Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 95 (2026) 504–511

Transportation Research Procedia

www.elsevier.com/locate/procedia

Euro Working Group on Transportation Annual Meeting 2025 - EWGT2025

Synthetic Resampling Algorithm for Response Class Imbalance in Supervised Learning: Application to Road Accident Severity Prediction

Filomena Mauriello[a], Massimo Aria[b], Roberta Siciliano[c], Francesco Galante[a], Alfonso Montella[a]

[a]*Department of Civil, Architectural and Environmental Engineering, University of Naples Federico II*
[b] *Department of Economics and Statistics, University of Naples Federico II*
[c] *Department of Electrical Engineering and Information Technology, University of Naples Federico II*

Abstract

Road traffic injuries are a leading cause of death worldwide and are projected to become even more critical by 2030. Understanding the factors influencing crash severity is essential for developing effective safety interventions. However, crash data often suffer from severe class imbalance, especially when distinguishing between fatal and non-fatal accidents. Traditional machine learning algorithms tend to perform poorly under these conditions, favoring the majority class and misclassifying critical minority cases. To address this, we propose a novel resampling algorithm—SONCA (Synthetic Over-sampling for Numerical and Categorical variables)—designed to balance datasets containing mixed data types. Unlike existing oversampling methods, SONCA handles numerical, ordinal, nominal, and dichotomous. We evaluated SONCA using both parametric (Logit) and non-parametric (CART) models on imbalanced datasets: PTW-ISTAT. The original models failed to detect the minority class effectively, while models estimated on SONCA-resampled data showed substantial improvements in True Positive Rate, G-mean, and Fmeasure. These results demonstrate SONCA's potential as a flexible, model-agnostic preprocessing tool for addressing class imbalance in diverse real-world scenarios

Keywords: Unbalanced data; Road accident; Accident severity; Supervised Learning

1. Introduction

Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally and are predicted to become the seventh leading cause of death by 2030. Identifying the factors that influence crash injury severity and understanding their impact are crucial for the planning and implementing highway safety improvement programs. Additionally, the EU Road Safety Policy Framework 2021–2030 significantly emphasises serious injury crashes, aiming to halve their number by 2030 (Rella Riccardi et al., 2022).

10.1016/j.trpro.2026.02.064

However, several studies have established that crashes are rare events and that the crash and non-crash cases are extremely imbalanced (Laureshyn and Varhelyi 2018; Oh et al., 2010; Cai et al., 2020). The problem of class imbalance in road traffic accident data becomes even more critical when analyzing accident severity, especially for severe and fatal crashes (Fiorentini and Losa, 2020; Hans et al., 2024).

A dataset is considered imbalanced when the response variable classes are not approximately equally distributed; one class significantly outweighs the other, making it more challenging to identify and predict the underrepresented events (He et al., 2009). This problem exists in many real-world domains, such as medical diagnosis of particular cancer, oil spill detection, network intrusion detection, fraud detection (Chawla, 2003; Guo et al., 2008), and human behaviours (Song et al., 2013). In all these cases, the minority class is often the most critical to identify accurately. For example, in medical diagnosis, patients with rare diseases such as cancer typically belong to the minority class. If a cancer patient is misclassified as healthy, they will not receive the necessary treatment, potentially leading to disease progression and severe health consequences (Cao et al., 2011).

This imbalance complicates predictive modelling, as traditional machine learning algorithms tend to favour the majority class, resulting in biased predictions and poor classification of severe crashes (Fernandeset al., 2019, He et al., 2009).

Imbalanced datasets significantly impact the performance of classification models, particularly in non-trivial learning problems (Hancock et al., 2023; Ndour & Dossou-Gbété, 2012).

The reasons for the poor performance of the existing classification algorithms on imbalanced data sets are (Kotsiantis et al., 2006; Werner et al, 2023): (a) They are accuracy driven, i.e., their goal is to minimise the overall error to which the minority class contributes very little; (b) They assume that there is equal distribution of data for all the classes; (c) They also assume that the errors coming from different classes have the same cost (Guo et al., 2008; Ganganwar, 2012; Loyola-González et al., 2017).

Given these challenges, accurately classifying minority events (e.g., severe crashes) requires specialised techniques to address class imbalance.

