Time Series Classification using Imbalanced Learning for Real-time Safety Assessment

Christos Katrakazas[1],Constantinos Antoniou[1], George Yannis[2]

[1]Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Munich, Germany

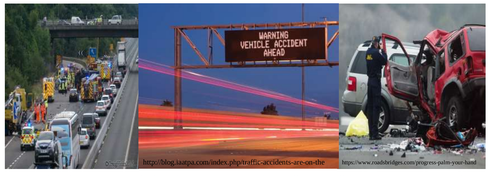
[2]Department of Transportation Planning and Engineering, National Technical University of Athens, Greece

Paper No.: 19-04457

ABSTRACT

The probability of estimating a **traffic collision** happening in real-time primarily depends on comparing traffic conditions just before a collision with traffic conditions during normal operations. Most studies however utilize aggregated traffic data and are not concerned with the dynamic nature of collisions or the **imbalance of safety databases** which can lead to erroneous real-time predictions. In this study, this is overcome through the use of **raw speed time series data of variant duration** (1-minute to 5-minute time series data) from a **driving simulator experiment** and the use of **imbalanced learning** techniques. Two classifiers are then employed to examine the proposed idea: (i) **Random Forests (RFs)** – an ensemble classifier and (ii) **Neural Networks (NNs)** – a popular classifier in the literature. These classifiers are tested on the original time series data, as well as on time-series treated with the imbalanced learning techniques of **undersampling and its integration with oversampling**. The main results demonstrate the viability of using raw speed time series data for real-time safety assessment and the **superiority of time series with 4-minute duration** in the classification results. Furthermore, RFs perform well even in 1-minute time series data while the **classification results can be enhanced by up to 40%** from imbalanced learning approaches. It is also demonstrated that the classification results outperform similar approaches in the literature. However, real-world traffic data and the use of more sophisticated classifiers are expected to provide more effective predictions.

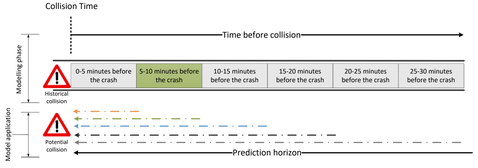
INTRODUCTION



A dominating approach for detecting unsafe traffic conditions is the comparison of traffic situations just prior to traffic collision occurrences on a segment with the traffic conditions at normal situations on the same segment.

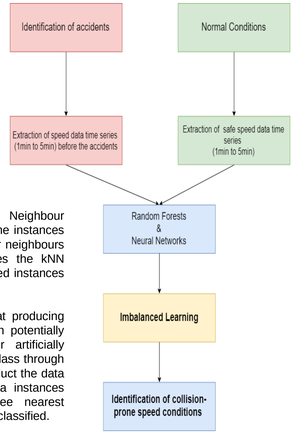
Three major limitations of real-time collision prediction models are:

1. the imbalance of the datasets because safe traffic condition cases are over-illustrated against collision-prone conditions due to the rarity of collision events.
2. The absence of approaches considering the dynamic nature of collision occurrences
3. The use of aggregated traffic data which affects the prediction horizon



METHODOLOGY

The limitations mentioned are overcome in this paper by considering highly disaggregated time-series data and utilizing two imbalanced learning techniques; Repeated Edited Nearest Neighbours (RENN) and Synthetic Minority Oversampling Technique (SMOTE) integrated with Edited Nearest Neighbours (ENN).



RENN utilizes the Edited Nearest Neighbour (ENN) algorithm repeatedly until all the instances in the dataset have a majority of their neighbours within the same class. ENN applies the kNN algorithm and removes all misclassified instances from the training dataset.

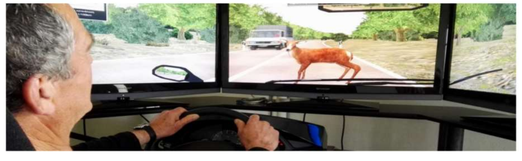
SMOTE integrated with ENN aims at producing well-defined class clusters which can potentially improve classification results. After artificially generating instances of the minority class through SMOTE, ENN is implemented to conduct the data cleaning in depth and removes data instances from both classes when the three nearest neighbours of a data instance are misclassified.

DATA COLLECTION & PROCESSING

Data utilized in this study were collected using a FOERST 32 Driving Simulator consisting of 3 LCD wide screens (40'' Full HD), a driving position and a support motion base. The simulator is based at the Department of Transportation Planning and Engineering of NTUA in Athens, Greece.

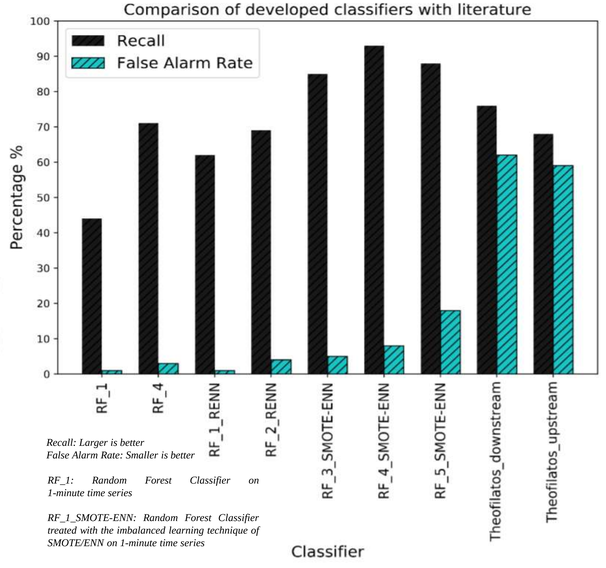
- 2,1 km long rural single carriage way
  - 3m lane width
  - 41 horizontal curves
- 279 driving sessions
  - 169 collision events

