TRANSPORTES
www.revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
www.anpet.org.br

Evaluation of database balancing techniques for road accident severity classification employing Artificial Neural Network

Maria Lígia Chuerubim[1], Leonardo N. Ferreira[2], Alan D. B. Valejo[3], Bárbara Stolte Bezerra[4], Giuliano Sant'Anna Marotta[5], Irineu da Silva[6]

[1]Universidade Federal de Uberlândia, UFU; Escola de Engenharia de São Carlos, USP, São Paulo, Brasil, marialigia@ufu.br
[2]Laboratório Associado de Computação e Matemática Aplicada, INPE, São Paulo, Brasil, leonardo.ferreira@inpe.br
[3]Escola de Engenharia de São Carlos, USP, São Paulo, Brasil, alan@icmc.usp.br
[4]Universidade Estadual Paulista, UNESP, São Paulo, Brasil, barbara.bezerra@unesp.br
[5]Universidade de Brasília, UNB; Universidade Federal de Uberlândia, UFU, São Paulo, Brasil, marotta@ufu.br; marotta@unb.br
[6]Escola de Engenharia de São Carlos, USP, São Paulo, Brasil, irineu@sc.usp.br

Recebido:
8 de janeiro de 2020
Aceito para publicação:
21 de maio de 2020
Publicado:
15 de dezembro de 2020
Editor de área:
Flávio Cunto

Keywords:
Imbalanced data.
Accident severity.
Classification and Artificial Neural Networks.

Palavras-chave:
Dados desbalanceados.
Severidade do acidente.
Classificação e Redes Neurais Artificiais.

DOI:10.14295/transportes.v28i5.2271

ABSTRACT
An inherent feature of road accident databases is the imbalance between the number of observations associated with accidents with fatal and non-fatal victims of injuries concerning to accidents without victims. This particularity led to the adoption of corresponding balancing techniques, which can resample classes and attributes. Therefore, it ensures that there is no over-adjustment of the data in classification problems. This study investigates the influence of different balancing methods such as undersampling, oversampling and SMOTE on the classification process of road accident severity adopting an Artificial Neural Network approach. The results obtained indicate that all methods used were able to effectively adjust the balance between the minority and majority classes. Balancing leads to a better performance of the classifier, shown by the efficient adjustment of the data to the model, as the gain in the quality and accuracy of the classification process, especially, considering sampling techniques such as SMOTE.

RESUMO
Uma característica inerente aos bancos de dados de acidentes rodoviários refere-se ao desequilíbrio existente entre o número de observações associadas às ocorrências dos acidentes com vítimas fatais e não fatais, em relação aos acidentes sem vítimas. Essa particularidade conduz à necessidade da aplicação de técnicas de balanceamento, que possibilitam a reamostragem de classes e atributos. Assim, assegura-se que não haja um super ajuste dos dados em problemas de classificação. Este trabalho investigou a influência de diferentes métodos de balanceamento como undersampling, oversampling e SMOTE no processo de classificação da severidade de acidentes rodoviários pela abordagem de Redes Neurais Artificiais. Os resultados obtidos indicam que o balanceamento proporciona um ganho significativo na taxa de acerto da classificação das classes minoritárias. Verifica-se um melhor ajuste do classificador ao modelo e o ganho na qualidade e acurácia do processo de classificação, principalmente, quando são utilizadas técnicas de sobre amostragem como a SMOTE.

1. INTRODUCTION

One of the main challenges faced in data mining and machine learning is the imbalance between the number of observations and the categories of classes found in databases (Alejo *et al.*, 2013). Especially when dealing with multiclass problems, such as road safety, this particularity may

affect the Artificial Neural Network (ANN) performance in the prediction and classification process of the level of traffic accident severity (Krawczyk, 2016; Sain & Wulan, 2015).

Imbalance, also called imbalanced databases, occurs when there are majority and minority classes, i.e. when the number of instances of a given class is much higher than in other classes (Alejo *et al*., 2013). They often occur in real-world problems such as road accident databases, in which the predominant class corresponds to observations related to accidents without victims (WOV) and the less frequent class to accidents with victims of fatal and non-fatal injuries (WV).

Supervised learning in Multi-Layer Perceptron (MLP) using the "backpropagation" algorithm is the most used approach for solving non-linear problems with non-homogeneous distribution (Alejo *et al*., 2013). However, when dealing with imbalanced databases, the algorithm tends to converge more slowly and ends up making the learning process biased, neglecting the less recurring classes in the database (Alejo *et al*., 2013).

In the literature, the problem of class imbalance is minimised by using database balancing or sampling techniques such as undersampling, oversampling and SMOTE (Synthetic Minority Over-sampling Technique), (Li *et al*., 2017). These techniques reduce the number of samples of the majority class and increase the magnitude of the samples of the minority class to carry out the rebalancing process of the database.

Undersampling reduces the population of the majority class and implies loss of information. This means elimination of samples belonging to the dominant class, which may affect the quality of the classification process (Alejo *et al*., 2013). Oversampling balances class distribution by increasing the population of the minority class by random replication of the samples present in these classes and, in general, does not imply in any loss of information as no sample is discarded (Alejo *et al*., 2013).

In this context, one of the most used sampling techniques is SMOTE, in which the minority class is balanced considering each one of its instances and the introduction of synthetic samples based on neighbourhood criteria, such as the Euclidean distance, using the nearest neighbour algorithm (Bolón-Canedo *et al*., 2014). In practice, it calculates the characteristics of these neighbours to create new synthetic data without duplicating data, that is, new data is generated. For more details, it is recommended to consult Fawcett (2006).

Although techniques that use oversampling and SMOTE are the most promising in database balancing, they have a higher computational cost when compared to undersampling techniques in the classification process, especially when using classifiers based on similarities such as the nearest neighbour. In addition, superabundant sampling can cause overweight of the data and provide “pseudo-accuracy" in the classification process (Salunkhe, 2016).