To mitigate class imbalance issues, two primary approaches have been developed:(a) Cost-sensitive learning (at the algorithm level) and (b) Sampling technique (at the data level) (Guo et al., 2008). At the algorithmic level, solutions try to adapt existing classifier learning algorithms to strengthen learning concerning the small class. Two common methods, Boosting and Cost-sensitive learning, are used in this approach (Guo and Viktor, 2004; Maheshwari et al., S., 2011; Sonak et al, 2016). In particular, the goal of cost-sensitive learning is to minimise the cost of misclassification. Cost-sensitive learning methods enforce emphasis on the minority class by manipulating and incorporating learning parameters such as data-space weighting class-dependent cost matrix and Receiver Operating Characteristics (ROC) threshold into conventional learning paradigms. These solutions alter the original class distribution at the data level, driving the bias towards the minority or positive class. They consist of resampling the original data set, either by over-sampling the minority class or by under-sampling the majority class, until the classes are approximately equally represented (Cieslak et al., 2008).

Synthetic data generation is a more recent and promising approach to handling imbalanced data. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling generate artificial minority class examples to enhance model learning (Menardi & Torelli, 2010). These methods improve generalisation and reduce overfitting risks associated with simple oversampling.

However, a critical limitation of these techniques is their reliance on Euclidean distance, which is not well-suited for datasets containing both numerical and categorical variables (Velleman & Wilkinson, 1993; Murphy et al., 2024; Selman et al., 2024). Euclidean distance assumes continuous numerical features, making it problematic for categorical variables without meaningful numerical relationships. Assigning arbitrary numerical values to categories can distort similarity measures, leading to inaccurate classifications.

In this paper, we introduce a resampling method for dealing with response class imbalance in supervised learning with predictor variables of any type, namely numerical, ordinal, nominal, and dichotomous variables. Both parametric and non-parametric models were employed to assess the effectiveness of the SONCA algorithm in improving predictive performance on imbalanced data. Specifically, the Logit and CART models were used to analyse crash severity based on the PTW-ISTAT datasets. The response variable, crash severity, is binary and exhibits a high degree of imbalance: injury crashes represent 98.18% of the total, while fatal crashes account for only 1.82%.

2. Synthetic Over-sampling for Numerical and Categorical variables (SONCA) algorithm

Given a dataset comprising n observations, where n denotes the number of units in the sample. The dataset includes a response vector $Y_{n\times1}$ and a predictor matrix $X_{n\times q}$, where:

- The response vector $Y_{n\times1} = (y_1,y_2,...,y_n)'$, where each $y_i \epsilon \{0,\ 1, \dots, c, \dots, C\}$ representing the response class for observation i, and C+1 is the total number of response categories;
- The predictor matrix $X_{n\times q}$, where q is the total number of predictor variables, is composed of both numerical and categorical predictor variables. Accordingly, $X_{n\times q}$ is divided into two sub-matrices: $X_{Num} \in R^{n\times p}$: the sub-matrix of p numerical predictors; $X_{Cat} \in R^{n\times(p-q)}$: the sub-matrix of q-p categorical predictors.

Suppose that one of the categories in Y is underrepresented. Using SONCA, it is possible to generate a synthetic dataset to balance the response variable across all categories. Each observation, randomly drawn from the original dataset, is replaced by another observation selected from the entire predictor matrix. The selection process follows a probability function that is inversely proportional to the distance between the randomly drawn observation and all other observations in the predictor matrix. As a result, observations with smaller distances have a higher probability of being chosen.

In order to calculate the distance $d(x_i; x_{i^*})$ and thus use the SONCA algorithm, it is necessary to preprocess the matrix of q-p categorical predictors, X_{Cat}, using complete disjunctive coding. In other words, disjunctive coding consists of creating, for each variable, as many columns as there are levels (categories), where each column represents an indicator for each level. In this way, the submatrix X_{Cat}, which has q-p categorical predictors, is transformed into a submatrix X_{Binary} with n_{binary} dichotomous variables.

