Multi-Objective Sample Weighting for Imbalanced Data Classification

Szymon Wojciechowski[0000−0002−8437−5592]

Department of Systems and Computer Networks,
Wrocław University of Science and Technology,
Wrocław, Poland
szymon.wojciechowski@pwr.edu.pl

Abstract. An unequal sampling of classes in imbalanced datasets may lead to overfitting the model towards the majority class. Moreover, especially in probabilistic classifiers, the lack of prior probability compensation may negatively impact the classification of the minority class. Therefore, an essential element of building such a model for imbalanced data classification is sample weighting providing a more reasonable distribution estimation. However, the most commonly used approach is assigning equal weights to dense samples and the outliers, which in consequence, may lead to incorrect distribution estimation. Hence, it can be assumed that there is a more suitable sample weighting method for probabilistic classifiers. This work employs a multi-objective optimization algorithm to assign weights regarding models' sensitivity, precision, and specificity, providing a better-suited solution for imbalanced data classification. The article defines an optimization procedure for the addressed problem, evaluates the proposed method with other *state-of-the-art* methods, and outlines possible further research directions.

Keywords: Classification · Imbalanced Data · Multi-Objective Optimization

1 Introduction

The classification of imbalanced data is one of the most popular topics in machine learning community. Such data are characterized by a particular difficulty regarding the uneven representation of classes. This may result in the underrepresentation of one of them, which leads to in poorly fitted models, especially in terms of the minority class detection in such tasks as fraud detection or medical diagnosis [3]. Among the most frequently used techniques to deal with imbalanced data, one can find algorithms for oversampling the minority class, undersampling of the majority class, or a *hybrid methods* [4]. However, regardless of the chosen method, the purpose of preprocessing is to change the *prior probability* of a given problem – which in the case of some classifiers can also be achieved by proper sample weighting.

Often, weighting is based on original data imbalance, which might not be a perfect strategy in some cases. For example, giving some weight to outlier observation as the original distribution can cause a decision boundary shift. This leads to the following research question - *is it possible to propose a better procedure of assigning weights for samples in imbalanced problems*?

2 Algorithm

The proposed solution for better assignment of weights is based on a multi-objective optimization algorithm. The considered task can be defined as an optimization problem, determining a classification model characterized by the highest quality. However, evaluation of algorithms for imbalanced data classification requires specific metrics, often formulated in a way, that the error in the minority class is equally included as one in the majority class.

Commonly used *Gmean* metric is defined as geometric mean of *sensitivity* and either *specificity* or *precision*. This overcomes the limitation of single-metric optimization, which (i) is often limited to only two out of three basic metrics and (ii) depends on their aggregation. Therefore multi-criteria optimization tasks may give certain advantages in this kind of problem. Defining several metrics, which are simultaneously maximized in the optimization process, results in creating a number of interdependent solutions converging to *pareto-optimal* solutions. One of advantages is, that models can go beyond limitations of aggregated metrics, but finally, only one solution can be used, which requires a procedure for selecting most useful model.

To define given problem formally it will be assumed that the solution $s = w_1, w_2, ...w_n$ belongs to feasible solutions space S and it is defined as a series of weights w corresponding to n samples in training set. Optimization objective is to maximize three basic metrics estimated on validation set:

$$\text{maximize} \quad f_{sns}(S), \quad f_{spc}(S), \quad f_{prc}(S) \tag{1}$$

where *sns*, *spc*, *prc* translate to *sensitivity*, *specificity* and *precision* respectively. Obtained weights are further used in *Gaussian Naive Bayes* classifier training for estimating μ_y and σ_y using maximum likehood of:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2}$$

Proposed algorithm – Multi-Objective Sample Weigthing (MOSWE) – is solving given problem using MOEAD with population size of 25 objects, limited to 10 000 evaluations. The reference directions were generated using the *Das-Dennis* algorithm [6]. To obtain a single model as a final classifier, the solution was selected using *Compromise Programming* multi-criteria decision making method.

3 Results

The experiments were conducted in stratified 5×2 CV protocol with paired t-test [1]. The proposed method was implemented in Python programming language supported by *scikit-learn* and *pymoo* packages [2]. Algorithm was compared with baseline classifier and other commonly used preprocessing techniques: *Random Undersampling* (RUS), *Cluster Centroids* (CC), *Random Oversampling* (ROS) and *Synthetic Minority Oversampling* (SMOTE) implemented in *imbalanced-learn* package [5]. For the evaluation, 21 datasets from the KEEL repository[1] were used.

Table 1. Balanced Accuracy results.