In this research, we present a study on the technique of balancing, using supervised learning with the backpropagation algorithm, before and after adopting undersampling, oversampling and SMOTE techniques. Based on this study, the data rescheduling method is discussed, which was selected a priori and used with supervised learning adopting an ANN with MLP. In addition, this research discusses the impact of processing imbalanced and balanced data in the classification process and studies about road accident severity in order to contribute to the theoretical and practical argumentation of the multiclass problem.

This article is divided into 5 sections. Section 1 presents the contextualization and objectives of the proposed research. Section 2 discusses the literature on the problem of imbalance in road accident databases. In this section, an analysis is made of the methodologies adopted and the

limitations found in road safety studies. The Section 3 discusses presents an approach of learning by ANNs. Section 4 highlights the methodological approach adopted to carry out this study. It includes the selection of the analysis variables, the unbalanced database balancing step and the ANN approach. Section 5 presents the results obtained based on the proposed ANNs and discusses the results obtained in the classification process based on performance measures. Finally, the conclusion of the study and the recommendations for future work are presented in Section 6.

2. UNBALANCING PROBLEM ON ROAD TRAFFIC DATABASE

As previously mentioned, road accident severity classification poses a huge challenge for traffic engineering research due to the imbalance between the classes found in the database. This problem occurs when a dataset is dominated by the main class (for example, accidents WOV) to the detriment of classes considered rarer or minority, such as accidents WV. Furthermore, there are additional experimental complications due to the variations of the database size, which is a factor that leads to significant errors in estimating the level of traffic accident injuries (Bolón-Canedo *et al*., 2014; Wang *et al*., 2014).

The traditional pre-processing techniques used to minimise or solve this question consist of undersampling methods, which create a subset from the original imbalanced database by eliminating instances. The oversampling methods, that create a subset based on the original database using the replicating process of certain instances or by creating new instances from the pre-existing classes in the database; and, finally, the hybrid methods, that combine these two sampling methods (Bolón-Canedo *et al*., 2014).

Wang *et al*. (2014) emphasise that in the literature on traffic accident severity, most of the studies use classification accuracy to measure the quality of a given classifier as a decision tree, ANN, Bayesian Networks, etc. (Chen *et al*., 2016) when using an imbalanced dataset. Therefore, these authors used a database with 3,105 traffic accidents in Beijing, China, from 2008 to 2010, of which 1,996 were nonfatal accidents (64.3%) and 1,109 were fatal accidents (35.7%). They also used data mining techniques with the k-means algorithm to identify the risk factors related to accident severity based on twelve variables that include information about the accident severity, the time of the accident and the driver's characteristics (e.g. age, gender, seat belt use), accident type, cause of the accident, type of vehicle involved, highway characteristics (signposting, central or non-central reservation and number of lanes). They also point out the importance of using the ROC (Receiver Operating Characteristic) curve to identify the classifier performance and class distribution in the database.

The ROC curve expresses the relationship between the sensitivity and the specificity of the model, and it is widely used for the evaluation of classifiers (Fawcett, 2006; Prati, Batista & Monardi). The sensitivity corresponds to the proportion of true positives, that is, it evaluates the model's ability to correctly classify an observation. The specificity is equivalent to the proportion of true negatives, that is, the model's ability to predict observations incorrectly.

Fouladgar *et al*. (2017) applied balancing and neighbourhood relations in databases obtained from traffic control stations in Northern California, USA, totalling data obtained from 39,000 detectors, to track congestion levels in a road network in real-time Using a database containing the 10-year history of flow data on expressways, as well as data related to incidents such as accidents, meteorological information, road closure and obstruction, etc., the authors generated traffic flow prediction models for the studied routes.

Delen; Sharda; Bessonov (2006); Mussone; Ferrari; Oneta (1999) used balancing techniques in the road accident data processing step and classification algorithms based on ANN modelling to reduce computational costs and enhance classifier performance. These authors mention the class distribution problem in traffic databases and the implications of this characteristic to investigate the road accident observational phenomenon.

3. ARTIFICIAL NEURAL NETWORKS LEARNING

The principle of Artificial Neural Networks (ANN) is to minimise the Mean Squares Error (MSE) provided by equation (1):

$$MSE = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{k} (t_{i,j} - a)^2, \quad (1)$$

where t and a are the observed and estimated parameters, respectively; K is the number of neurons of the output layer and N the size of test dataset (Chang, 2005).

In this work, the SPSS software was used to construct ANN with Multiple Layers Perceptron (MLP), based on the "backpropagation" algorithm, which corrects the weights in all layers, starting from the output layer to the input layer, using the mean square root error, in two phases: the phase forward and the backward phase. In the phase forward, each variable of the database is stored in a neuron of the network (Facelli *et al.* 2011; Yuan *et al.*; 2019).

The weight adjustment in MPL using the back-propagation algorithm can be obtained by equation (2), (Facelli *et al.* 2011):

$$w_{j,l}(t+1) = w_{j,l}(t) + \eta x^j \delta_l \quad (2)$$

where, $w_{j,l}$ represents the weight between the l^{th} neuron and the j^{th} input attribute or the j^{th} output from the neuron in the previous layer; δ_l is the error associated with the l^{th} neuron; x^j corresponds to the input received by this neuron, that is, the j^{th} input attribute or the j^{th} output from the neuron in the previous layer; and η is the model's learning rate.

In the back-propagation algorithm, the error associated with a neuron in an intermediate layer is estimated as the sum of errors of the neurons in the following layer, whose inputs are connected to it. The weighing occurs based on the weights attributed to these connections.
This way, the error calculation will depend on the layer in which the neuron is located, as shown in equations (3) and (4), (Facelli *et al.* 2011):

$$\delta_l = f_a^{'} e_l, \text{ if } n_l \in \text{layer}_{\text{output}}, \quad (3)$$

$$\delta_l = f_a^{'} \sum w_{l,k} \delta_k, \text{ if } n_l \in \text{layer}_{\text{intermediate}}, \quad (4)$$

where, n_l represents the lth neuron; $f_a^{'}$ is the partial derivative of the neuron activation function; and e_l is the squared error of the output neuron when its answer (y_q) is compared to

$$e_l = \frac{1}{2} \sum_{q=1}^{k} (y_q - \hat{f}_q)^2 \quad (5)$$

The weight adjustment is defined by $f_a^{'}$. This derivative measures the contribution of each weight to the network error for each variable used in the analysis.

