Machine Learning process for injury severity prediction and Traffic accidents classification

Zheng Xiang, Chang Li, Lee Chen, Bing Pan, Don Chen, Lixuan Zhang
Faculty of Computer Science and Information System, Universiti Teknologi MARA (UiTM), Malaysia

ABSTRACT

Traffic accidents constitutes the first cause of death and injury in many developed countries. However, traffic accidents information and data provided by public organisms can be exploited to classify these accidents according to their type and severity, and consequently try to build predictive model. Detecting and identifying injury severity in traffic accidents in real time is primordial for speeding post-accidents protocols as well as developing general road safety policies. This article presents a case study of traffic accidents classification and severity prediction in Spain. Raw data are from Spanish traffic agency covering a period of six years ranging from 2011 to 2015. To this end, are compared three different machine learning classification techniques, such as Gradient Boosting Trees, Deep Learning and Naïve Bayes.

KEYWORDS: Data Mining, Deep Learning, Gradient Boosting Trees, Naive Bayes, Machine Learning, Data fusion, Traffic Accident, Emergency Management, Open Data

1.0 INTRODUCTION

Engineers and researchers in the automobile industry have tried to design and build safer automobiles, but traffic accidents are unavoidable. The cost of deaths and injuries due to traffic accidents have a great impact on society [1-5]. Traffic accidents and their severity are the result y several factors ranging from driver behavior, roads characteristics, vehicles types, to weather conditions, to cite few of them. At present and unfortunately, traffic accidents are one of the most life-threatening dangers to human being. According to the Spanish traffic agency (hereafter DGT) [6-10], are reported 102,362 accidents with victims during 2016. These accidents caused 1,810 deaths during the accidents or up to 30 days after, 9,755 people were hospitalized, and 130,635 were injured with no hospitalization. These figures show an increase of 7% in deaths with respect to 2015. Many strategies can be deployed to reduce deaths during traffic accidents, and one of them, consists in speeding post- accidents attention. In this sense, predicting accidents severity could be a key for quick response to such accidents [11-17]. In this sense, data mining approach combined with other techniques for knowledge discovery is an encouraging way for understanding, classifying and even for predicting injury severity. Indeed, recently attention have been paid for determining factors that significantly affect the severity of traffic accidents and several approaches have been used to study this problem. For example, traffic accidents analysis based on machine learning paradigm in presented in, and the same authors presented the same study using decision trees and neural networks. Traffic accidents classification using adaptive regression trees is presented [18-26]. A study linking traffic accidents with the distance with respect to the zones of residence related to Lothian region in Scotland is presented [27-31]. Traffic accidents applying data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity, body injury and property damage, in Korea is presented. In study is presented a study related to driver injury severity in traffic accidents at signalized intersections in central Florida area (US). A study linking speed limit increase with fatal crash rate and deaths on freeways in Washington State is presented in. The work presents a statistical analysis of accident severity on rural freeways based on nested logic formulation as a means for determining accident severity given that an accident has occurred. In project is applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers' fatality risk, where the main founding is that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities. Finally, a study of the relationship between drivers' age, gender, vehicle mass and driving speed can be found in [31-40]. This article presents a case study of traffic accidents classification and severity prediction in Spain. Data used are from Spanish traffic agency and they cover a period ranging from

2011 to 2015. In this study are compared three different machine learning classification techniques, namely gradient boosting trees, deep learning and naive Bayes. The remaining of the article is organized as follow: Section 2 presents the structure of data-mining process and describes the datasets used, its sources and its most important features. Section 3 describes more details about the problem and the pre- processing of data to be used are presented, followed; in Section 4 by a short description the different machine learning paradigm used and performance analysis and Section 5 compiles the conclusions and presents some future work for improving the proposed study.

2.0 DATAMINING PROCESS AND DATASET

The development of the classification model to predict the severity of traffic accidents is based on three stages and processes. The first stage comprises the process of obtaining the data needed to feed the classification model. The second stage focuses on the process of feature engineering for generating structured data. Finally, the third stage corresponds to the development of the predictive models using the different algorithms for obtaining patterns and knowledge. The hierarchical scheme of these three processes is summed up in "Fig. 1".

