Data complexity and classification accuracy correlation in oversampling algorithms

Joanna Komorniczak JOANNA.KOMORNICZAK@PWR.EDU.PL
Paweł Ksieniewicz PAWEL.KSIENIEWICZ@PWR.EDU.PL
Michał Woźniak MICHAL.WOZNIAK@PWR.EDU.PL
Department of Systems and Computer Networks
Wrocław University of Science and Technology

Editor: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak and Shuo Wang.

Abstract

Purpose: This work proposes the hypothesis that data oversampling may lead to dataset simplification according to selected data difficulty metrics and that such simplification positively affects the quality of selected classifier learning methods.
Methods: A set of computer experiments was performed for 47 benchmark datasets to make the hypothesis plausible. The experiments considered five oversampling methods, five classifiers, and 22 metrics for data difficulty assessment. The experiments aim to establish: (a) whether there is a relationship between resampling and change in the difficulty of the training data and (b) whether there is a relationship between changes in the values of training set difficulty metrics and classification quality.
Results: Based on the obtained results, the research hypothesis was confirmed. It was indicated which measures correlate with selected classifiers. The experiments showed the relationship between the change of assessed difficulty measures after oversampling and the classification quality of selected models.
Conclusion: The obtained results allow using the selected measures to predict whether a given oversampling method leads to favorable modifications of the learning set for a given type of classifier. Showed relationship between difficulty measures and classification will allow using the mentioned measures as a learning criterion. For example, guided oversampling can treat the modification of the learning set as an optimization task. During the oversampling process, no estimation of classification quality metrics will be required, but only an evaluation of the training set difficulty. This may contribute to the proposition of computationally efficient methods.
Keywords: oversampling, data complexity, imbalanced data, pattern classification

1. Introduction

The paper considers the problem of evaluating imbalanced data preprocessing. The unequal number of training examples in each class concerns most real-world classification tasks and has been of interest to the scientific community for more than thirty years. Initially, the inequality in the number of objects in each class and the fact that most canonical classifier learning methods cause the returned classification model to be biased toward the majority class was the main source of the difficulty of training classifiers on imbalanced data. Many methods based on data preprocessing and so-called algorithm-level solutions have been proposed to offset this disparity, such as using randomized over- and undersampling or simple

guided strategies. However, it is easy to show examples of conditional class distributions that, despite the disparity in the representation of the different fractions, do not pose any problem in building a high-quality model. In that case, any interference with the class distributions is unnecessary and will not improve classification quality.

Previous works have recognized that in addition to the disparity mentioned above, a more critical problem is the mutual distribution of minority and majority class data, their overlap, and the formation of small unrepresentative minority class clusters. Several taxonomies of the difficulty of minority class distributions were proposed. Mainly, they divide objects into safe objects (i.e., minority class objects forming homogeneous clusters) and unsafe objects (i.e., minority class objects surrounded by minority and majority class objects). The most popular taxonomy is based on the composition of an object's five nearest neighbors and classifies it in the case of an inhomogeneous neighborhood into a borderline, rare, or outlier category. Based on this approach, guided preprocessing methods began to be designed, which take into account, for example, object types during the oversampling process (Sáez et al.) or focus on one selected fraction, such as BorderlineSmote (Han et al., 2005).

However, it should be noted that most of the guided methods try to use some more or less legitimate heuristics. Some authors try to avoid using the "*rule of a thumb*" by treating preprocessing of imbalanced data as an optimization task. E.g., Khoshgoftaar et al. (2010) proposed using an evolutionary algorithm for the undersampling. García and Herrera (2009) treated the undersampling as a multi-criteria optimization problem. Also, several guided oversampling strategies were supported by metaheuristic algorithms, such as Li et al. (2020).

An additional problem is the lack of reliable metrics to assess the quality of the resulting model without information about the mutual validity of the minority and majority classes. Such information would simplify the problem by taking the expected value of the loss function (overall risk) as a criterion. Unfortunately, in most tasks, we do not have this information, and the most popular metrics can lead to unwarranted conclusions about the quality of the evaluated methods (Brzeziński et al., 2020; Stapor et al., 2021).