In this work, the first subset uses 70% of the data for ANN training and the second subset uses 30% of the data for the ANN test. These subsets are used iteratively for learning, validation and cross-validation without repetition. In this study, the input layer contains 39 neurons. The variable "type of accident" is represented by 10 neurons; the variable "weather condition," by 5 neurons; the variable "accident cause," by 4 neurons; and the variable "mileage," by 1 neuron. The other variables, such as visibility, road profile, road geometry, pavement condition, and period are represented with 3 neurons each. Horizontal signal and vertical signal are represented with 2 neurons each.

For the transfer of information between the input-layer neurons and the hidden layer, the hyperbolic tangent activation function was used, in which the selection of the network architecture was performed automatically. In addition, for the transfer of information between the hidden layer and the output layer, the Softmax activation function was used, in which all the independent variables are categorical.

4. MATERIALS AND METHODS

4.1. Road accident database

The database used in this study consists of 4,259 traffic accident occurrences observed from 2009 to 2016 on the Dom Pedro I highway (SP-065) in the municipality of Campinas, Brazil between km 125 and km 146. From 2009 to 2012, a period leading up to improvements on the highway and the construction of marginal roads in some segments of the stretch under study, 2,909 accidents were observed (68.30%) and from 2013 to 2016, after introducing traffic accident countermeasures, there were 1,350 accidents (31.70%).

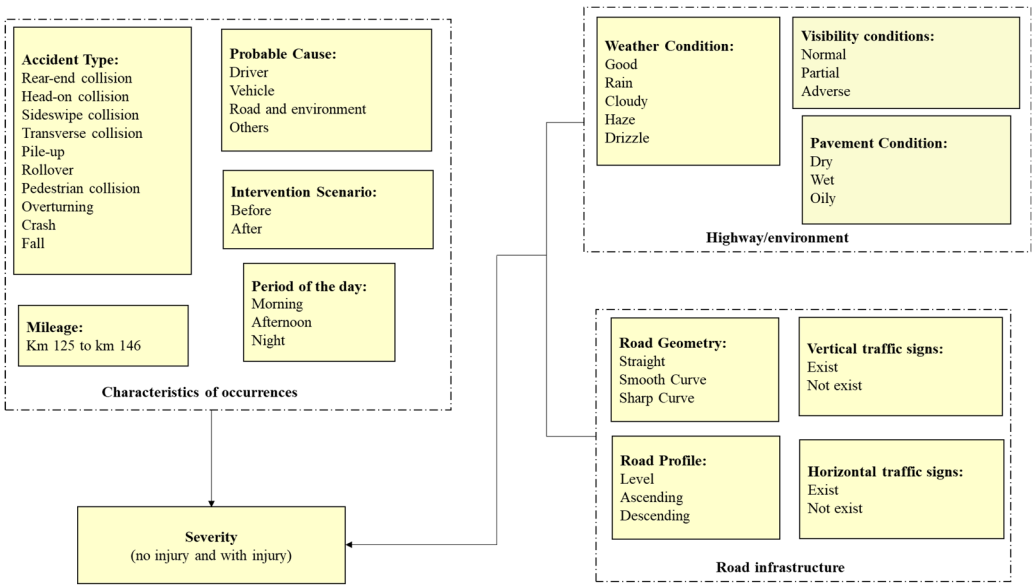


Figure 1. Variables selected for accident severity classification

In order to characterize the accident severity according to the infrastructure, road conditions/environment and driver, twelve variables were selected that represent the information related to the accident type, weather conditions, visibility conditions, road profile, road geometry, pavement condition, the period of the day, probable cause, horizontal traffic signs, vertical traffic signs, intervention scenario and the mileage (km). They can be found in the categories shown in Figure 1.

4.2. Database balancing and classification

The methodology used in this experiment consists of initially applying database balancing techniques to solve or minimise the class imbalance problem found in road accident databases. In order to do this, three basic and already traditional sampling techniques were used in the computation area, namely undersampling, oversampling and SMOTE, in order to mitigate the performance of the ANN algorithm in the prediction process and the difficulties in classifying road accident severity.

To do this, an exploratory analysis of the original and imbalanced database was carried out to identify, a priori, the majority and minority classes. Furthermore, later on, each balancing technique was performed individually resulting in different balanced databases. The undersampling technique reduced the number of observations related to accidents WOV (majority class), resulting in a balanced database with 99 occurrences, of which 58 corresponded to accidents WOV (58.59%) and 41 to accidents WV (41.41%).

The oversampling technique increased the size of the accident database by increasing the number of observations related to accidents WV (minority class). The resulting balanced database totaled 8,416 occurrences of which 47.83% corresponded to accidents WOV and 52.17% to accidents WV. Moreover, the SMOTE technique produced a balanced database with 353 observations due to synthetic observations in the minority class, of which 61.19% were accidents WOV and 38.81% were accidents WV.

Then the accident severity modelling was done for each dataset (original or imbalanced, undersampling, oversampling and SMOTE) by the multi-layer ANN (Multi-Layer Perceptron) approach using the supervised learning technique called "backpropagation". For ANN iterative learning, we extracted the training and test subsets from the cross-validation with 70.0% and 30.0% of the data, respectively. Considering information transfer between the input layer neurons and the hidden layer, the hyperbolic tangent activation function was used. The Softmax activation function was used for the information transfer between the hidden layer and the output layer.

5. EXPERIMENTS AND RESULTS

The results of this study indicated significant differences between the approaches used in the road accident database balancing.

Initially, we classified the accident severity using the imbalanced database and ANN approach with MLP adopting the backpropagation algorithm. In this phase, an overall accuracy of 76.7% and the best accuracy of only 5.0% for accidents WOV and 95.0% for accidents WV were obtained, as shown in Table 1.

The most important variables in the classification of accidents WOV and WV, based on the imbalanced database, were accident type (100.0%), probable cause (68.1%), road geometry (50.2%) and pavement condition (45.2%). The other variables exhibited importance below 40.0%. Figure 2 shows the ROC curve based on the 12 predictor variables, in which the average values of the order of 0.641 can be observed for the target variables (WOV and WV).

