10000	0.5708	0.6033	0.5972	0.5709	0.4149	0.5982	0.5661	0.4156
KNN classifier	neighbors：6	0.6007	0.5994	0.5998	0.6026	0.5886	0.5882	0.5884	0.5917
Decision Tree classifier	entropy,max_depth：12	0.8783	0.8769	0.8836	0.8774	0.8789	0.8775	0.8841	0.878
	gini,max_depth：12	0.8801	0.8863	0.8899	0.8887	0.8806	0.8867	0.8904	0.8892
Random forest classifier	n_estimator:10	0.8591	0.8496	0.8551	0.857	0.8586	0.849	0.8546	0.8565
	n_estimator:20	0.8739	0.8781	0.8794	0.8809	0.8739	0.8781	0.8794	0.8809
	n_estimator:40	0.8835	0.8849	0.8848	0.8885	0.8837	0.8851	0.885	0.8887
	n_estimator:100	0.8901	0.8905	0.8891	0.8933	0.8904	0.8908	0.8893	0.8935
XGBoost classifier	max_depth：12	0.9374	0.9372	0.9380	0.9373	0.9376	0.9374	0.9382	0.9375

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

It is shown the models of SGD and KNN in the gradient descent linear classification do not perform well on the training set in TABLE III. Further, the data confounding and weak linear correlation are explained. The learning correlation of the adjacent reference data is not significant cause the KNN method to perform poorly. The complexity and diversity of the data reduce the training effect of SGD and KNN. In contrast, tree classification algorithm is widely used in data sets to achieve better training results. The decision tree model performs better than the linear model and KNN at glance. Since the random forest algorithm uses multiple subtrees to estimate, there is a certain relationship between performance of algorithm and the number of estimated subtrees. However, with the increase of the estimated subtrees, the gain of training decreases. Especially when the tree number exceed 200, training time became unbearable, and the score hardly improve.

At the maximum tree depth of 12 training conditions, XGBoost showed excellent training results, which is beyond the decision tree and random forest about 5%. Because of Taylor's second derivative term in loss function calculation, XGBoost may find direct optimization in more data details. Never the less, to introduce regularization term avoids overfitting of data, this make XGBoost training effect significantly better than the rest of the tree classification algorithm.

The trained algorithm model is used to test on the test set, calculation for classifying acc scores as shown in TABLE IV. It can be seen the test results are basically consistent with the model training evaluation conclusions, XGBoost take the best performance in classification precision.

TABLE IV. The parameter setting of each algorithm and the comparison of acc value on the test set

Classification model	parameter characteristic	ACC scores
SGD linear classifier	Max iteration：10000	0.574
KNN classifier	neighbors：6	0.608
Decision Tree classifier	entropy,max_depth：12	0.878
	gini,max_depth：12	0.890
Random forest classifier	n_estimator:10	0.864
	n_estimator:20	0.882

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

	n_estimator:40	0.893
	n_estimator:100	0.895
XGBoost classifier	max_depth：12	0.941

For further comparative study the classification effects of decision trees, random forests, and XGBoost, we use the roc graph to compare the trend precision and recall. Fig. 5 shows the three tree algorithm in the Serverity (class = 'serious') for the accident on the ROC curve, which shows that performance of XGBoost is superior to other two algorithms.

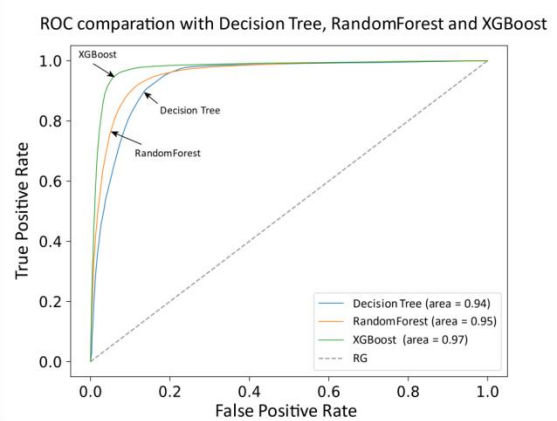

Fig 5: Comparison of ROC curves of tree classification algorithms

IV. CONCLUSION

In this paper, we take the traffic accident data in California from 2016 to 2019 as the study sample which is derived from US-ACCIDENT[1]. Our strategy has been experienced data preprocessing, such as data cleaning, null value fill up, data split, One-hot encoding, feature scaling. Finally we feed it in the several training model and test performance of these models, the following conclusions are drawn.