The synthetic observations are generated following the steps listed below:

1. Randomly extract the response category assumed by the $i-th$ observation $y_i = c$ for $c = 0, \dots, C$, assigning each category a uniform probability of 1/(C+1), where C+1 represents the total number of response categories;
2. Select an observation xi from the subset of the predictor matrix consisting of units that belong to the extracted response category, $X|(Y = c)$, with probability $p = 1/nc$ where nc is the number of observations in the given category;
3. Compute the distances $d(x_i; x_{i*})$ between x_i and all other observations $x_{i*} \in X - x_i$, where i *= 1,2, . . . , i − 1, i + 1, … n, using the weighted Euclidean distance (Greenacre, 2008; Schultz and Joachims, 2003). Weighted Euclidean distance assigns different weights to numeric and categorical variables, allowing for a balanced comparison between heterogeneous observations and preventing a single type of variable from dominating the distance calculation:

   $$d(x_i; x_{i*}) = \sqrt{\sum_j w_j \left(x_{i,j} - x_{i*,j}\right)^2}$$

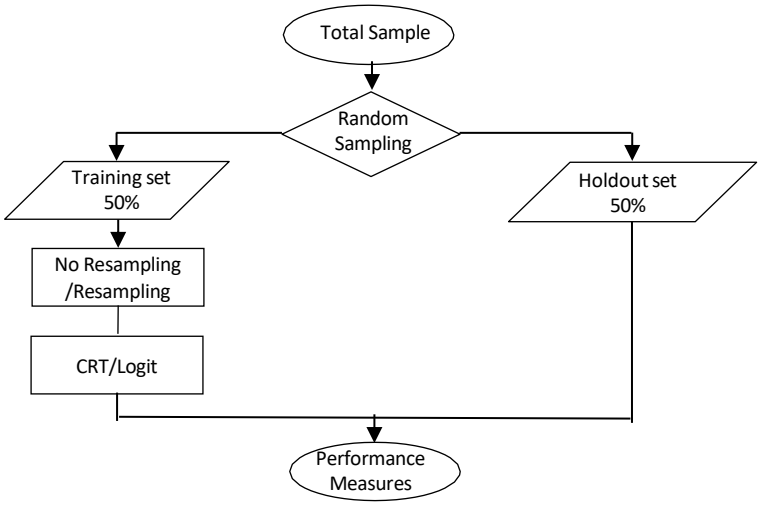
   where wj is the weighting coefficient, that is, it indicates the weight attributed to the variable j. If the j-th variable is of numeric type, $w_j = 1/\sigma^2$ is the inverse of the j-th variance, instead if the j-th variable is of categorical type then $w_j = 1/c_j$ where $c_j = \frac{n_j}{n_{binary}}$ with n_j number of categories of j-th variable and n_{binary} total number of categories;
4. Estimate a probability distribution for x_i using a monotonically decreasing function of the computed distances: $P(x_{i*}|x_i) \propto f(d(x_i, x_{i*}))$: the triangular and Gaussian distribution functions;
5. Randomly draw a new synthetic observation x^{SONCA} from the estimated probability distribution;
6. Repeat steps 1–5, from 1 to [(C+1)×m] times iterations, where m is the number of synthetic observations to generate for each response category.

3. Methodological process

To assess the effectiveness of the SONCA algorithm, the dataset was randomly split into two independent sets: a Training set (50% of observations) for model estimation, and a Holdout set (remaining 50%) for validation. The training set was resampled using SONCA to balance class distribution. Two models, Logit and CART, were applied

to three versions of the training data: (1) the original dataset, (2) SONCA-resampled data with a triangular distribution, and (3) SONCA-resampled data with a Gaussian distribution. Model performance was evaluated on the holdout set using standard metrics: True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Precision, F-measure, and G-mean.

Fig. 1 Phases of the learning and validation process



3.1. PTW – ISTAT data

The study data of PTW – ISTAT were the micro data provided by the Italian National Institute of Statistics (Istat).