It should be noted that the values of the Area under the Curve (AUC), in the range of $0.7 \leq AUC < 0.8$, provide a model with acceptable discrimination, while values between $0.8 \leq AUC < 0.9$ show highest performances. Finally, values of $AUC \geq 0.9$ showed excellent modelling (Hosmer & Lemeshow, 2000).

Table 1 – Classification performance adopting ANN with MLP using the imbalanced database

		Predicted		
Subsets	Injury level			
		WOV	WV	(%) Best accuracy
Training	WOV	2,163	46	97.9%
	WV	643	106	14.2%
	(%) Total	94.9%	5.1%	76.7%
Test	WOV	954	21	97.8%
	WV	282	44	13.5%
	(%) Total	95.0%	5.0%	76.7%

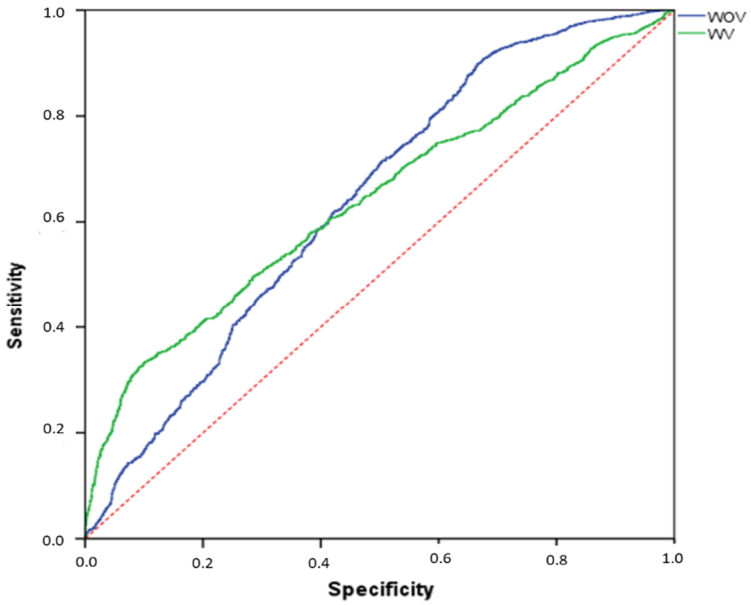


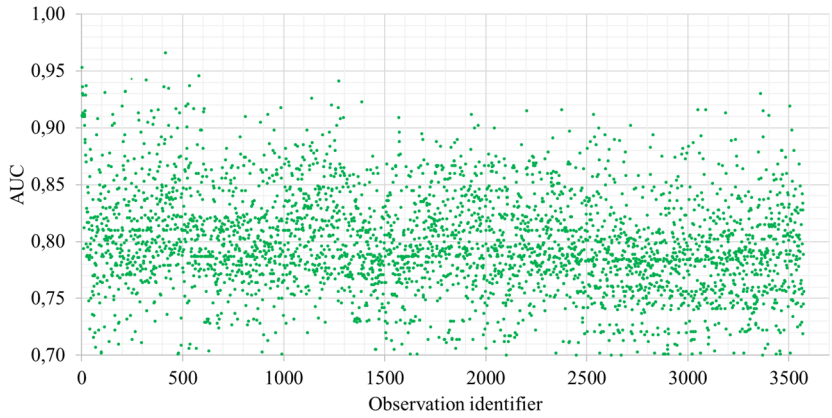
Figure 2. ROC curve obtained for the Accident Severity dependent variable using the imbalanced database

Figures 3a and 4a present, respectively, the scores obtained for the predictions of categorical severity variables considering accidents WOV and WV using the original database. The scores in the graphs illustrated in Figures 3, 4, 6, 7, 9, 10, 12 and 13, represent a variation of the AUC according to the number of presentations.

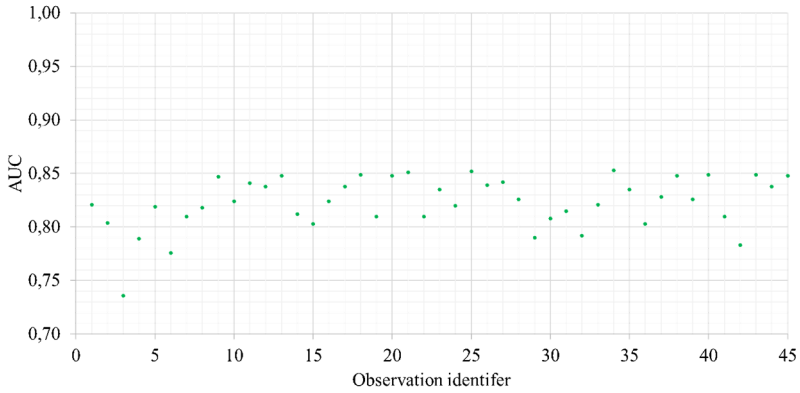
The accident severity modelling (WOV and WV) by undersampling resulted in an overall precision of 82% according to the twelve predictor variables. However, it exhibited a low accuracy for the minority class of 28.6%, while for the majority class the best accuracy rate was of the order of 71.4%, as shown in Table 2.

Figure 5 shows the ROC curve for each dependent variable (WOV and WV) using the undersampling technique. When analysing the ROC curve graph (Figure 6), it can be observed that the predicted values had mean values of 0.895 for both accidents WOV and WV. The most important variable in the process of predicting severity using undersampling was the probable cause (100.0%), followed by accident type 57.6%, visibility conditions (54.0%), road profile (46.7%) and period of day (46.0%). The other variables used in the modelling presented importance below 40.0%.

