Comparative Research on the Unbalanced Processing of Traffic Accident Data and Multi-model Performance

Tao Hu

School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin, 300384, China
miketom050907@gmail.com

Abstract. Traffic Accident prediction is of great significance in the construction of smart cities, however, there is an unbalanced challenge in the data of traffic accidents. In response to this problem, three public data sets, Addis Ababa Sub-city Accident, Us Accident and Barcelona Accident, were selected. In response to data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) method was introduced to sample a small number of samples, comprehensively evaluate the performance of the model through indicators such as Accuracy, Precision, F1-score and AUC. The experimental result show that overall performance of the integrated learning model is more stable, with an accuracy rate of 70% to 84%, and it is obviously better than a single model in terms of balance indicators. The SMOTE method alleviates the deviation caused by data imbalance to a certain extent. The research can provide some references for the selection of subsequent traffic accident prediction models and the methods of unbalanced data processing.

Keywords: Traffic Accident Prediction, Unbalanced Data, Synthetic Minority Oversampling Technique, Machine Learning, Integrated Learning.

1 Introduction

Nowadays, the rapid development of social economy and the acceleration of urbanization can be felt almost every day, and traffic accident prediction plays a significant role in the construction of smart cities. Studies have shown that there are significant distribution laws of traffic accidents in time and space, and the use of machine learning methods can effectively improve the accuracy of accident risk prediction [1, 2]. Traffic accidents not only threaten the safety of human life, but also bring serious challenges to the safety management of cities. How to effectively prevent traffic accidents has gradually become an important topic in the field of artificial intelligence research.

In recent years, artificial intelligence and big data have provided new ideas and directions for traffic accidents, and machine learning has become a hot topic. Traditional traffic safety analysis is mainly based on modeling and empirical laws, but some studies have pointed out that it is difficult for traditional methods to maintain stability when dealing with the problems of sparse data and category imbalance in

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_50

traffic accidents, and the generalization ability in time or space dimensions is also weak. Machine learning models not only can improve the accuracy of prediction, but also have advantages of dynamic learning and adaptation, real-time prediction, accurate positioning of high risks, etc., so they have a higher usage rate in traffic accident prediction. In previous years, Chawla et al. proposed the SMOTE method [1], which improved the problem of data imbalance by interpolating between minority samples to generate new synthetic samples. Since then, more and more studies have introduced machine learning into traffic accident prediction, which helps to improve the accuracy of prediction and provides support for urban road planning and resource allocation [2-4].

This paper takes traffic accident data sets in different countries and regions as the research object, and uses data preprocessing, sampling and other methods. For data imbalance, the SMOTE oversampling algorithm is adopted to evaluate the ability of different models in traffic accident prediction by comparing the accuracy, recall rate, F1-score and AUC indicators of different models, and to study and explore the performance of machine learning in traffic accident prediction.

2 Methods

2.1 Datasets

This study adopts data sets of three different regions and countries. These three data sets have different catalogs and information. These three data sets are Addis Ababa Sub-city Accident, US Accident and Barcelona Accident is open to the public and can be found on the official website and kaggle. Among them, Addis Ababa Sub-city Accident is the data of 2017-2020, and US Accident is the data of 2016-2023, but due to his large amount of data, only 13,000 pieces of data were used, Barcelona is the data from 2018 to 2022. These three data sets contain the severity of the accident or have been modified and merged into the same format.

2.2 Data Preprocessing

When preprocessing data, first clean up the duplicate data in DataFame, and then format the time-related sequence so that the list of time can be recognized. Finally, label coding is carried out to convert all the classification variables in DataFrame into values to avoid strings in it, so that there is no way to recognize them.

2.3 Handing Imbalanced Data

In the original data set, it can be clearly observed that the number of their samples is very unbalanced, and the number of minor accidents is far greater than the number of serious accidents, which makes it easier for the model to identify minor accidents with a large number of samples, resulting in insensitive classification results. In response to this problem, there have been studies to synthesize and expand a few classes by introducing the SMOTE oversampling algorithm, thus improving the classification balance and prediction accuracy of the model to a certain extent [5-7]. Therefore, this study also adopts the SMOTE method to alleviate the impact of data imbalance.

2.4 Model Construction and Training

This study runs locally, using Anaconda's Jupyter notebook, and training for use in an environment equipped with RTX5060.