Table 1. Distribution of datasets for PTW-Istat

Variable	**Code**	**Total**	**%**	**Variable**	**Code**	**Total**	**%**
Total		254’575	100.00	Alignment			
Severity				Tangent	Tang	103’775	40.76
Fatal	F	4’626	1.82	Curve	Cu	18’231	7.16
Injury	Inj	249’949	98.18	Intersection	Int	129’860	51.01
Crash Type				Other	Oth	2’709	1.06
Angle	Ang	107’140	42.09	Pavement			
Sideswipe	Sidsw	47’306	18.58	Dry	Dry	224’636	88.24
Rear-End	RE	27’843	10.94	Other	Oth	5’120	2.01
Head-On	Ho	18’171	7.14	Wet	Wet	24’819	9.75
Run-Off-The-Road	Ror	17’214	6.76	Lighting			
Hit Pedestrian	Hped	11’098	4.36	Day	D	195’710	76.88
Falling From The Vehicle	FfV	8’547	3.36	Night	N	58’865	23.12
Hit Stopped Vehicle	HsV	8’122	3.19	Area			
Hit Obstacle In Carriageway	HobsCar	6’112	2.40	Rural	R	29’999	11.78
Other	Oth	3’022	1.19	Urban	U	224’576	88.22
Road Type				Weather Conditions			
Motorway	Mot	4’228	1.66	Clear	Cl	223’509	87.80
Rural Municipal	RM	4’801	1.89	Foggy	Fo	926	0.36
Rural National	RN	9’438	3.71	Other	Oth	14’763	5.80
Rural Provincial	RP	11’532	4.53	Rainy - Snow	RS	15’377	6.04
Urban National	UN	9731	3.82				
Urban Provincial	UP	11585	4.55				
Urban Municipal	UM	203260	79.84				

The dataset comprises 254,575 road crashes involving at least one powered two-wheeler (PTW). Based on an initial exploratory analysis and a review of relevant literature, a subset of 7 categorical variables was selected from the 159

available fields, focusing on those most relevant to crash severity. The response variable, Severity crash, is highly imbalanced: only 1.82% of the crashes resulted in a fatal outcome, while the remaining 98.18% were classified as injury crashes.

4. Results

In the Table 2, the coefficients and good of fit of logit models, estimated on the three training datasets, are reported.

Table 2. Logit: parameter estimates and goodness of fit measures

Variables	Original Dataset	SONCA-Gaussian Dataset	SONCA-Triangular Dataset
Intercept	-4.643 (0.056)	-2.209 (0.088)	-1.605 (0.081)
Road type			
Mot	1.227 (0.124)	4.931 (0.333)	3.534 (0.206)
RM	1.253 (0.109)	3.028 (0.365)	2.597 (0.281)
RN	1.881 (0.069)	3.658 (0.141)	3.803 (0.165)
RP	2.048 (0.061)	3.968 (0.161)	2.659 (0.153)
UN	1.017 (0.095)	2.212 (0.172)	-0.25 (0.116)
UP	1.165 (0.082)	3.94 (0.149)	5.347 (0.27)
UM			
Alignment			
Cu	0.408 (0.064)	1.084 (0.126)	0.998 (0.134)
Int	-0.24 (0.051)	-0.47 (0.075)	-1.177 (0.073)
Oth	0.475 (0.158)	-2.273 (0.511)	0.413 (0.219)
Tg			
N	0.517 (0.046)	0.252 (0.059)	0.607 (0.067)
D			
Weather condition			
Fo	0.479 (0.277)	4.107 (0.672)	1.131 (0.678)
Oth	not significant	1.776 (0.169)	1.146 (0.162)
RS	not significant	0.881 (0.256)	-0.712 (0.213)
Cl			
Pavement			
Oth	-1.409 (0.263)	-3.539 (0.765)	not significant
Wry	-0.212 (0.123)	not significant	1.615 (0.193)
Dry			
Crash type			
FfV	-0.377 (0.145)	2.823 (0.362)	0.648 (0.19)
Ho	0.784 (0.065)	2.476 (0.101)	3.707 (0.107)
HobsCar	0.905 (0.097)	4.249 (0.126)	3.758 (0.176)
Hped	0.62 (0.1)	3.216 (0.095)	2.628 (0.09)
HsV	not significant	1.626 (0.132)	2.137 (0.127)
Oth	not significant	2.569 (0.457)	1.187 (0.163)
RE	-0.402 (0.083)	-0.788 (0.125)	-0.229 (0.122)
Ror	0.202 (0.076)	0.514 (0.177)	-1.589 (0.195)
Sidsw	-0.726 (0.082)	-2.81 (0.166)	-3.402 (0.227)
Angle			
Goodness of fit			
log likelihood null model	-11617.940	-11088.040	-11090.050
log likelihood full model	-10345.720	-4480.457	-3987.659
R^2 McFadden	0.110	0.596	0.640

Note: standard errors of the parameter estimates are reported in parenthesis; Italics indicates baseline indicator variable for the i-th variable.