Unfortunately, there is a lack of explicit guidance that attempts to indicate whether a given preprocessing method is worth using for a given decision problem and, if so, for which learning method will a resulting dataset lead to a classifier of acceptable quality. The typical structure of computer experiments aiming to evaluate the proposed preprocessing method is to test it on a set of benchmark databases for several classifier learning methods. Such a process is computationally costly, especially if preprocessing is treated as an optimization task. For example, when using evolutionary algorithms, each potential solution requires quality assessment, i.e., learning a classifier and estimating its quality.

In this paper, we will consider whether it is possible to use data complexity measures to assess the difficulty of a decision problem before and after data preprocessing and indicate the relationship between the chosen measure and the target classification model. In this work, we want to make plausible the hypothesis that *it is possible to indicate, without training a classifier, whether a given method of data preprocessing leads to such a modification of the training set, which will result in an increase in prediction quality for a given classifier trained on a data after preprocessing*. If this hypothesis is confirmed, we will obtain a simple tool to indicate which preprocessing method is suitable for which dataset and classifier type and can be used as an optimization criterion for guided preprocessing methods.

2. Related works

The scope of data complexity measures applications, according to the available literature, is rather wide, ranging from applications in signal data (Li et al., 2022), through the dominant application in meta-learning (Rivolli et al.), extending the context of difficult data analysis (Goethals et al., 2022), or applications in supporting the classification models of imbalanced data classification (Barella et al.).

However, the most heavily exploited research area using complexity measures is *meta-learning*, where a metric problem assessment is used to identify the processing pipeline that should be used to construct an optimal recognition model (Rivolli et al.). Particularly interesting is the work on spectrogram profiling Chinese liquor, the achievements of which allow for efficient recognition of fake vintage liquors (Zhang et al., 2022), rejecting not promising recognition strategies through meta-learning before building appropriate models. Complexity measures are most often used as meta-features for a variety of *AutoML* solutions (Alcobaça et al., 2020), being the basis for learning appropriate problem representations (Rakotoarison et al., 2021), or in practical applications such as selecting a model that takes into account prior risk knowledge of construction accidents (Li et al., 2021).

Assessing the problem's difficulty often appears in the literature as a tool to extend the analysis of results. This applies to research in the field of digital image processing, when assessing the impact of the thermal image scale on the difficulty of the classification task solved by LSTM networks (Reuß et al., 2021), and to the issue of model explainability, when assessing the impact of increasing the explainability of the model on changes in problem difficulty was measured by the $F1v$, $N3$ and $L1$ measures (Goethals et al., 2022). There is also a noticeable trend taking into account changes in the complexity of the problems after the resampling phase for the imbalanced data classification (Kong et al., 2019).

The primary scope of research on imbalanced data classification is based on evaluating the relationship between the resampled training set and the test set (Dogo et al., 2021). As part of the research on the correlation of measures with the quality of recognition in an imbalanced environment, (Barella et al.) proposed versions of metrics adapted to the disturbed prior distribution. Practical applications in such environments allow determining the size of the synthetic set in oversampling or prior recommendation of the most appropriate recognition methods for a given problem (Costa et al., 2020). It is worth mentioning (Santos et al., 2022) where the authors evaluated the difficulty of simultaneous class imbalance and overlapping.

As an issue integrating all areas of research in data complexity measures, the study of the relationship between model quality and classification task difficulty can be distinguished. The base analyses, conducted both on small collections of benchmark data (Camacho-Urriolagoitia et al., 2022), on dimensionality reduction (Morán-Fernández et al.) and imbalanced sets (Barella et al.) initially confirm the possibility of making it plausible. This provides the basis for conducting research on selecting preprocessing models and optimizing their configuration, an example of which may be the work of (Bartz et al., 2021).

A set of data complexity measures is presented in Lorena et al. (2019). This work describes 22 measures, divided into six categories and implemented in *ECoL* package for *R*, *DCoL* package for *C++* and in *problexity* package for *Python* language. All the measures are presented in Table 1. The original implementation of measures with a single asterisk

(*) was slightly modified by authors of *ECoL* package, and for those with a double asterisk (**), authors have selected the required parameters.