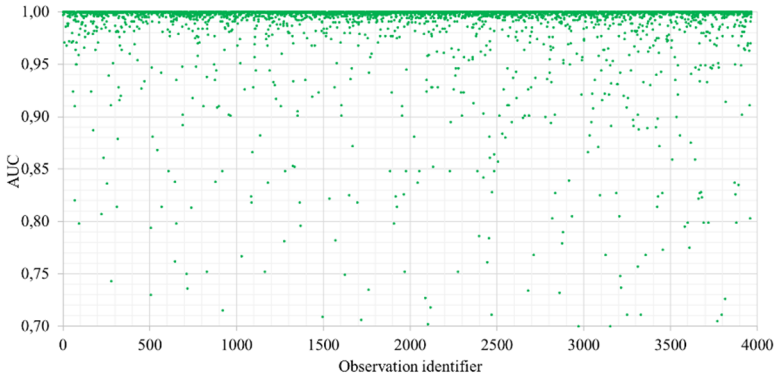
Figures 3b and 4b present, respectively, the scores obtained for the predictions of categorical severity variables considering the accidents WOV and WV using the undersampling technique.



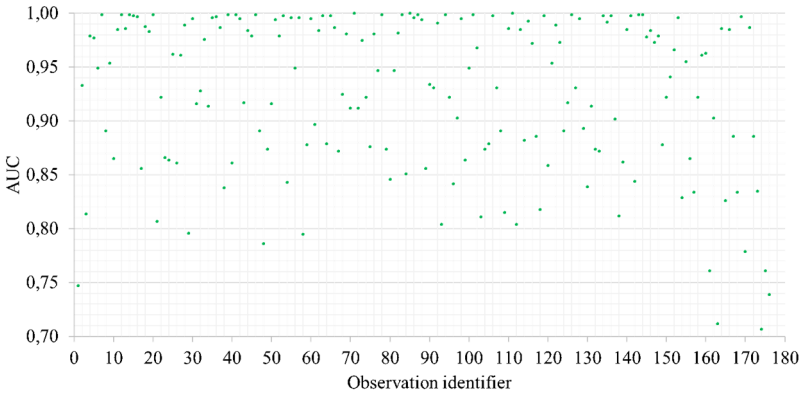
(a)



(b)

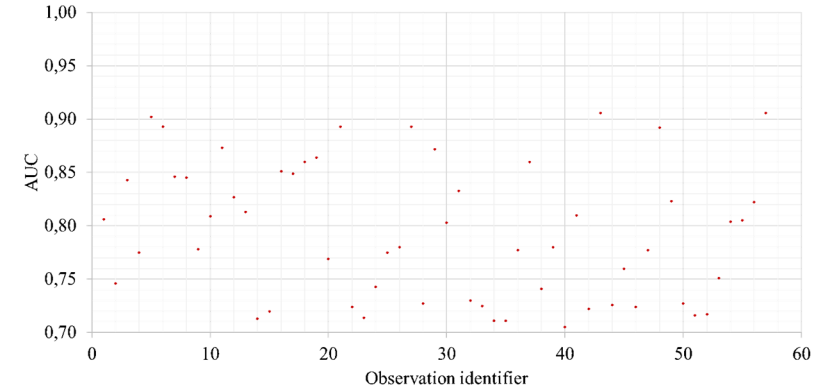


(c)

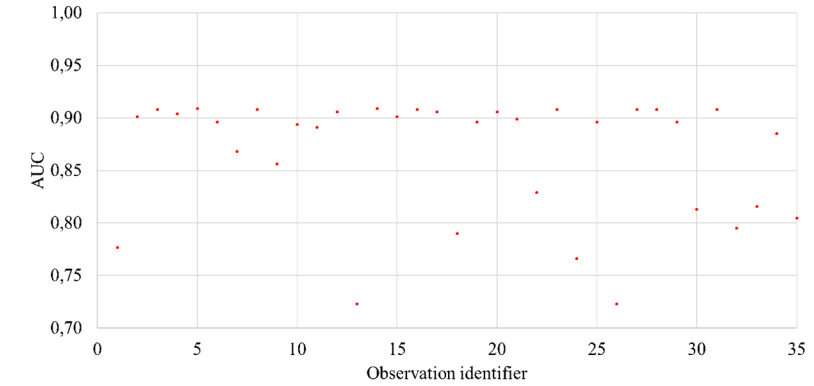


(d)

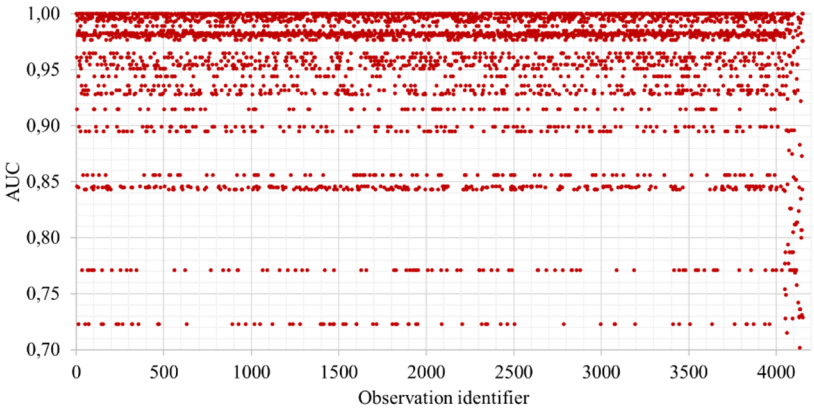
Figure 3. Scores obtained from the prediction of accidents WOV using: (a) Imbalanced database; b) Undersampling; c) Oversampling; d) SMOTE



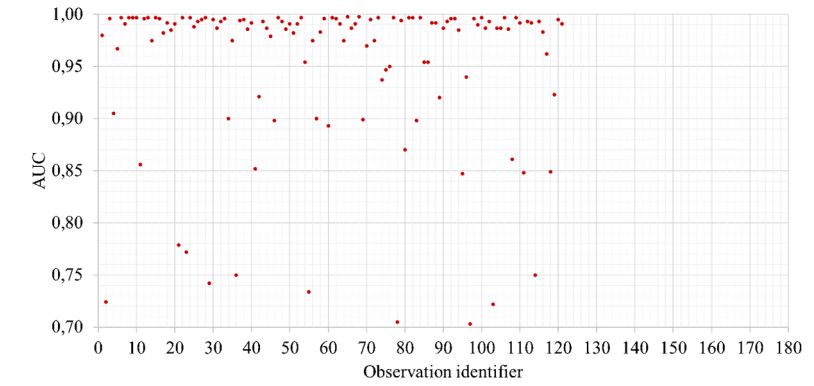
(a)