Significant differences are observed in the coefficients, starting from the intercept values. In the model estimated on the non-resampled dataset, the intercept is equal to -4.643, while in the models estimated on the resampled datasets,

the values increase to -2.209 and -1.605. This highlights how, in the original dataset, the estimated model has a very low base probability of predicting the positive class that coincides with the rare class (fatal crashes), consistent with the strong imbalance of the distribution. The resampling has, therefore, contributed to re-balancing the distribution of the classes, increasing the ability to predict events belonging to the rare class. Furthermore, increases in the coefficients associated with different road typologies, alignments and weather conditions are observed, which are more marked in the resampled datasets. For example, the motorway category goes from 1.227 in the model with the non-resampled dataset to 4.931 and 3.534 in the SONCA-Gaussian and SONCA-Triangular models. These differences are reflected in a greater explanatory capacity of the models estimated on the resampled dataset, as confirmed by the values of R^2 McFadden's. For the model estimated on the original dataset, the R^2 is equal to 0.110, showing a low predictive power, while for the models estimated with the resampled dataset, R^2 is equal to 0.596 and 0.640, highlighting that the balanced datasets have a greater predictive power.

The tree diagrams estimated with the three datasets and the related histograms of the variables' importance are reported in the Fig. 2.

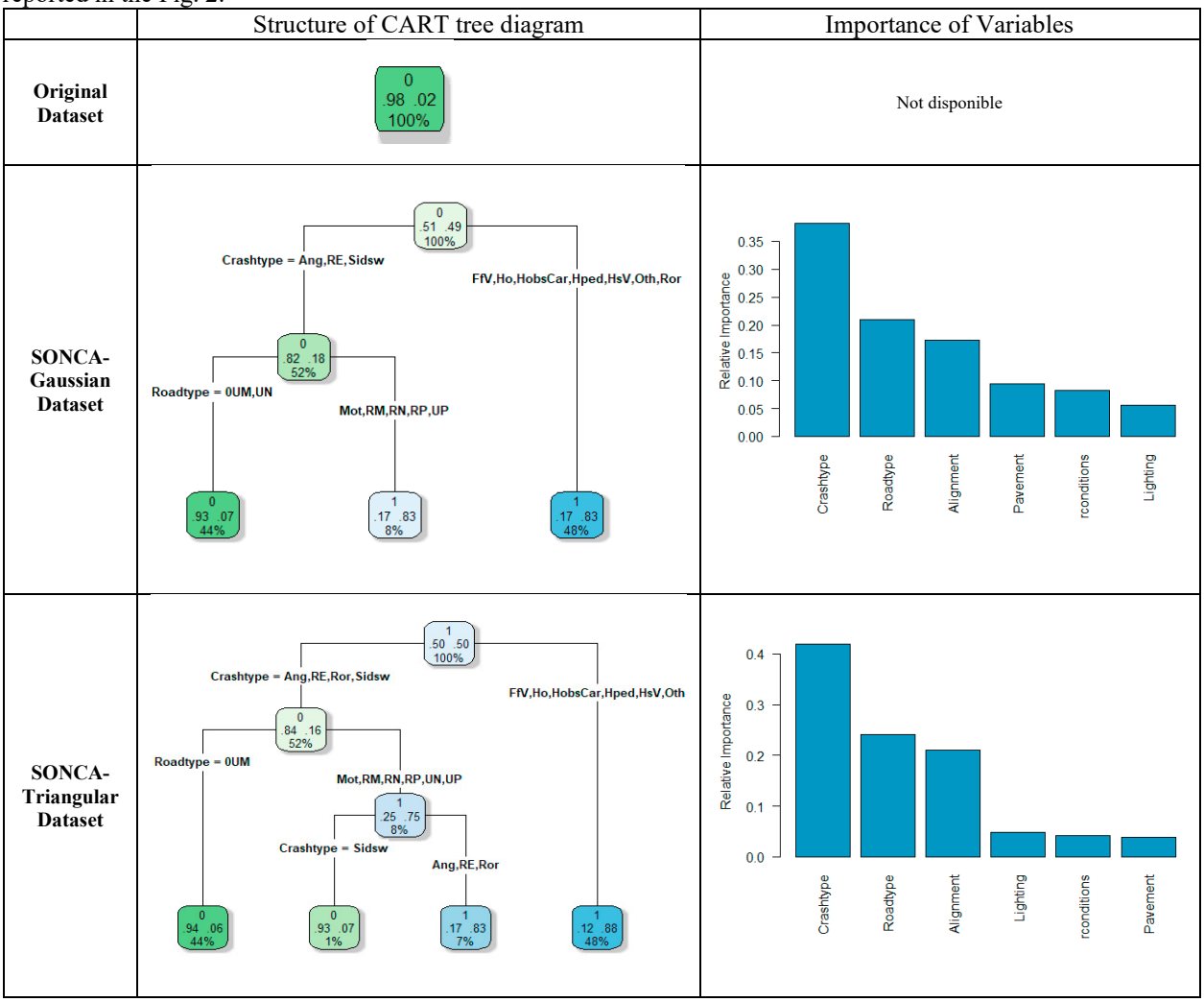


Fig. 2 Result of CART models

In particular, it is observed that the decision tree built using the original, highly imbalanced dataset is reduced to the root node only, which provides a constant prediction of the majority class (injury crashes). In this case, no

discriminative rule is learned to identify the minority class (fatal crashes), and it is impossible to compute variable importance since the algorithm performs no splits. In contrast, the trees built using the resampled datasets through the SONCA algorithm have three and four terminal nodes, respectively, allowing for improved predictive ability towards the rare class and for the calculation of variable importance based on the features used in the decision splits.
The performance measures calculated on the validation dataset, using the Logit and CART estimated on the original training dataset and resampled datasets using SONCA with Gaussian and triangular probability distributions, are reported in the Table 3.

Table 3. Measure of performance

Measure of perfomance	Logit			CART		
	Original	SONCA-Gaussian	SONCA-Triangular	Original	SONCA-Gaussian	SONCA-Triangular
TPR	0.000	0.629	0.576	0.000	0.717	0.675
TNR	1.000	0.724	0.724	1.000	0.613	0.658
FPR	0.000	0.276	0.276	0.000	0.387	0.342
FNR	1.000	0.371	0.424	1.000	0.283	0.325
Precision	0.000	0.040	0.037	0.000	0.033	0.035