Table 1: Utilized measures of classification problem complexity assessment

CATEGORY	MEASURE	SYMBOL
Feature-based	Maximum Fisher's discriminant ratio	$F1$*
	Directional vector maximum Fisher's discriminant ratio	$F1v$
	Volume of overlapping region	$F2$
	Maximum individual feature efficiency	$F3$
	Collectove feature efficiency	$F4$
Linearity	Sum of error distance by linear programming	$L1$
	Error rate of linear classifier	$L2$
	Non linearity of linear classifier	$L3$**
Neighborhood	Fraction of borderline points	$N1$
	Ratio of intra/extra class NN distance	$N2$
	Error rate of NN classifier	$N3$
	Non linearity of NN classifier	$N4$**
	Fraction of hyperspheres covering data	$T1$*
	Local set average cardinality	LSC
Network	Density	$density$**
	Clustering Coefficient	$ClsCoef$**
	Hubs	$Hubs$**
Dimensionality	Average number of features per points	$T2$*
	Average number of PCA dimensions per points	$T3$
	Ratio of the PCA dimension to the original dimension	$T4$
Class imbalance	Entropy of class proportions	$C1$
	Imbalance ratio	$C2$

3. Experimental evaluation

We conducted a series of experimental studies to prove the hypothesis presented in the introduction by answering the following research questions:

RQ1: Do the known oversampling methods affect the change in the value of the data difficulty measures?

RQ:2 Is there a relationship between the change in the value of the selected difficulty measure caused by oversampling the training data and the classification quality of the chosen models?

3.1. Experimental setup

The experiments were implemented in the Python programming language using the *scikit-learn* (Pedregosa et al., 2011), *imbalance-learn* (Lemaître et al., 2017) libraries and the *problexity* module (Komorniczak and Ksieniewicz, 2022). The experiments examined the relationships between all data difficulty measures and five classification quality metrics: balanced accuracy (BAC), *F1 score*, area under the curve (AUC), recall, and precision.

Forty-seven publicly available binary data sets were used for the classification task. The tested problems were characterized by a different imbalance ratio, a different number of instances, and a different number of features defining the problem space. The difficulty of the examined data sets was also analyzed in the context of the classification problem complexity to confirm the diversity of the selected collection.

The evaluation was performed with 5×2 cv protocol (Stapor et al., 2021). The training set was subjected to resampling, then the measures of problem difficulty were calculated, and the classifier was trained using post-resampling data. The classification quality was tested on the test data without resampling. Each experimental loop was repeated ten times to minimize the influence of data bias, initializing classifiers and resampling methods with different random states. Due to the high imbalance ratio in some datasets, not all oversampling algorithms can generate synthetic data. In the event of an error in the resampling procedure, the dataset was not considered during the analysis. The experiments considered five oversampling methods and five classification algorithms.

3.2. Results and lessons learned

Let's present the results of the experiments and answer the research questions.

3.2.1. **RQ1:** Do the known oversampling methods affect the change in the value of the data difficulty measures?

The first experiment investigated the effect of oversampling on the data difficulty. Figure 1 shows the mean values of the measures with the standard deviation in all tested sets before and after resampling. In order to improve the readability of the chart, it shows only two resampling methods: *ROS* and *SMOTE*, and the base values, without modifying the dataset. Conclusions for the remaining resampling methods do not differ from those presented below. Nevertheless, a summary for all tested methods has been included in supplementary materials[1].

For the *feature-based* measures, only the $F1$ metric shows a substantial change. This measure describes the overlapping of values across classes. Adding synthetic instances to the set has a much smaller impact on changing the remaining criteria in this category. In the case of the $F1v$ measure, slight changes are observed, most often concerning the increased problem difficulty. The $F2$ measure examines the classes' maximum and minimum values of instance characteristics. The presented oversampling methods generate synthetic instances within the distribution of a given class, so the values taken into account and thus the metric's value will remain unchanged. The $F3$ and $F4$ measures test the ability of a single feature (in the case of $F3$) or sets of features ($F4$) to separate problem classes. Minor changes appearing in these metrics' values usually concern an increase in difficulty. They will result from losing the feature's ability to separate problem classes after adding synthetic samples.

Each metric shows a data difficulty increase in the category of *Linearity* measures. The measures in this category test the linear separability of the problem. Assuming the class areas in the original set overlap, generating synthetic instances will increase the error of the *SVM Linear Classifier*. The mistakes made by the classifier will consequently increase the difficulty of the problem in the area of these metrics.