Time-Series Length (minutes)	Collision-prone speed conditions:Normal Conditions ratio
1	1:13
2	1:6
3	1:4
4	1:3
5	1:2



CLASSIFICATION RESULTS

Figure 1: Comparison of the best developed classifiers with the results of Theofilatos et al. (2018)

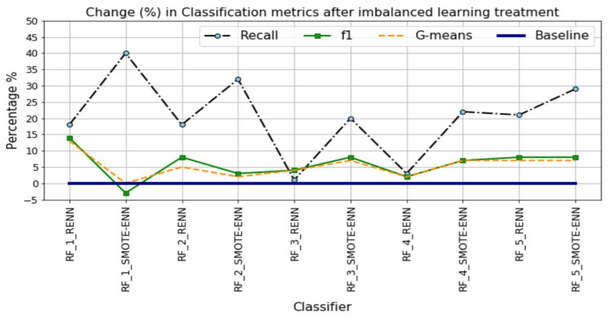


Recall: Larger is better
False Alarm Rate: Smaller is better

RF_1: Random Forest Classifier on 1-minute time series

RF_1_SMOTE-ENN: Random Forest Classifier treated with the imbalanced learning technique of SMOTE/ENN on 1-minute time series

Figure 2: Percentage change in classification metrics when compared with the untreated classification results



- RFs generally outperform NNs
- The best results are indicated for the 3- and 4-minute time series data
- The integration of oversampling with undersampling (SMOTE-ENN) results in better classification performance, than only undersampling the majority class (RENN).
- NNs perform better in terms of identifying correctly collision-prone conditions in shorter time series (i.e. consisting of 1-, 2- or 3-minute measurements), while RFs perform better when the duration of time series increases.
- Looking at the overall performance of the classifiers, as depicted in the f1-score and G-means, it is understood that the best results are obtained with RFs and SMOTE-ENN using a 3-minute f1=82.76% G-means 82.79%) or a 4-minute time series (f1=86%, G-means=86.3%), guaranteeing correct identification of both collision-prone speed time series, as well as safe conditions.
- Even without imbalanced treatment, the original 4-minute series outperforms the majority of the developed classifiers.
- RFs when combined with undersampling are able to identify almost 70% of collision-prone speed conditions with a very small false alarm rate even when 1-minute or 2-minute speed data are used.

CONCLUSIONS

- This paper proposes the classification of raw speed time series data before collision events using imbalanced learning. The approach intents to overcome two problems: i) the use of aggregated data, as raw time-series data are utilized; and ii) the imbalance of safety databases through the use of imbalanced learning
- The data used were obtained through a driving simulator experiment in Athens, which took place in order to assess the driving performance of drivers with cognitive impairment.
- Five speed time series of variant duration (spanning from 1-minute to 5-minutes) were constructed in order to test the effect of duration in the classification results.
- Regarding imbalanced learning, two techniques were utilized, namely undersampling of the majority class (i.e. safe traffic conditions) and oversampling of the minority class (i.e. collision-prone traffic) integrated with undersampling.
- It was shown that RFs in general lead to better classification results when compared to NNs, and the treatment with imbalanced learning can enhance results up to 40% even when 1-minute time series are utilized for real-time classifications

KEY REFERENCES

1. Abdel-Aty, M., and A. Pande. The viability of real-time prediction and prevention of traffic accidents. In *Efficient Transportation and Pavement Systems* (Al-Qadi, T. Sayed, Alnuaimi, and Massad, eds.), Taylor & Francis Group, London, pp. 215–226.
2. Xu, C., P. Liu, and W. Wang. Evaluation of the predictability of real-time crash risk models. *Accident Analysis and Prevention*, Vol. 94, 2016, pp. 207–215.
3. He, H., and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, 2009, pp. 1263–1284.
4. Katrakazas, C., M. Quddus, and W. H. Chen. A Simulation Study of Predicting Real-Time Conflict-Prone Traffic Conditions. *IEEE Transactions on Intelligent Transportation Systems*, 2017, pp. 1–12.
5. Theofilatos, A., G. Yannis, C. Antoniou, A. Chaziris, and D. Sermpis. Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach. *Journal of Transportation Safety and Security*, Vol. 9962, No. March, 2017, pp. 1–20.
6. Breiman, L. Random Forests. *Machine learning*, Vol. 45.1, 2001, pp. 5–32.
7. Christodoulou, C., and M. Georgiopoulos. *Applications of Neural Networks in Electromagnetics*. Artech House, Inc., Norwood, MA, USA, 2001.
8. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357.
9. Wilson, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 2, No. 3, 1972, pp. 408–421.
10. Yannis, G., J. Golias, C. Antoniou, S. Vardaki, P. Papantoniou, D. Pavlou, S. Espié, G. Kalisperakis, S. G. Papageorgiou, G. Tsivgoulis, A. Bonakis, N. Andronas, I. Papatriantafyllou, A. Liozidou, D. Kontaxopoulou, A. Economou, and M. Kosmidis. *Distract: Causes and Impacts of Driver Distraction: A Driving Simulation Study, Deliverable 4: Driving Simulator Experiment*. 2014.

ACKNOWLEDGEMENTS

The research was funded by the EU H2020 NOESIS Project (Project Number: 769980).