Fmeasure	0.000	0.076	0.070	0.000	0.063	0.067
Gmean	0.000	0.675	0.646	0.000	0.663	0.666

Performance measures confirm the findings of the literature (King and Zeng, 2001; Chawla, 2003; Cieslak and Chawla, 2008; Menardi e Torelli, 2009), showing that classification algorithms such as Logit and CART, when estimated using the original unbalanced dataset, fail to predict the minority class, as evidenced by TPR and Precision values equal to 0 and a TNR of 1. Models estimated on resampled datasets using the SONCA algorithm improve their ability to classify the rare class. The TPR values are 0.629 and 0.576 for Logit, and 0.717 and 0.675 for CART. Both F-measure and G-mean improve significantly; for Logit, the highest values were obtained with the dataset balanced using the Gaussian distribution (0.076 and 0.675), while for CART, the highest values were achieved with the dataset balanced using the triangular distribution (0.067 and 0.666).

5. Conclusion

Both the parametric model (Logit) and the non-parametric model (CART) demonstrated an apparent inability to accurately predict crash severity when the distribution of the response variable is highly imbalanced. The findings are consistent with the existing literature: traditional classification algorithms tend to favour the majority class, completely overlooking the minority class—which, in the context of road safety, is precisely the most critical.
The analysis showed that resampling through the SONCA algorithm significantly enhances predictive performance, even under conditions of extreme imbalance. SONCA is an easy tool for data balancing; it acts as a pre-processing phase, allowing the learning system to receive the observations as if they belonged to a well-balanced data set. Therefore, it can be applied to any supervised classification method. SONCA is a balancing algorithm through which it is possible to obtain a new synthetic dataset to balance the response variable for each class. The synthetic dataset is obtained so that each observation, randomly extracted from the original dataset, is replaced by another observation selected within the entire predictor matrix. The random choice is performed by assuming a probability function, triangular or Gaussian, inversely proportional to the distance between the observation randomly extracted from the original dataset and all the observations of the predictor matrix.
Unlike other algorithms in the literature, SONCA makes it possible to treat numerical and categorical datasets.
However, the SONCA algorithm, for large datasets with a high number of observations and variables, could have a high computational cost. In particular, SONCA requires weighted Euclidean distance calculations for each observation relative to every other observation in the dataset. This step, repeated for each observation in the new dataset, can significantly increase the computational load. However, the high computational load is offset by the method's simplicity.