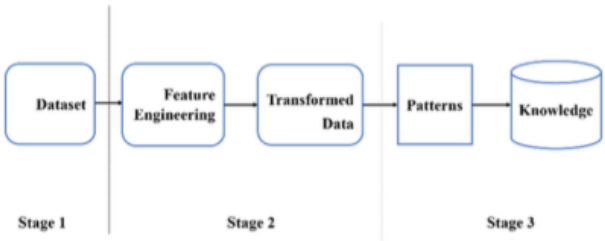

Figure 1. Data mining system for a classification model to predict the severity of traffic accidents

Data used in this study are from the DGT, a public organism for, through its open web site [2]. The initial datasets for the study contain microdata. As an important characteristic of the dataset, it is worth mentioning, that from the year 2011, in the DGT statistical portal, is included the variables that correspond to the determination of the traffic injuries of the people involved in the accident during the 24 hours and during the following 30 days to the accident. These datasets are structured in three related files:

Accidents Dataset. A file containing annual data of traffic accidents, structured on 38 attributes, such as luminosity, road type, weather conditions, accident type, day of the week, region, total victims, total injuries, etc. This file contains more than 83,000 records for 2015 year.

• Vehicles Dataset. A file containing data of vehicles involved in traffic accidents. This file is based on 10 attributes, such as vehicle type, vehicle state, number of occupants, year of registration, vehicle code, code of the accident, etc. During 2015, more than 163,000 records were stored.

• People Dataset. A file containing data about people involved in traffic accidents, and it considers if they are pedestrians, drivers or occupants. This file is structured on 27 attributes, where some of them are: age, gender, serious injury 24 hours, serious injury 30 days, slight injury 24 hours, slight injury 30 days, dead 24 hours, dead 30 days, speed violation, pedestrian infraction, passenger code, pedestrian code, driver code, code vehicle, etc. For example, they were registered more than 219,000 records for 2015 year.

The microdata of accidents, vehicles and people dataset, where refers to the year of occurrence of traffic accidents, are related to each other of the following way:

• Accidents with Victims and Vehicles: 1 to n ratio, through the variable ID_ ACCIDENT.

• Vehicles and People: relationship 1 to n, through 2 variables, ID_ACCIDENT e ID_VEHICLE.

• Accidents with Victims and People: 1 to n ratio, through the variable ID_ACCIDENT.

To unify the mentioned datasets, data for each year have been downloaded separately and then merged in a unique file. These new attributes being the following:

• People Dataset: [id_person_norm {id_driver; id_passenger; id_ pedestrian}; id_vehicle_norm {id_accident; id_vehicle}]

• V ehicles Dataset: [id_vehicle_norm {id_accident; id_vehicle}]

These relationships are based on which for each record of the accidents dataset there may be one or more vehicles involved. Also for each vehicle involved there may be one or more occupants of the vehicle. The result of this unification is a unique dataset of microdata containing 58 attributes and 1,018,204 records.

3.0 FEATURE ENGINEERING

The construction of the final dataset has considered the elimination of useless attributes, irrelevant for the research and the introduction of new attributes necessary for the objective. On the new dataset that contains 58 attributes and 1,018,204 records, two new aggregated attributes are generated that will be added to the dataset. These new attributes are:

• Age of the vehicle. It is an integer attribute. Is the result of the difference between the year of the accident and the year of registration of the vehicle.

• Severity Accident. It is a binary attribute. Value 0 means that it is not a serious accident and the aggregate data is obtained if the values of the attribute SERIOUS INJURY 30D and attribute DEAD 30D are equal to 0. Value 1 means that it is a serious accident and the aggregate data is obtained if the value of the attribute SERIOUS INJURY 30D field is equal to 1 or if the value of the attribute DEAD 30D is equal to 1 or if both attributes are equal to 1.

The new generated attribute “Severity accident”, will be the category column that will be used to predict if an accident is serious or not. The criterion of success for this datamining investigation is the discovery of classification rules for the severity of accidents that would differentiate accidents that are serious from those that are potentially not serious. Before applying the machine learning algorithms on the created dataset, the target will be prepared using the following techniques associated with the data mining, as detailed below.

A. Dealing with unbalanced target

After performing an analysis of the majority class "Severity Accident", it is observed that the attribute is unbalanced, as it appears in " Fig. 2 ". 1,000,122 records are obtained for the value 0 of the attribute (no serious accident), compared to 18,082 records for the value 1 of the same attribute (serious accident).

Most Machine Learning algorithms do not perform well when working with highly unbalanced examples. To improve the numerical proportion of unbalanced data, a subsampling [16] technique is applied to the dataset. This technique involves obtaining a smaller amount of data from the majority class, without modifying the number of elements of the minority class.

A random sampling is carried out with 18,000 records to balance the data presented previously and prepare the dataset to select the most appropriate attributes to generate the machine learning model.