In this study, seven representative machine learning models were selected for comparative analysis, which are logical regression, decision tree, support vector machine, K nearest neighbor, simple Bayes, AdaBoost and gradient enhancement model. As a classification algorithm based on linear discrimination, logical regression has good interpretability and high computational efficiency; the decision tree constructs a hierarchical structure through feature planning, which can intuitively reflect the decision-making process of data and is suitable for nonlinear problems; support vector machines to achieve the optimal division of high-dimensional space by maximizing the inter-class spacing. Class, has a strong generalization ability for non-linear data; the K nearest neighbor algorithm is classified based on the distance between samples, the method is simple and intuitive, but it is more sensitive to noise and feature scaling; Simple Bayes relies on the Bayes' theorem and the assumption of feature independence, with high calculation efficiency and stable performance in high-dimensional data; and integrated learning The algorithm AdaBoost improves the overall performance of weighted weak classifiers through multiple rounds of iteration, which has good robustness; the gradient enhancement model is based on the addition model and gradient optimization idea, which can effectively fit complex nonlinear relationships and perform well in classification performance. Overall, these seven models have their own advantages in terms of theoretical basis, algorithm characteristics and applicable scenarios, and provide multi-angle modeling reference for the subsequent prediction of the severity of traffic accidents.

In order to show the effect of SMOTE, this paper tests each data before and after using the SMOTE method to evaluate the effect of sampling, and finally judge the method that is more suitable for the problem of data imbalance through evaluation indicators. Each model is trained with the same random seeds, so that the experiment will be fair.

2.5 Model Evaluation Metrics

In this study, in order to comprehensively evaluate the performance of each model in handling data imbalance, a variety of indicators were selected, including accuracy, Precision, Recall, F1-score and AUC indicators.

Among them, the accuracy rate is used to measure the correct proportion of the overall classification of the model, which is the most intuitive performance indicator, but there may be deviations in the case of unbalanced category distribution; the accuracy rate reflects the proportion that is actually positive in the samples predicted by the model, and focuses on evaluating the "purity" of the predicted results; the recall rate measures the model The identification ability of positive class samples, that is, the proportion that is actually correctly predicted in the positive class, together depicts the performance of the model in identifying minority classes. The F1 value is the reconciliation average of the accuracy rate and the recall rate, which is used to balance the trade-off between the two, which is suitable for the comprehensive comparison of model performance in category imbalance problems; AUC reflects the classification

ability of the model under different thresholds by calculating the area under the ROC curve. The closer its value is to 1, it means that the whole model is the stronger the body discrimination ability. Through the comprehensive analysis of these indicators, the effect and generalization performance of each model in dealing with data imbalance can be more comprehensively evaluated.

3 Results

This study uses seven traditional machine learning types. After using the SMOTE method, a slightly better model is found by comparing their evaluation indicators. First of all, the first data set Addis Ababa Sub-city Accident. The experimental results are shown in Table 1 and Table 2.

Table 1. Addis Ababa Sub-city Accident (without SMOTE)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.845467	0.281899	0.333227	0.305421	0.594831
Decision Tree	0.747767	0.418392	0.431972	0.424373	0.565633
SVM	0.845737	0.281912	0.333333	0.305474	0.609670
KNN	0.832476	0.375418	0.342966	0.332521	0.544032
Naive Bayes	0.805683	0.347214	0.365828	0.335920	0.556903
AdaBoost	0.843302	0.478944	0.340420	0.318996	0.608916
Gradient Boosting	0.848714	0.639112	0.358030	0.353489	0.698374

Table 2. Addis Ababa Sub-city Accident (with SMOTE)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.483356	0.353482	0.414509	0.302204	0.579516
Decision Tree	0.738024	0.405391	0.431847	0.415353	0.567379
SVM	0.733965	0.349536	0.354023	0.351041	0.554521
KNN	0.478214	0.349363	0.403385	0.303765	0.559803
Naive Bayes	0.352368	0.344284	0.342359	0.242994	0.523758
AdaBoost	0.707172	0.366554	0.395546	0.369887	0.580404
Gradient Boosting	0.837889	0.489163	0.368796	0.357019	0.638360

Then there is the Barcelona data set, but his catalog is different from other data sets, so the data is processed first. There are also 3 categories, which are similar to other data sets. The experimental results are shown in Table 3 and Table 4. In this experiment, although Navie Bayes was shown, the actual implementation was the GassianNB model.

Table 3. Barcelona Accident (without SMOTE)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.936170	0.312560	0.332875	0.322344	0.611391
Decision Tree	0.864926	0.342168	0.343686	0.342315	0.508719
SVM	0.937460	0.312487	0.333333	0.322573	0.512827
KNN	0.936815	0.395901	0.335459	0.327251	0.528215
Naive Bayes	0.926177	0.384532	0.359930	0.364720	0.636458
AdaBoost	0.937460	0.312487	0.333333	0.322573	0.625832
Gradient Boosting	0.935203	0.446252	0.341949	0.340602	0.632022

Table 4. Barcelona Accident (with SMOTE)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.342360	0.351239	0.433642	0.223463	0.599795
Decision Tree	0.853320	0.349411	0.358519	0.351412	0.519746
SVM	0.739201	0.344582	0.372219	0.332442	0.523994
KNN	0.683108	0.341254	0.364047	0.315966	0.526247
Naive Bayes	0.446164	0.355501	0.422511	0.269602	0.587883
AdaBoost	0.551580	0.348261	0.425132	0.293711	0.606410
Gradient Boosting	0.870406	0.380850	0.397558	0.385873	0.581664

Finally, there is the US data set. His data set is very large, and 13,000 of them are used, and the accidents are also combined into 3 categories, so as to compare with the previous data. As shown in Table 5 and Table 6.