1. https://github.com/w4k2/complexity/blob/main/LIDTA/suplementary.pdf

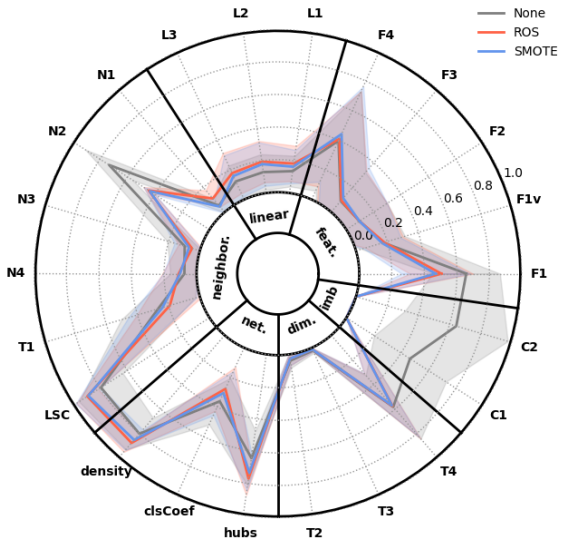

Figure 1: Resampling effect on complexity measures

The *neighborhood* measures include $N1$-$N4$, $T1$ and LSC metrics. $N1$ describes the *fraction of borderline points* based on the *minumum spanning tree* spread over the points. Out of the analyzed oversamplers, only ROS causes an increase in this measure. The metric is determined by the number of edges in the MST between instances of opposite classes. If the resampled points are at the same location of a multidimensional space, both will be considered as instances lying on the class area boundary. This behavior will be reflected in the increase of metric value. The remaining oversampling algorithms will have little effect on the MST covering the problem instances. The largest metric value change is visible for $N2$ – determined by the distance to the nearest neighbor of the same and opposite class. Out of tested oversampling methods, a decrease in this measure is visible, which occurs due to the generation of the synthetic samples in the vicinity of the original ones. The intra-class distances are decreasing, affecting the metric value. The $N3$ measure decreases for the same reason – this measure is expressed by the error rate of the NN classifier. The $N4$ measure will increase for problems where the class areas overlap. The $T1$ metric value will decrease when the hyperspheres generated for synthetic instances cover more instances of the problem than the hyperspheres for the original points. The last measure in this category – LSC – measures the distances to the closest neighbor of the sample and the closest instance of the opposing class. Neighbors closer to the instance than the closest enemy are considered. The value of this metric usually increases as a result of oversampling. When classes have an overlap region, the generated instances may lie close to the opposite class instance.

We generated a graph based on the normalized *Gower distances* between instances to determine the *network measures*. For *density* and *hubs*, the difficulty values after resampling increases. It may result from generating a graph where synthetic instances, lying at a large distance from each other, will not be connected by edges with the rest of the class instance. This will happen when the minority class points are spread. In the case of *clfCoef* measure, the connections between a given instance's neighbors are taken into account, divided by the number of possible connections between the neighbors, depending on the size of the

neighborhood. In the case of a large dispersion of class instances and the lack of edges in the graph, their absence will not negatively affect the measure.

For the *Dimensionality* measures, there are no notable differences between the set before and after resampling. However, the most visible change is for the *class imbalance* measures – the oversampled classes will be equally numerous.

Answering the **RQ:1**, the known oversampling methods affect the value of the problem difficulty measures. In *Feature based* category, there is a substantial change in measure describing the overlapping of values across classes. In *Linearity* measures, there is a stable increase in difficulty since synthetic samples reduce the problem's linear separability. *ROS* is also increasing the $N1$ *Neighborhood* metric, while all oversampling methods are highly influencing the $N2$ and $N3$ value, describing the distance to the nearest neighbor of the same and opposite class and *NN* classifier error. There is also an increase in *density* and *hubs* measures of *Network* category and a complete reduction of all metrics from *Class imbalance* category of the main assumptions of used balancing techniques.

3.2.2. **RQ:2** Is there a relationship between the change in the value of the selected difficulty measure caused by oversampling the training data and the classification quality of the chosen models?

The second experiment analyzed the problem difficulty and the classification quality after oversampling.