Acknowledgments

Data were partially analised within the Italian PRIN (Progetto di Rilevante Interesse Nazionale) 2022 PNRR - Research Project E53D23017300001 *FINGERTIPS*.

References

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y., 2020. Real-time crash prediction on expressways using deep generative models. Transportation Research Part C: Emerging Technologies, 117, 102697. https://doi.org/10.1016/j.trc.2020.102697

Chawla, N.V., Lazarevic, A., O. Hall L., Bowyer, K., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003.

Cieslak, D.A., & Chawla, N.V., 2008. Learning Decision Trees for Unbalanced Data. In Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I (pp. 241–256). Springer-Verlag.

Fernandes, E. R., & de Carvalho, A. C. 2019. Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. Information Sciences, 494, 141-154. https://doi.org/10.1016/j.ins.2019.04.052

Fiorentini, N., & Losa, M., 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), 61. https://doi.org/10.3390/infrastructures5070061

Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4).

Greenacre, M. (2007). Correspondence Analysis in Practice (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420011234

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G., 2008. On the Class Imbalance Problem. Natural Computation, ICNC'08, 4, 192-201. https://doi.org/10.1109/ICNC.2008.871

Han, W., Zhou, X., Zhang, A., Zhao, L., & Pan, G., 2024. AE-TabNet: A Two-stage Framework for Traffic Accident Severity Prediction with Imbalanced Data. In Proceedings of the 2024 2nd International Conference on Frontiers of Intelligent Manufacturing and Automation. https://doi.org/10.1145/3704558.3704581

He, H., & Garcia, E. A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9). https://doi.org/10.1109/TKDE.2008.239

King G, Zeng L, 2001. Logistic regression in rare events data. Political Anal 9:137–163

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P., 2006. Handling imbalanced dataset: A review. GESTS International Transactions on Computer Science and Engineering, 30.

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L., 2023. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. Knowl Inf Syst 65, 31–57 (). https://doi.org/10.1007/s10115-022-01772-8

Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. Neurocomputing, 175, 935-947. https://doi.org/10.1016/j.neucom.2015.04.120

Laureshyn, A., & Varhelyi, A., 2018. The Swedish Traffic Conflict Technique - Observer's manual.

Maheshwari, S., Agrawal, J., & Sharma, S., 2011. A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms. International Journal of Scientific & Engineering Research, 2(7), 1-5.

Menardi, G., & Torelli, N., 2012. Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-012-0295-5

Murphy, K., López-Pernas, S., & Saqr, M., 2024. Dissimilarity-Based Cluster Analysis of Educational Data: A Comparative Tutorial Using R. In: Saqr, M., & López-Pernas, S., eds) Learning Analytics Methods and Tutorials. Springer, Cham. https://doi.org/10.1007/978-3-031-54464-4_8

Sonak, A., Patankar, R., & Pise, N., 2016. A new approach for handling imbalanced dataset using ANN and genetic algorithm. In 2016 international conference on communication and signal processing (ICCSP), pp. 1987-1990). IEEE. https://doi.org/10.1109/ICCSP.2016.7754521

Ndour, M., & Dossou-Gbété, S., 2012. Analysis of classification performance in imbalanced datasets. Journal of Applied Statistics, 39(5), 1021-1032.

Hancock, J.T., Khoshgoftaar, T.M. & Johnson, J.M., 2023. Evaluating classifier performance with highly imbalanced Big Data. J Big Data 10, 42 https://doi.org/10.1186/s40537-023-00724-5

Oh, J., Kim, E., Kim, M., & Choo, S., 2010. Development of conflict techniques for left-turn and cross-traffic at protected left-turn signalized intersections. Safety Science, 48(4), 460-468.

Schultz, M., & Joachims, T. (2003). Learning a distance metric from relative comparisons. Advances in neural information processing systems, 16.

Selman, C. J., Lee, K. J., Ferguson, K. N., Whitehead, C. L., Manley, B. J., & Mahar, R. K., 2024. Statistical analyses of ordinal outcomes in randomized controlled trials: A scoping review. Trials, 25(1), 241. https://doi.org/10.1186/s13063-024-07779-8

Velleman, P. F., & Wilkinson, L., 1993. Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. The American Statistician, 47(1), 65-72.