B. Attributes selection

This stage focuses on the quality of the data, specifically on the quality of the attributes of the dataset. Only the attributes that are selected will be used as input for the machine learning model. The attributes that provide the lowest value, that is, the useless attributes, are discarded, using the following quality

indicators for each attribute of the dataset:

• Bad quality attributes. These attributes must be eliminated from the dataset, applying the following rules: More than 70% of all values in these attributes are missing; the attributes are identifiers, where for each row of the dataset they generate a different value; the attribute is constant with more than 90% of all values are equal.

• Attribute that has a very low or very high correlation with respect to the categorical attribute used to predict whether an accident is serious or not: Low correlation, a correlation of less than 0.01%; High correlation, a

correlation of more than 50%. Generally, attributes with high correlation are preferred, but not if the high correlation occurs due to a direct cause-and-effect relationship with the data you want to predict.

Applying these rules to the dataset, 34 attributes are extracted that will be used for the predictive model. The new dataset contains the following attributes: year of circulation permit, position, use of the safety belt, helmet use, shunting, infringement speed, infringement opening, infringement summary, infringement lighting, age, sex, year, year of vehicle registration, month of vehicle registration, vehicle type, anomaly none, pneumatic anomaly, tire rebound anomaly, direction anomaly, brakes anomaly, number of occupants of the vehicle, zone, grouped area, restricted visibility, time, day of the week, autonomous community, municipality, road, road network, type via, type intersection, vehicle's age.

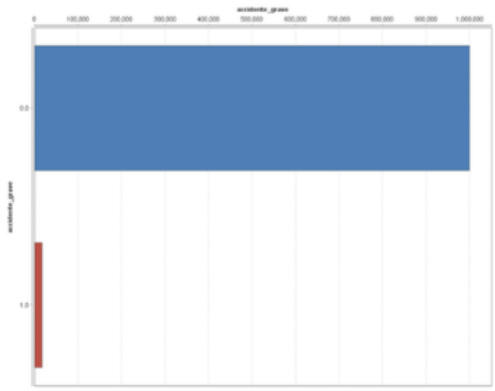

Figure 2. Severity Accident Attribute

4.0 RESULT

To build the classifier developed in this article, three different machine learning classification techniques have been used, such as those that will be presented in this section: Naive Bayes, Gradient Boosting Trees and Deep Learning. Naive Bayes algorithm is one of the most used classifiers for its simplicity and speed in the construction of a probabilistic model, based on the Bayes Theorem [17], also known as a conditional probability theorem. It is a supervised classification and prediction technique that constructs models that predict the probability of possible results. Naive Bayes is a high-bias, low- variance classifier, and it can build a good model even with a small dataset. The fundamental assumption of Naive Bayes is that, given the value of the label (the class), the value of any attribute is independent of the value of any other Attribute. If the Bayes Theorem is used in a machine learning problem, it is because a posteriori probability of any hypothesis consistent with the dataset is estimated, to select the most probable hypothesis. Given an example x represented by k values, the Naive Bayes classification algorithm is based on finding the most probable hypothesis that described.

In the experiments that have been carried out, the algorithms described in the previous section have been parameterized, as described below:

• Naïve Bayes. No type of parameterization has been required.

• Gradient Boosting Trees. The number of trees used is twenty, for the approach function.

• Deep Learning. The activation functions that have been used for this algorithm have been: Hyperbolic Tangent Function, Rectifier Linear, Choose the maximum coordinate of the input vector and Exponential Rectifier Linear. In addition, this algorithm is parameterized with two hidden layers and with ten epochs, that is, ten forward passes and ten backward passes of all the training examples of the dataset used.

The results that have been obtained in each metric with the machine learning algorithms discussed above are shown in "Table I".:

Machine Learning Algorithm	Metrics		
	Accuracy	*Precision*	*F-measure*
Naïve Bayes	0,7689	0,7448	0,7797
Gradient Boosted Trees	0,8712	0,8501	0,8750
Deep Learning (10 epochs - rectifier)	0,8704	0,8477	0,8745
Deep Learning (10 epochs -tanh)	0,8775	0,8519	0,8818
Deep Learning (10 epochs - Maxout)	0,8758	0,8635	0,8779
Deep Learning (10 epochs - ExpRctifier)	0,8747	0,8620	0,8759

Table I. Comparative of machine learning techniques

"Fig. 4" shows graphically the results described above for each machine learning algorithms. After comparing the different results, the best result obtained for traffic accidents classification is that provided by the Deep Learning algorithm (10 epochstanh) in its different metrics.