The Figure 2 presents the data difficulty expressed in the $F1$ and $L3$ measures and the classification quality for all tested classifiers. The presented values were subjected to a standard normalization. The Figure shows the results for *SMOTE* oversampling. A similar summary for other measures of difficulty and other oversamplers is available in the supplementary materials.

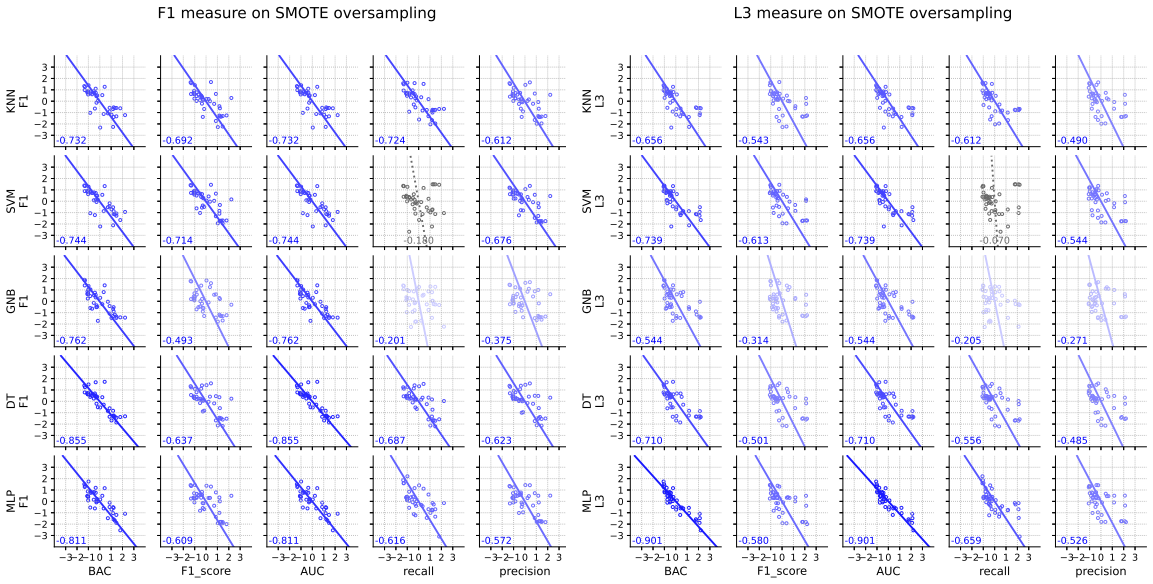

Figure 2: Correlation of F1 and L3 measures. The classifiers are presented in rows, while the columns show the classification quality metrics.

The values on the bottom of each chart show the *Pearson correlation coefficient* value. The line visible in each of the graphs was determined using linear regression, and its color depends on the value of the correlation coefficient, where *blue* means a strong negative correlation, *red* – a strong positive. The closer the color is to white, the smaller the correlation between the values. If the correlation coefficient lies between -0.2 and 0.2, the line determined by linear regression becomes gray and dotted – the correlation is not significant. The color of points is determined by correlation coefficient in a analogous way.

In Figure 2 the strong negative correlation can be noticed for each classifier in the case of *BAC*, *F1 score* and *AUC* metrics. For *precision* and *recall* metrics, for *SVM* and *GNB* classifiers, the correlation coefficient has a less significant value. The negative correlation can be understood as the following relationship between the data difficulty and the quality of the classification: the more difficult the problem is (large values on the y axis), the lower the quality of the classification (low values on the x-axis), while for simpler data sets (low values on the y axis) the quality of the classification is higher (high values on the x-axis).

The experiment results are consistent with intuition – the classifiers achieve better results for sets in which the training data after resampling was characterized by low difficulty.

Another analysis concerned the relative values of the examined measures. For the equal division of the set into folds, the original training set was used directly to build the classification model. The obtained classification quality was the base value for the analyzed classification metrics. The baseline value of the difficulty was the measures calculated for the original data set. Baseline values were subtracted from the classification quality and problem difficulty after resampling. The obtained results, presented in a similar form in Figure 3, show the relative value of the difficulty and quality of the classification in relation to the data set without resampling. A summary of other difficulty measures and oversampling methods is available in supplementary materials.