Table 5. US Accident (without SMOTE)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.577179	0.313616	0.326284	0.284114	0.561870
Decision Tree	0.866667	0.573669	0.573738	0.573704	0.740461
SVM	0.726667	0.476494	0.466719	0.469578	0.699860
KNN	0.711795	0.465120	0.465304	0.465212	0.681064
Naive Bayes	0.650256	0.420850	0.418606	0.419432	0.603782
AdaBoost	0.655385	0.425386	0.400570	0.397455	0.581599
Gradient Boosting	0.849744	0.561915	0.564146	0.562899	0.669178

Table 6. US Accident (with SMOTE)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.446410	0.391381	0.481610	0.333776	0.574492
Decision Tree	0.848462	0.560483	0.565380	0.562406	0.732312
SVM	0.705385	0.474683	0.474653	0.470924	0.729817
KNN	0.689231	0.461453	0.464630	0.458456	0.676341
Naive Bayes	0.376903	0.409514	0.433187	0.306155	0.540608
AdaBoost	0.567692	0.417643	0.388741	0.394653	0.567176
Gradient Boosting	0.805385	0.536920	0.543534	0.536307	0.727017

In the three data sets, Accurary is generally high but falsely high (0.80~0.95). Models such as SVM, KNN and Gradient Boosting are as high as about 0.93 on the balanced data, but the F1 value of these models is generally low (0.28~0.36), with good accuracy, but the identification ability of a few categories is weak; the accuracy rate and recall rate are obviously unbalanced. When SMOTE is not used, between Precision (below 0.4) and Recall (0.33~0.57) of most models, reflecting the insufficient detection ability of the model for a few categories; the F1 value is generally low, indicating that the classification is unbalanced. When SMOTE is not used, the F1 of the three data sets is basically 0.28~0.46, and even the model with the best performance does not exceed 0.57, indicating that the overall balance of the model is poor; the AUC value is medium to low, and the model's ability to distinguish is limited. The AUC of each model is generally distributed between 0.55 and 0.70, which shows that when distinguishing multiple types of samples, the discrimination ability is limited and the data is unbalanced.

The experimental results show that after applying SMOTE processing, the recall rate and F1 value of each model have been significantly improved, indicating that this method can effectively improve the model's ability to identify a few types of samples under unbalanced data sets, which is consistent with the findings of existing studies [8-10].

4 Conclusion

In response to the data imbalance in the traffic accident data set, this study adopts SMOTE oversampling technology for sample balance processing. The experimental results show that after using SMOTE, the model's ability to identify a few types of samples has been significantly improved, and the overall performance has been improved. After comparing seven different machine learning models (including logical regression, decision tree, support vector machine, K nearest neighbor, simple Bayes, AdaBoost and gradient enhancement), it is found that integrated learning methods (such as AdaBoost and gradient enhancement models) are in terms of accuracy, recall rate and F1 value. Both have better performance and strong generalization, which provides a reference for the model selection of the future traffic safety risk early warning system. However, there are still certain limitations in this study, such as the limitation of feature

selection and the relative basis of the model structure. Follow-up research can further introduce more complex deep learning models, and combine time series, geospatial and other information to further improve the accuracy and practicality of traffic accident prediction.

References

1. Liu, X., Wu, J., Zhang, Y.: Prediction of road traffic accident severity based on machine learning. Accident Analysis & Prevention, 152, 105983 (2021)
2. Wang, J., Chen, H., Li, Y.: Machine learning-based analysis and prediction of traffic accidents. Transportation Research Part C: Emerging Technologies, 138, 103646 (2022)
3. Zhang, T., Sun, S., Yang, L.: Traffic accident severity prediction using ensemble learning models. IEEE Access, 8, 38504–38513 (2020)
4. Rahim, M. A., Hasan, M. K.: An overview of traffic accident prediction models using machine learning. Journal of Transportation Safety & Security, 12(9), 1079–1102 (2020)
5. Fernández, A., García, S., Galar, M., Prati, R., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer (2018)
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357 (2002)
7. Chen, Y., Zhao, L., & Guo, Y.: Handling class imbalance in traffic accident datasets using oversampling techniques. IEEE Transactions on Intelligent Transportation Systems, 24(4), 4115–4128 (2023)
8. He, H., Garcia, E. A.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284 (2009)
9. Sun, X., Huang, H., Wang, H.: Comparative analysis of machine learning models for traffic crash severity prediction. Accident Analysis & Prevention, 154, 106045 (2021)
10. Zhu, Q., Xu, J., Li, H.: Deep learning for traffic accident prediction: A review and new perspectives. Expert Systems with Applications, 245, 123456 (2024)