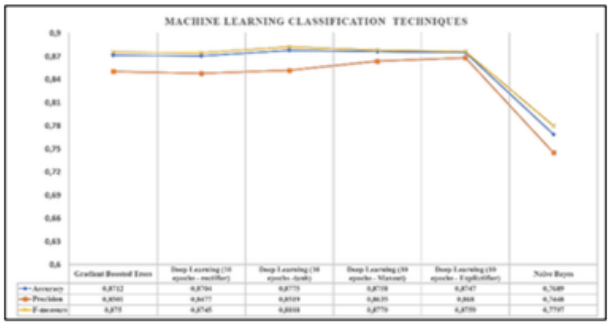

Figure 4. Results Graph

5. CONCLUSIONS

Traffic accidents constitutes the first cause of death and injury in many developed countries. However, traffic accidents information and data provided by public organisms can be exploited to classify these accidents according to their type and severity, and consequently try to build predictive model. Detecting and identifying injury severity in traffic accidents in real time is primordial for speeding post-accidents protocols as well as developing general road safety policies. This article presents a case study of traffic accidents classification and severity prediction in Spain. Raw data are from Spanish traffic agency covering a period of six years ranging from 2011 to 2015. To this end, are compared three different machine learning classification techniques, such as Gradient Boosting Trees, Deep Learning and Naïve Bayes.

The objective of this article has been to compare different machine learning classification techniques, as Naïve Bayes, Gradient Boosted Trees and Deep Learning, to develop a classification model that

determines if an accident is serious or not serious from a dataset provided by the Spanish traffic agency covering a period of six years ranging from 2011 to 2015. This classification model could support Spanish Traffic Agency to make decisions in traffic control activities, such as helping to understand the behavior of the driver, in addition to other associated problems that cause accidents where there are dead or seriously injured.

REFERENCES

[1] Dimitrijevic, Branislav, et al. Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling. No. CAIT-UTC-NC62. Rutgers University. Center for Advanced Infrastructure and Transportation, 2020.

[2] Li, Chang, et al. "Machine learning for text mining based on prediction of occupational accidents and safety risk calculation." Australian Journal of Engineering and Applied Science 13.6 (2020): 11-17.

[3] Bahrami, Javad, Viet B. Dang, Abubakr Abdulgadir, Khaled N. Khasawneh, Jens-Peter Kaps, and Kris Gaj. "Lightweight implementation of the lowmc block cipher protected against side-channel attacks." In *Proceedings of the 4th ACM Workshop on Attacks and Solutions in Hardware Security*, pp. 45-56. 2020.

[4] Chen, Lee, et al. "Machine learning established by using crowdsourced investigation vehicle data for forecast of expressway crash risk ." International Journal of Applied Science and Information Science 11.8 (2020): 356-363.

[5] Ahmadinejad, Farzad, Javad Bahrami, Mohammd Bagher Menhaj, and Saeed Shiry Ghidary. "Autonomous Flight of Quadcopters in the Presence of Ground Effect." In *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 217-223. IEEE, 2018.

[6] Zhang, Lixuan, et al. "Machine Learning Models established toward the Car Smash Injury Difficulty." European Journal of Applied Engineering and Basic Sciences 19.17 (2020): 4678-4685.

[7] Bozorgasl, Zavareh, and Mohammad J. Dehghani. "2-D DOA estimation in wireless location system via sparse representation." In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 86-89. IEEE, 2014.

[8] Xiang, Zheng, et al. "Machine Learning process for injury severity prediction and Traffic accidents classification." International Journal of Management System and Applied Science 23.12 (2020): 997-1003.

[9] Amini, Mahyar, and Aryati Bakri. "Cloud computing adoption by SMEs in the Malaysia: A multi-perspective framework based on DOI theory and TOE framework." Journal of Information Technology & Information Systems Research (JITISR) 9.2 (2015): 121-135.

[10] Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine learning applied to road safety modeling: A systematic literature review." Journal of traffic and transportation engineering (English edition) 7.6 (2020): 775-790.

[11] Amini, Mahyar. "The factors that influence on adoption of cloud computing for small and medium enterprises." (2014).

[12] AlMamlook, Rabia Emhamed, et al. "Comparison of machine learning algorithms for predicting traffic accident severity." 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, 2019.

[13] Amini, Mahyar, et al. "Development of an instrument for assessing the impact of environmental context on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.10 (2014): 129-135.

[14] Rezapour, Mahdi, Amirarsalan Mehrara Molan, and Khaled Ksaibati. "Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models." International journal of transportation science and technology 9.2 (2020): 89-99.