	NONE	MOSWE	RUS	CC	ROS	SMOTE
ecoli1	0.780	0.836	0.778	0.726	0.764	0.817
	4, 5	4	—	—	4	1, 4, 5
ecoli2	0.682	0.854	0.786	0.625	0.664	0.773
	4	1, 4, 5, 6	1, 4, 5	—	4	1, 4, 5
ecoli3	0.841	0.884	0.747	0.827	0.832	0.865
	—	1, 4, 5, 6	—	—	—	1, 4, 5
glass0123vs456	0.859	0.864	0.876	0.826	0.873	0.871
	4	4	4	—	1, 4	4
glass0	0.707	0.707	0.704	0.706	0.701	0.701
	—	—	—	—	—	—
glass1	0.657	0.660	0.656	0.643	0.659	0.653
	4	—	—	—	4	—
glass6	0.872	0.872	0.864	0.855	0.875	0.865
	—	4	—	—	4	—
haberman	0.575	0.559	0.608	0.561	0.621	0.600
	—	—	1, 2	—	1, 2	—
iris0	1.000	0.998	1.000	1.000	1.000	1.000
	—	—	—	—	—	—
new-thyroid1	0.987	0.979	0.973	0.948	0.977	0.978
	3, 4, 5, 6	4	4	—	4	4
new-thyroid2	0.987	0.979	0.974	0.955	0.976	0.976
	3, 4, 5, 6	4	4	—	4	4
page-blocks0	0.695	0.716	0.714	0.669	0.695	0.699
	4	1, 4, 5, 6	4	—	4	1, 4, 5
pima	0.714	0.713	0.718	0.719	0.726	0.731
	—	—	—	—	1, 2	1, 2, 3
segment0	0.896	0.914	0.890	0.860	0.887	0.892
	4, 5, 6	all	4	—	4	4
vehicle0	0.735	0.761	0.753	0.760	0.756	0.755
	—	1	1	1	1	1
vehicle1	0.677	0.697	0.670	0.627	0.671	0.675
	4	all	4	—	4	4
vehicle2	0.724	0.806	0.729	0.687	0.755	0.736
	4	all	4	—	1, 4	4
vehicle3	0.674	0.683	0.665	0.653	0.667	0.672
	—	all	—	—	—	—
wisconsin	0.966	0.965	0.965	0.964	0.966	0.966
	—	—	—	—	—	—
yeast1	0.519	0.529	0.529	0.518	0.515	0.523
	5	1, 4, 5, 6	—	5	—	1, 4, 5
yeast3	0.578	0.652	0.657	0.547	0.551	0.599
	4, 5	1, 4, 5, 6	—	—	—	1, 4, 5

Table 1 presents mean *balanced accuracy score* obtained by evaluated methods. Additionally, the list of algorithms from which referenced method was statistically better is presented below each score. It can be observed that MOSWE achieved significantly better results for the sets `segment0`, `vehicle1`, `vechicle2` and `vehicle3`, although it should be also noticed, that for the sets `ecoli2`, `ecoli3`, `page-block0`, `yeast1` and `yeast3` difference between MOSWE and RUS was not statistically significant. In the remaining datasets, algorithms did not outperformed each other, although most of the algorithms were achieving better score than CC. Interestingly, the baseline algorithm for the `new-thyroid1` and `new-thyroid2` was better then other methods, with the exception of MOSE, where there was no statistical significance. Moreover, RUS and ROS algorithms

[1] https://sci2s.ugr.es/keel/imbalanced.php

Table 2. Balanced Accuracy mean rank results.

NONE	MOSWE	RUS	CC	ROS	SMOTE
3.810	4.667	3.381	1.762	3.476	3.905

were significantly better than MOSWE only for the `haberman` dataset. Very good results of the proposed method, can be also observed in Table 2, which present the average ranking of all the methods. The observations made, allows to state that proposed method can significantly improve the results with regards to other *state-of-the-art* methods.

4 Conclusions

Conducted experiment shows that the proposed method achieves excellent results, and four datasets present the best *balanced accuracy score*. Also, there was only one dataset, for which other methods outperformed MOSWE.

It is also worth mentioning that the conducted research did not consider an important element in the classification of imbalanced data – the cost of an error made concerning the minority class is often higher than for the majority class. For example, *false-positive* test results in fraud detection are more dangerous than *true-negative*. Nevertheless, due to the pool of solutions obtained from the optimization algorithm, selecting the appropriate model may be adjusted with the expert knowledge regarding the said cost. In addition, the method is adaptable – as long as all solutions are stored, it is possible to change the model to one that corresponds to the significance of recognition of a given class set by an expert.

Acknowledgment

This work was supported by the Polish National Science Centre under the grant No. 2019/35/B/ST6/04442.

References

1. Alpaydin, E.: Combined 5 × 2 cv f test for comparing supervised classification learning algorithms. Neural Computation **11**(8), 1885–1892 (1999)
2. Blank, J., Deb, K.: Pymoo: Multi-objective optimization in python. IEEE Access **8**, 89497–89509 (2020)
3. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer International Publishing, Cham (2018)
4. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016)
5. Lemaître, G., Nogueira, F., Aridas, C.K.: 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)
6. Qingfu Zhang, Hui Li: MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. IEEE Transactions on Evolutionary Computation **11**(6), 712–731 (2007)