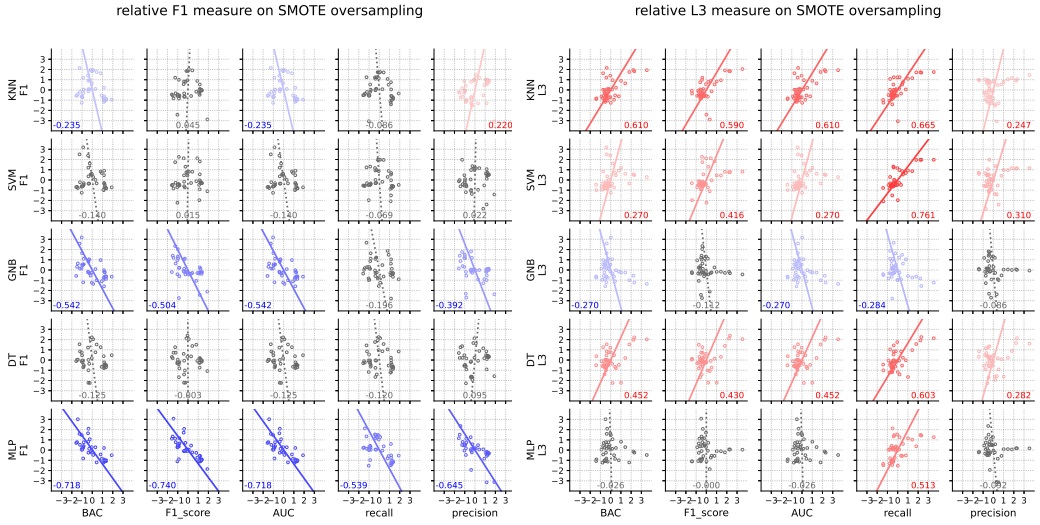

Figure 3: Correlation of relative F1 and L3 measures

In the case of the $F1$ measure, shown on the left side of the Figure 3, the strongly negative correlation coefficient was preserved only in the *MLP* classifier. For the *GNB* classifier, a negative coefficient for aggregated metrics can also be seen. In the case of the

other classifiers, the dependence of the relative change in the problem's difficulty and the classification quality is small. However, in the $L3$ measure, shown on the right side of the Figure, there was a significant change in the correlation coefficient. The absolute values presented in Figure 2 showed a strong negative correlation. Relative values, however, have a strong positive or close to zero correlation. Strong positive correlation is observed for the *KNN*, *SVM* and *DT* classifiers.

In the relative analysis, a low coefficient means that the decrease in the problem difficulty in the context of a given metric was associated with increased classification quality. A high coefficient implies that the increased difficulty in understanding a given metric is related to a rise in classification quality.

As presented in the first experiment – resampling has a positive effect on some measures ($F1$, $N2$, $clsCoef$, $C1$, $C2$) and negatively on others ($L1$-$L3$, LSC, $density$). Linear measures, including the presented $L3$, are measures for which resampling causes the difficulty value to increase. Usually, resampling leads to an increase in the quality of classification for imbalanced data, so despite the increase in difficulty, we observed an improvement in the classification result. A negative correlation is the expected result for measures where the problem's difficulty after resampling increases.

Figure 4 presents a summary of the correlation coefficient for the tested classifiers and difficulty measures. Each heatmap shows the correlations for a different classification quality metric. On the left side it is possible to observe an absolute values (example in the Figure 2), on the right side a relative values (example in the Figure 3). The values presented on the heatmap were averaged for all tested resampling methods.