[15] Amini, Mahyar, et al. "The role of top manager behaviours on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.1 (2014): 490-498.

[16] Rahim, Md Adilur, and Hany M. Hassan. "A deep learning based traffic crash severity prediction framework." Accident Analysis & Prevention 154 (2021): 106090.

[17] Amini, Mahyar, and Nazli Sadat Safavi. "Critical success factors for ERP implementation." International Journal of Information Technology & Information Systems 5.15 (2013): 1-23.

[18] Siebert, Felix Wilhelm, and Hanhe Lin. "Detecting motorcycle helmet use with deep learning." Accident Analysis & Prevention 134 (2020): 105319.

[19] Amini, Mahyar, et al. "Agricultural development in IRAN base on cloud computing theory." International Journal of Engineering Research & Technology (IJERT) 2.6 (2013): 796-801.

[20] Yang, Yang, et al. "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods." Physica A: Statistical Mechanics and Its Applications 595 (2022): 127083.

[21] Amini, Mahyar, et al. "Types of cloud computing (public and private) that transform the organization more

effectively." International Journal of Engineering Research & Technology (IJERT) 2.5 (2013): 1263-1269.
[22] Fu, Yuchuan, et al. "A decision-making strategy for vehicle autonomous braking in emergency via deep reinforcement learning." IEEE transactions on vehicular technology 69.6 (2020): 5876-5888.
[23] Amini, Mahyar, and Nazli Sadat Safavi. "Cloud Computing Transform the Way of IT Delivers Services to the Organizations." International Journal of Innovation & Management Science Research 1.61 (2013): 1-5.
[24] Alkinani, Monagi H., Wazir Zada Khan, and Quratulain Arshad. "Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges." Ieee Access 8 (2020): 105008-105030.
[25] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Heuristic Solution For IaaS Cloud Placement Problem Without Migration." International Journal of Computer Science and Information Technologies 6.11 (2014): 25-30.
[26] Wahab, Lukuman, and Haobin Jiang. "Severity prediction of motorcycle crashes with machine learning methods." International journal of crashworthiness 25.5 (2020): 485-492.
[27] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Solution For IaaS Cloud Placement Problem Using Simulated Annealing." International Journal of Computer Science and Information Technologies 6.11 (2014): 52-57.
[28] Muhammad, Khan, et al. "Deep learning for safe autonomous driving: Current challenges and future directions." IEEE Transactions on Intelligent Transportation Systems 22.7 (2020): 4316-4336.
[29] Sadat Safavi, Nazli, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 70-75.
[30] Cai, Qing, et al. "Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data." Transportation research part A: policy and practice 127 (2019): 71-85.
[31] Sadat Safavi, Nazli, Nor Hidayati Zakaria, and Mahyar Amini. "The risk analysis of system selection and business process re-engineering towards the success of enterprise resource planning project for small and medium enterprise." World Applied Sciences Journal (WASJ) 31.9 (2014): 1669-1676.
[32] Wahab, Lukuman, and Haobin Jiang. "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity." PLoS one 14.4 (2019): e0214966.
[33] Sadat Safavi, Nazli, Mahyar Amini, and Seyyed AmirAli Javadinia. "The determinant of adoption of enterprise resource planning for small and medium enterprises in Iran." International Journal of Advanced Research in IT and Engineering (IJARIE) 3.1 (2014): 1-8.
[34] Ziakopoulos, Apostolos, and George Yannis. "A review of spatial approaches in road safety." Accident Analysis & Prevention 135 (2020): 105323.
[35] Safavi, Nazli Sadat, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 61-66.
[36] Mannering, Fred, et al. "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis." Analytic methods in accident research 25 (2020): 100113.
[37] Khoshraftar, Alireza, et al. "Improving The CRM System In Healthcare Organization." International Journal of Computer Engineering & Sciences (IJCES) 1.2 (2011): 28-35.
[38] Mokhtarimousavi, Seyedmirsajad. "A time of day analysis of pedestrian-involved crashes in California: Investigation of injury severity, a logistic regression and machine learning approach using HSIS data." Institute of Transportation Engineers. ITE Journal 89.10 (2019): 25-33.
[39] Abdollahzadegan, A., Che Hussin, A. R., Moshfegh Gohary, M., & Amini, M. (2013). The organizational critical success factors for adopting cloud computing in SMEs. Journal of Information Systems Research and Innovation (JISRI), 4(1), 67-74.
[40] Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine learning applied to road safety modeling: A systematic literature review." Journal of traffic and transportation engineering (English edition) 7.6 (2020): 775-790.