(b)



(c)



(d)

Figure 4. Scores obtained from the prediction of accidents WV using: **(a)** Imbalanced database; **b)** Undersampling; c) Oversampling; d) SMOTE

Table 2 – Undersampling performance

		Predicted		
Subsets	Injury level			
		WOV	WV	(%) Best accuracy
Training	WOV	28	7	80.0%
	WV	10	26	72.2%
	(%) Total	53.50%	46.5%	76.1%
Test	WOV	15	0	100.0%
	WV	5	8	61.5%
	(%) Total	71.4%	28.6%	82.0%

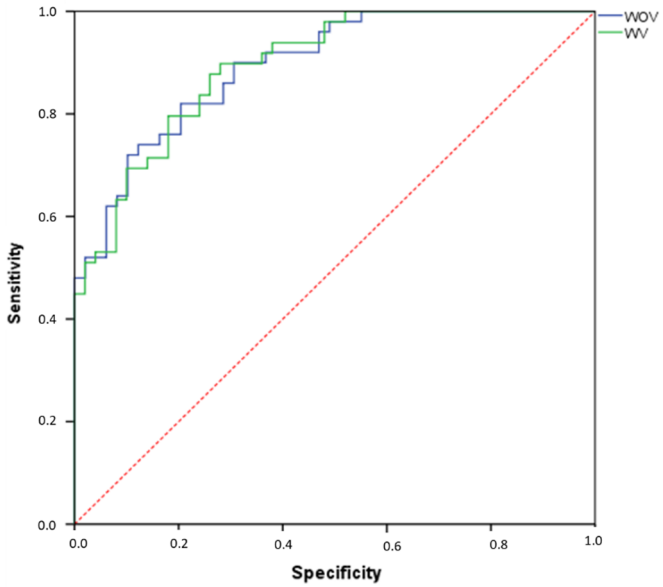


Figure 5. ROC curve obtained for the Accident Severity dependent variable using the undersampling technique

Considering only the predictions with $AUC \geq 0.7$, it can be observed that the classification of accidents WOV (Figure 3b) presented an average score of 0.82 and minimum and maximum values of, respectively, 0.74 and 0.85. Predictions for accidents WV (Figure 4b) resulted in an average score of 0.87 and minimum and maximum values of, respectively, 0.72 and 0.91.

The accident severity modelling (WOV and WV) by oversampling resulted in an overall precision of 97.4%. Although this sampling made it possible to improve the overall precision of the classification when compared to the undersampling technique, it is very probable that a super fit of the data to the model occurred, since there was no confusion or classification error of accidents WV, as can be seen in Table 3.

Table 3 – Performance of the classification of accident severity

		Oversampling Sampling		
		Predicted		
Subsets	Injury level			
		WOV	WV	(%) Best accuracy
Training	WOV	2,826	116	96.1%
	WV	0	2,916	100.0%
	(%) Total	48.2%	51.8%	98.0%
Teste	WOV	1,199	67	94.7%
	WV	0	1,292	100.0%
	(%) Total	46.9%	53.1%	97.4%

Figure 8 shows the ROC curve for each dependent variable (WOV and WV) as a function of the oversampling technique.

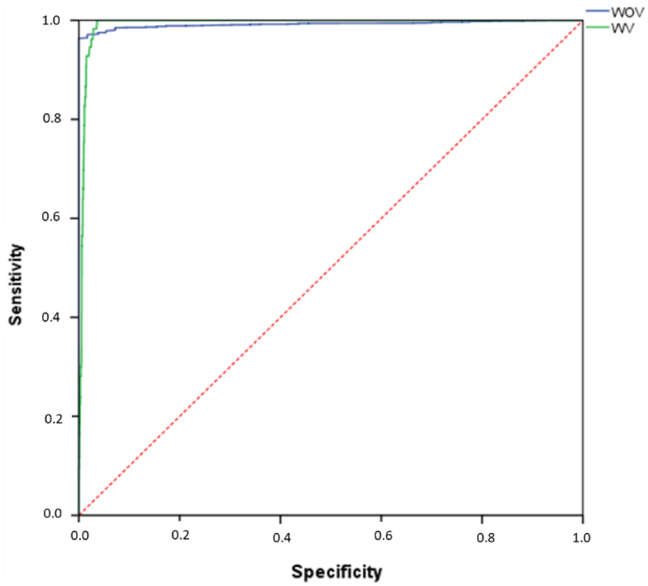


Figure 8. ROC curve obtained for the Accident Severity dependent variable using the oversampling technique.

In the ROC curve (Figure 8), it can be observed that the predicted values presented mean values of 0.992 for both accidents WOV and WV. The most important variables in the prediction of severity using oversampling were accident type (100.0%), probable cause (93.8%), mileage (71.5%), visibility conditions (54.6%), weather conditions (46.3%), route (44.9%) and period of day (44.0%). The other variables used in the modelling exhibited importance below 40%.

Figures 3c and 4c present, respectively, the scores obtained for the predictions of categorical severity variables considering accidents WOV and WV using oversampling.

In the oversampling technique considering only the predictions with $AUC \geq 0.7$ it can be observed that the classification of accidents WOV (Figure 3c) presented an average score of 0.99 and minimum and maximum values of, respectively, 0.70 and 1.00.

The predictions for accidents WV (Figure 4c) resulted in an average score of 0.95 and minimum and maximum values of, respectively, 0.70 and 1.00.

The sampling of the database by the SMOTE technique resulted in an overall accuracy of 90.4%. Table 4 presents the classification error for accidents WOV and WV considering the 12 predictors.