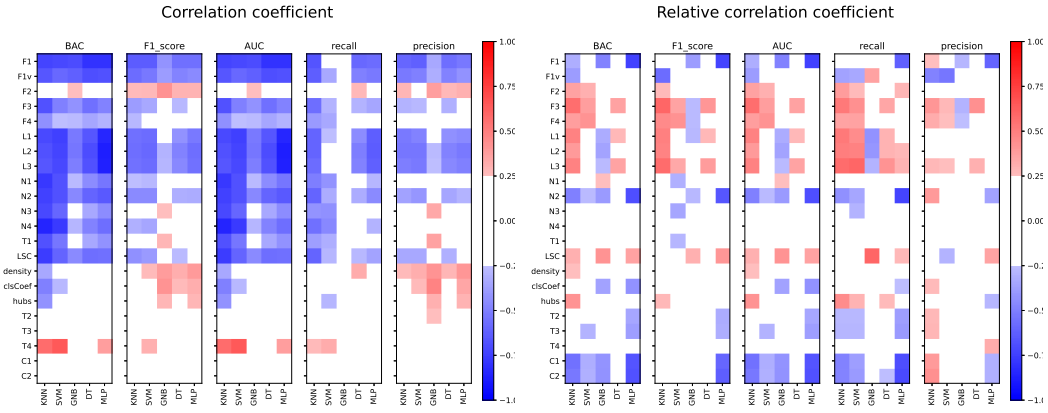

Figure 4: Mean non-relative and relative correlation coefficient

For absolute correlations, a significant part of the metrics shows a negative coefficient value, especially for *BAC* and *AUC* measures. For the relative correlation, some metrics from the feature-based group and all from the linear group show a positive correlation coefficient value, especially for the *KNN* and *SVM* classifiers.

Answering the **RQ:2**, there is a relationship between the change in the difficulty measure value caused by oversampling and the classification quality of the chosen models. The strong negative correlation is possible to observe, especially for aggregated metrics of *BAC*, *F1* and *AUC*. Base metrics dedicated to imbalanced data

also show the negative correlation, but with a lower coefficient for *SVM* and *GNB* models. This observation concludes that problems with lower complexity calculated according to established measures give a higher model's predictive quality.

Also, relative analysis was conducted, validating if a change in training data complexity leads to increasing the model quality. In the case of $F1$ measure, this trend was confirmed since, in relative comparison, the negative correlation is preserved. However, for the $L3$ measure – showing nonlinearity of linear classifier – a significant change in correlation occurred for some classifiers, showing an increase of predictive abilities proportional to problem difficulty. The final experiment emphasized that for an identified group of data complexity measures ($F1$, $N2$, $clsCoef$, $C1$, $C2$), there is a significant negative correlation between the data difficulty and model quality. It leads to the conclusion that reduction of data complexity – calculated within this identified group – leads to increased model's predictive quality.

4. Conclusion

The main objective of this study was to analyze the behavior of the data difficulty metrics before and after the oversampling procedure. It was hypothesized that there is a relationship between the change in the value of selected training set difficulty metrics and the classification quality of selected models.

The results of the experimental study supported the hypothesis. Particularly interesting behavior was observed for the $F1$ metrics for *MLP* and *GNB* and $L3$ metrics for *SVM*, *DT*, *KNN*, as well as for *GNB*. It turns out that there is a strong relationship between the change in their values for the training set of nuts and after oversampling and the quality of classification of the above models (evaluated on the validation set). This allows us to start working on quick methods for evaluating whether it is worth oversampling a given set for a given classification model. It also allows using the mentioned data difficulty metrics in guided oversampling design, where the preprocessing task is treated as an optimization task. Then the mentioned metrics can be used as an optimization criterion. The advantage of this approach is the speed of evaluating solutions during the optimization process, without the need to train classifiers in each iteration and estimate their quality. This is one of the directions of work to be undertaken by the authors of this article.

Acknowledgments

This work was supported by the Research Fund of Department of Systems and Computer Networks, Faculty of ICT, Wroclaw University of Science and Technology and by the Polish National Science Centre under the grant No. 2019/35/B/ST6/04442.

References

Edesio Alcobaça, Felipe Siqueira, Adriano Rivolli, Luís Paulo F Garcia, Jefferson Tales Oliva, André CPLF de Carvalho, et al. Mfe: Towards reproducible meta-feature extraction. *J. Mach. Learn. Res.*, 21(111):1–5, 2020.

Victor H Barella, Luis PF Garcia, Marcilio CP de Souto, Ana C Lorena, and Andre CPLF de Carvalho. Assessing the data complexity of imbalanced datasets. *Information Sciences*.

Eva Bartz, Martin Zaefferer, Olaf Mersmann, and Thomas Bartz-Beielstein. Experimental investigation and evaluation of model-based hyperparameter optimization. *arXiv preprint arXiv:2107.08761*, 2021.