(1) Compared with SGD and KNN, tree classification algorithm performs better in the classification of large-scale diverse traffic accident data. (2) The classification effect of XGBoost is obviously better than decision tree and random forest algorithm model. It provides a better research strategy for future traffic accident prediction and classification.

Due to the complicated traffic scene[11], traffic data sets are mainly derived from different kinds of accident detectors and sensors apparatus, it means the data components with high complexity, sparsity and heterogeneity. In the process of classification prediction research, data preprocessing should be carried out according to the characteristics of the data, data training should be preferred to selection algorithm with excellent performance in balance like XGBoost. This study provides an important research means for public traffic safety, traffic accident prediction and traffic accident readiness. In addition, the characteristics of large-scale traffic

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

accident data still need to be further explored in the accuracy of classification.

ACKNOWLEDGEMENT

This research was supported by the 2018 basic competence improvement project of young and middle-aged teachers in Guangxi universities, "research on key technologies of urban traffic collaborative scheduling based on massive image big data" (No. 2018KY1272), phased research result.

REFERENCES

[1] Moosavi S, Samavatian M H, Parthasarathy S, et al. (2019) A Countrywide Traffic Accident Dataset

[2] Wang xiao fan, Zhu yong qiang (2019) Prediction of Road Traffic Accidents Based on Grey BP Neural Network Models. Journal of Baicheng Normal University 33(06): 36-40+51

[3] ZHANG Zhihao, YANG Wenzhong, YUAN Tingting, et al (2019) Traffic accident prediction based on LSTM neural network model. Computer Engineering and Applications 55(14): 249-253

[4] ZHU Boya, FU Xinsha, YANG Siqi, ZHU Zhenjie (2018)Prediction Model of Space-time Impact of Traffic Accidents Based on Nonlinear Regression and BP Neural Network, Highway Engineering 43(6):134-139

[5] LU yong, YAO shi wei, et al (2018) Short - time forecast of freeway traffic accidentHoliday traffic safety characteristics analysis. HIGHWAY 2018(11):224-227

[6] ZHAO Yue-feng, ZHANG Sheng-rui, MA Zhuang-lin (2018) Analysis of Traffic Accident Serverity on Highway tunnels Using the Partial Proportion Odds Model, China Journal of Highway and Transport 31(09):159-166

[7] ZHANG Jun,HU Zhenbo, ZHU Xinshan (2017) Real-time traffic accident prediction based on AdaBoost classifier. Journal of Computer Applications 37(1): 284-288

[8] HE Ke, YANG Shunxin, GAO Yonggang (2019) Prediction of Traffic Incident Duration in Tunnels Based on a PCA-RF Combined Model, Journal of Transport Information and Safety (5):26-32

[9] Chen T, Guestrin C. (2016) XGBoost: A Scalable Tree Boosting System

[10] Quinlan J R. (1986) Induction of decision trees. Machine Learning1(1):81-106

[11] LIU jie (2016) A kind of adaptive object tracking strategy base on multi-feature integration. Popular Science & Technology(11):9-12

[12] Elith J, Leathwick J R, Hastie T, et al. (2008) A working guide to boosted regression trees. Journal of Animal Ecology77(4): 802-813

ISSN: 0011-9342
Issue: 11 | Pages: 572 - 584

[13] Kabir F, Siddique S, Kotwal M R, et al. (2015) Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier. International Conference on Cognitive Computing and Information Processing 1-4

[14] Liao Y, Vemuri V R. (2002) Use of K-Nearest Neighbor classifier for intrusion detection. Computers & Security 21(5):439-448

[15] Chen T, He T, Benesty M. (2016) xgboost: Extreme Gradient Boosting

[16] Breiman L. Random forest. (2001) Machine Learning 45:5-32

[17] Liaw, A, Wiener, M, Liaw, A. (2002) Classification and Regression with Random Forest. R News23(23)

[18] Alaya M Z, Bussy S, Stéphane Gaïffas, et al. (2017) Binarsity: a penalization for one-hot encoded features. Journal of Machine Learning Research20:1-34

[19] Abdel-Aty M A, Radwan A E. (2000) Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention32(5):633-642

[20] Ki Y, Lee D. (2007) A Traffic Accident Recording and Reporting Model at Intersections. IEEE Transactions on Intelligent Transportation Systems8(2):188-194

[21] Evans L. (2010) Risk Homeostasis Theory and Traffic Accident Data. Risk Analysis6

[22] Bayam E, Liebowitz J, Agresti W. (2005) Older drivers and accidents: A meta analysis and data mining application on traffic accident data. Expert Systems with Applications29(3):598-629