Table 4 - Performance of the classification from SMOTE

		Predicted		
Subsets	Injury level			
		WOV	WV	(%) Best accuracy
Training	WOV	138	5	96.5%
	WV	20	96	82.8%
	(%) Total	61.0%	39.0%	90.3%
Test	WOV	53	4	93.0%
	WV	5	32	86.5%
	(%) Total	61.7%	38.3%	90.4%

Table 4 shows that the SMOTE sampling had a classification error of 61.7% for accidents WOV and 38.3% for accidents WV. Figure 11 shows the ROC curve for each dependent variable (WOV and WV) using the SMOTE sampling technique.

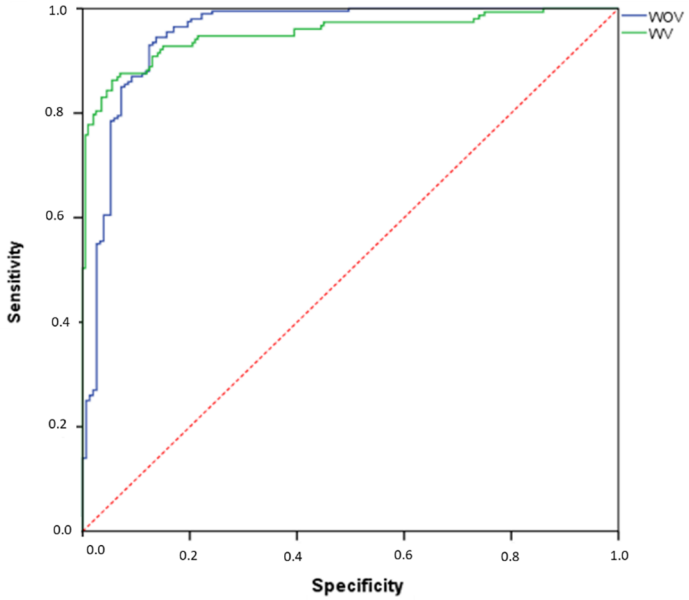


Figure 11. ROC curve obtained for the Accident Severity dependent variable using SMOTE sampling.

The ROC curve (Figure 11) using the SMOTE sampling presents mean values of 0.952 for the predictions of accident classes WOV and WV. The most important variables in the prediction process of severity using oversampling were Accident Type (100.0%) and Probable Cause (55.6%). The other variables presented importance below 40% in the modelling process.

Figures 12 and 13 presents, respectively, the scores obtained for the predictions of categorical severity variables considering accidents WOV and WV using SMOTE sampling.

The prediction of accident severity WOV (Figure 3d) with ANN using MLP and according to the SMOTE sampling technique resulted in an average score of 0.93 and a minimum and maximum value of, respectively, 0.71 and 1.00. For accidents WV (Figure 4d) resulted in an average score of 0.95 and minimum and maximum values of, respectively, 0.70 and 1.00.

As sampling by SMOTE uses neighbourhood criteria in the sampling process, it can be observed that the classification error is better distributed among the classes predicted for the severity dependent variable (WOV and WV). There is a significant performance improvement of the backpropagation classifier using ANN with MLP. In comparison to the oversampling techniques, SMOTE presents more promising and realistic results both in the quality of the classification and in relation to the computational gain, because it requires a shorter running time.

Table 5 (Appendix I) presents the classification of the severity of accidents WOV and WV considering the road infrastructure, environmental and accident characteristic variables. It can be seen that accidents WOV with the highest probability of occurrence are rear-end collisions (95.0%), sideswipe collisions (97.2%), pile-ups (71.7%) and crashes with fixed or mobile objects (100%), in good weather conditions (58.7%) and normal visibility conditions (58.7%) or partial visibility conditions (73.7%). These accidents were observed in the road profile with descending (55.1%), level (69.2%) and ascending (52.9%) stretches and occurred predominantly in road stretches with straight routes (62.2%) and dry road conditions (59.3%).

They present the probability of occurring in all periods of the day: morning (70.1%), afternoon (69.2%) and night (40.8%).

The main probable cause associated with accidents WV are drivers (73.6%) and other factors such as cyclists, pedestrians and animals on track, congestion, previous accidents and suicide (3.10%). They are frequent throughout the segment under study (km 125 to km 146), in stretches with good horizontal traffic signs (62.8%) and vertical traffic signs (62.8%). In the period prior to the interventions from the "Rota das Bandeiras" concessionaire in the stretch, the probability of occurrence was 56.8% and after that period the probability of occurrence was 72.9%.

The accidents WV with the highest probability of occurrence are head-on collisions (66.7%), transverse collisions (94.7%), pedestrian collisions (97.8%) and overturning (75.0%). They occur in good weather conditions (41.3%) or rain (26.3%). They occur in normal visibility conditions (41.3%), partial (30.8%) or adverse (14.3%). They were observed in all road profiles with a probability of 44.9% in descending stretches, 30.8% in level stretches and 47.1% in ascending stretches. They occurred predominantly in straight sections (37.8%) and sharp curves (63.6%), with dry road conditions (40.7%). They are more likely to occur at night (59.2%). The main causes of these accident types are: drivers (26.4%), pedestrians (100%) and other factors (36.9%). They are most frequently observed in the vicinity of km 128 (67%), km 134 (64.1%), km 137 (58.3%), km 139 (35.5%), km 141 (42.8%) and km 142 (53.5%). In the period prior to the intervention from the "Rota das Bandeiras" concessionaire in the stretch, there was a probability of occurrence of the order of 43.2% and after this period a probability of (27.10%), i.e., the number of accidents WV presented a natural decrease after introducing traffic accident countermeasures.

Overall, the results of the empirical analysis indicate compelling evidence that the use of balancing techniques has improved the performance of a grading algorithm evaluated on a road accident basis. These findings are important in reducing the effort required for future work and serve as guidelines for advanced research and the development of new solutions addressing machine learning problems defined in the context of road accidents.

6. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

In recent years, the scientific community has specifically addressed the problem of class imbalance in databases as it poses a challenge in many real-world applications. In the context of transport engineering, supervised classification of imbalanced traffic accident databases is a complex problem as most classifiers are sensitive to class distribution, favouring only the most frequent instances.

Therefore, dealing with knowledge-based database classification involves understanding the origin of the problem (whether linear or non-linear) and the distribution (whether normal or not normal) of the variables found in the database. In general, rare instances are the most relevant because they characterize exceptional occurrences in the database, such as instances associated with accidents with victims (fatal and non-fatal).