Dariusz Brzeziński, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczęch. On the dynamics of classification measures for imbalanced and streaming data. *IEEE transactions on neural networks and learning systems*, 31(8):2868–2878, 2020.

Francisco J Camacho-Urriolagoitia, Yenny Villuendas-Rey, Itzamá López-Yáñez, Oscar Camacho-Nieto, and Cornelio Yáñez-Márquez. Correlation assessment of the performance of associative classifiers on credit datasets based on data complexity measures. *Mathematics*, 10(9):1460, 2022.

Afonso José Costa, Miriam Seoane Santos, Carlos Soares, and Pedro Henriques Abreu. Analysis of imbalance strategies recommendation using a meta-learning approach. In *7th ICML workshop on automated machine learning (AutoML-ICML2020)*, pages 1–10, 2020.

Eustace M Dogo, Nnamdi I Nwulu, Bhekisipho Twala, and Clinton Aigbavboa. Accessing imbalance learning using dynamic selection approach in water quality anomaly detection. *Symmetry*, 13(5):818, 2021.

S. García and F. Herrera. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, 17(3):275–306, 2009.

Sofie Goethals, David Martens, and Theodoros Evgeniou. The non-linear nature of the cost of comprehensibility. *Journal of Big Data*, 9(1):1–23, 2022.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

Taghi M. Khoshgoftaar, Naeem Seliya, and Dennis J. Drown. Evolutionary data analysis for the class imbalance problem. *Intell. Data Anal.*, 14(1):69–88, jan 2010. ISSN 1088-467X.

Joanna Komorniczak and Paweł Ksieniewicz. problexity – an open-source python library for binary classification problem complexity assessment. —, 2022.

Jiawen Kong, Wojtek Kowalczyk, Duc Anh Nguyen, Thomas Bäck, and Stefan Menzel. Hyperparameter optimisation for improving classification under class imbalance. In *2019 IEEE SCCI*, pages 3072–3078. IEEE, 2019.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset complexity assessment based on cumulative maximum scaled area under laplacian spectrum. *Multimedia Tools and Applications*, pages 1–17, 2022.

Min Li, An Xiong, Lei Wang, Shaobo Deng, and Jun Ye. ACO resampling: Enhancing the performance of oversampling methods for class imbalance classification. *Knowledge-Based Systems*, 196:105818, 2020. ISSN 0950-7051.

Xin Li, Rongchen Zhu, Han Ye, Chunxiao Jiang, and Abderrahim Benslimane. Metainjury: Meta-learning framework for reusing the risk knowledge of different construction accidents. *Safety science*, 140:105315, 2021.

Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34, 2019.

Laura Morán-Fernández, Verónica Bólon-Canedo, and Amparo Alonso-Betanzos. How important is data quality? best classifiers vs best features. *Neurocomputing*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Herilalaina Rakotoarison, Louisot Milijaona, Andry Rasoanaivo, Michèle Sebag, and Marc Schoenauer. Learning meta-features for automl. In *International Conference on Learning Representations*, 2021.

Felix Reuß, Isabella Greimeister-Pfeil, Mariette Vreugdenhil, and Wolfgang Wagner. Comparison of long short-term memory networks and random forest for sentinel-1 time series based large scale crop classification. *Remote Sensing*, 13(24):5000, 2021.

Adriano Rivolli, Luís PF Garcia, Carlos Soares, Joaquin Vanschoren, and André CPLF de Carvalho. Meta-features for meta-learning. *Knowledge-Based Systems*.

José A Sáez, Bartosz Krawczyk, and Michał Woźniak. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*.

Miriam Seoane Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, Carlos Soares, Szymon Wilk, and João Santos. On the joint-effect of class imbalance and overlap: a critical review. *Artificial Intelligence Review*, pages 1–69, 2022.

Katarzyna Stapor, Paweł Ksieniewicz, Salvador García, and Michał Woźniak. How to design the fair experimental classifier evaluation. *Applied Soft Computing*, 104:107219, 2021.

Yinsheng Zhang, Zhengyong Zhang, and Haiyan Wang. A unified classifiability analysis framework based on meta-learner and its application in spectroscopic profiling data. *Applied Intelligence*, 52(8):8947–8955, 2022.