In this experimental study, the preliminary results obtained indicate the relevance of investigating the balancing method that best suits the treatment of the majority and minority classes found in a database. Although resampling methods are independent of the classifier used, when balancing a database through different techniques (undersampling, oversampling and SMOTE), there is an improvement in ROC (AUC) performance and a significant computational gain.

Although this study is limited to the available data, the results obtained show that there are statistically significant differences between the sampling techniques used. In a comparative analysis, it can be observed that the accuracy of the classifier increases when using balancing methods. The general accuracy obtained in the accident severity classification using an imbalanced database is only 76.7%, while for undersampling it is 82.0% and for oversampling and SMOTE techniques it is, respectively, 97.4% and 90.4%.

Although the undersampling technique provides an increase in the classification accuracy when compared imbalanced databases, it implies a loss of information as it excludes majority class samples from the data in the resampling process. Therefore, the relevant characteristics of the most frequent classes are lost in this process.

The oversampling and SMOTE resampling techniques are best suited to offset the class distribution of the database. However, it can be observed that the oversampling technique causes a super fit of the data to the model, which implies that the general accuracy of 97.4% is considered a pseudo-precision, which can be seen by the absence of confusion or error in accidents WOV, due to an increase in the minority class of data.

The SMOTE provides the distribution of the mean classification error between accidents WOV (61.7%) and accidents WV (38.3%) in a more balanced way since it used neighbourhood criteria, such as the nearest neighbour in the classification process of the variables. The classification using SMOTE presented an average score of 0.93 and minimum and maximum values between 0.70 and 1.00, respectively.

Recommendations for future work include studying different database balancing techniques and the application of hybrid approaches that jointly use concepts of subsampling and oversampling. In addition, although the process of rescheduling the database is independent of the classifier used, it is recommended to test variations of the backpropagation algorithm and different approaches based on networks such as Bayesian networks and complex networks.

In addition, it is recommended to explore, where possible, databases of different sizes in order to investigate the sensitivity of sampling techniques in relation to the total number of observations, the number of instances or attributes and the number of classes that are desired to perform the classification based on a target variable and a set of predictor variables.

ACKNOWLEDGEMENTS

The authors thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and the São Paulo Research Foundation (FAPESP) grants 2017/05831-9, 19/14429-5, 15/50122-0 and 19/07665-4 for the financial support to develop this research.

REFERENCES

Alejo, R.; Valdovinos, R. M. García, V. e J. H. Pacheco-Sanchez (2013) A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters*, v. 34, n. 4, p. 380–388. DOI: 10.1016/j.patrec.2012.09.003

Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A.; Benítez, J. M. e F. Herrera (2014) A review of microarray datasets and applied feature selection methods. *Information Sciences*, v. 282, p. 111–135. DOI: 10.1016/j.ins.2014.05.042

Chang, L-Y (2005) Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, v. 43, p. 541-557. DOI: 10.1016/j.ssci.2005.04.004

Chang, L. e H. Wang (2006) Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, v. 38, p. 1019–1027. DOI: 10.1016/j.aap.2006.04.009

Chen, C.; Zhang, G.; Qian, Z.; Tarefder, R. A. e Z. Tian (2016) Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, v. 90, p. 128–139. DOI: 10.1016/j.aap.2016.02.011

Delen, D.; Sharda, R. e M. Bessonov (2006) Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention*, v. 38, p. 434–444. DOI: 10.1016/j.aap.2005.06.024

Facelli, K.; Lorena, A. C.; Gama, J. e A. C. P. L. F, Carvalho (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC. 378p.

Fawcett, T. (2016) *Learning from Imbalanced Classes*. Available in: <https://www.svds.com/learning-imbalanced-classes/ >. Access: November/2018.

Fouladgar, M.; Parchami, M.; Elmasri, R. e A. Ghaderi (2017) Scalable Deep Traffic Flow Neural Networks for Urban Traffic Congestion Prediction. *International Joint Conference on Neural Networks (IJCNN),* p. 2251–2258. DOI: 10.1109/IJCNN.2017.7966128

Hosmer, D.W. e S. Lemeshow (2000) *Applied logistic regression*, 2nd Ed. John Wiley & Sons, New York.

Krawczyk, B. (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, v. 5, n. 4, p. 221–232. DOI: 10.1007/s13748-016-0094-0

Li, J.; Fong, S.; Wong, R. K.; Mohammed, S.; Fiaidhi, J. e Y. Sung (2018) A suite of swarm dynamic multi-objective algorithms for rebalancing extremely imbalanced datasets. *Applied Soft Computing Journal*, p. 1–22. DOI: 10.1016/j.asoc.2017.11.028

Mussone, L.; Ferrari, A. e M. Oneta (1999) An analysis of urban collisions using an artificial intelligence model. *Accident Analysis & Prevention,* 31, v. 31, p. 705–718. DOI: 10.1016/S0001-4575(99)00031-7

Prati, R. C.; Batista, G. E. A. P. A. e M. C. Monard (2008) Curvas ROC para avaliação de classificadores [Internet*]. IEEE Latin America Transactions*. 2008; 6 (2): 215-222.Available from: http://ieeexplore.ieee.org/stamp/stamp.do?arnumber=4609920&isnumber=4609907

Salunkhe, U. R. e S. N. Mali (2016) Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach*. International Conference on Computational Modeling and Security (CMS 2016),* v. 85, n. Cms, p. 725–732. DOI: 10.1016/j.procs.2016.05.259

Wang, C.; Qiu, C.; Zuo, X. e C. Liu (2014) An Accident Severity Classification Model Based on Multi-Objective Particle Swarm Optimization. *IEICE Trans. Inf. & Syst.,* n. 11, p. 2863–2871 DOI: 10.1587/transinf.2014EDP7069

Yuan, J., Abdel-Aty, M., Gong, Y. e Q. Cai (2019). Real-time crash risk prediction using long short-term memory recurrent neural network. Transportation research record, 2673(4), 314-326. DOI: 10.1177/0361198119840611