Generalization Bound for Imbalanced Classification

Evgeny Burnaev(✉)

Skoltech, Moscow 121205, Russia
e.burnaev@skoltech.ru
http://adase.group

Abstract. The oversampling approach is often used for binary imbalanced classification. We demonstrate that the approach can be interpreted as the weighted classification and derived a generalization bound for it. The bound can be used for more accurate re-balancing of classes. Results of computational experiments support the theoretical estimate of the optimal weighting.

Keywords: Imbalanced classification · Generalization bound · Resampling amount · Weighted classification

1 Introduction

In this paper, we consider the imbalanced binary classification, i.e. the case of two-class classification when one class (a minor class) has much less representatives in the available dataset than the other class (a major class). Many real-world problems have unavoidable imbalances due to properties of data sources, e.g. network intrusion detection and maintenance [1–3, 8], damage detection from satellite images [12, 13, 17], prediction and localization of failures in technical systems [4, 5, 7, 21, 22], etc. In these examples target events (diseases, failures, etc.) are rare and presented only in a small fraction of available data.

Often the main goal of the imbalanced classification is to accurately detect the minor class [11]. However, standard classification approaches (logistic regression, SVM, etc.) are often based on the assumption that all classes as equally represented [10]. As a consequence the resulting classification model is biased towards the major class. E.g., if we predict an event occurring in just 1% of all cases and the classification model always gives a "no-event" prediction, then the model error is equal to 1%. Therefore, the average accuracy of the classifier is high, although it can not be used for the minor class detection.

An efficient way to deal with the problem is to *resample* the dataset in order to decrease the class imbalance, as it was discussed in [6, 20]. In practice we can perform *oversampling*, i.e., add synthesized elements to the minor class, or perform *undersampling*, i.e., delete particular elements from the major class; or do the both types of samplings. There also exist other more delicate approaches to resampling.

A. N. Shiryaev et al. (Eds.): ICSM-5 2020, PROMS 371, pp. 107–119, 2021.
https://doi.org/10.1007/978-3-030-83266-7_8

Most of the resampling approaches takes as input the *resampling amount*, which defines how many observations we have to add or delete. In [6,20] they demonstrated that there is no "universal" choice of the resampling amount and the final classification accuracy significantly depends on a particular value we select for a dataset at hand.

The authors of [6,20] proposed to use either the cross-validation procedure [10] or the meta-learning procedure to select the resampling method and the resampling amount. However, these approaches are purely empirical and require to spend significant time for additional computational experiments due to the exhaustive search.

In this work we argue that the resampling approaches can be considered as a specific variant of the weighted classification: so to deal with a possible class imbalance when constructing a classifier we use a weighted error (risk) to stress the most important class. The question is how to select an appropriate weight value to up-weight the minor class. For that we estimate the theoretical generalization ability of the classifier with the weighted loss function and explore how it depends on the weighting scheme. We discuss how these findings can be used in practice when solving the imbalanced classification problem. In Sect. 2 we introduce the main notations and provide a theoretical problem statement. In Sect. 3 we prove the main result of the paper, namely, we obtain the generalization bound for the weighted binary classification and obtain an optimal weighting scheme. We propose the algorithm for the weighted classification based on the derived optimal weighting, and evaluate its empirical performance in Sect. 4. Results of computational experiments demonstrate usefulness of the proposed approach. We discuss conclusions in Sect. 5.

2 Problem Statement

Let us consider the formal binary classification problem statement, discuss how it can be interpreted as the weighted classification task in case we use the standard oversampling technique, and estimate the corresponding excess risk. Thanks to the estimate, we can characterize the influence of the weight (playing a role of the resampling amount) on the generalization ability of the classifier.

We denote by

- $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ a class of binary classifiers with a multi-dimensional input space $x \in \mathcal{X}$ and an output label space $\mathcal{Y}$. Here we consider $\mathcal{Y} = \{-1, +1\}$ for simplicity. E.g.

$$\mathcal{F} = \{f_{a,b} : f_{a,b}(x) = 2\mathbb{I}(\langle a, x\rangle + b \geq 0) - 1\};$$

- $\mathbb{P}$ a distribution on $\mathcal{X} \times \mathcal{Y}$;
- π a prior probability of a positive class, i.e.

$$\mathbb{P} = \pi\mathbb{P}_{x|y=+1} + (1-\pi)\mathbb{P}_{x|y=-1};$$

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ a training sample, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$;

– $\mathcal{R}_N(\mathcal{F})$ a Rademacher complexity of $\mathcal{F}$ [15]. Recall that the empirical Rademacher complexity of some family of functions $\mathcal{G}$ from $\mathcal{Z}$ to $[a, b]$ for a fixed sample $S = (z_1, \ldots, z_m)$ is equal to

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ are Rademacher random variables. Then the Rademacher complexity of $\mathcal{G}$ w.r.t. some distribution $\mathbb{P}$ on $\mathcal{Z}$ is defined as

$$\mathcal{R}_m(\mathcal{G}) = \mathbb{E}_{S \sim \mathbb{P}^m} \left[\hat{\mathcal{R}}_S(\mathcal{G}) \right].$$

We consider a zero-one loss function $L(\hat{y}, y) = \mathbb{I}_{\hat{y} \neq y}$. The theoretical risk is equal to $\mathbb{E}_{\mathbb{P}} L(f(x), y)$, so that the theoretically optimal classifier

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}} L(f(x), y).$$

The empirical risk has the form

$$\mathbb{E}_{\mathcal{D}} L(f(x), y) = \frac{1}{N} \sum_{j=1}^{N} L(f(x_j), y_j).$$

If we perform the oversampling the empirical risk can be represented as

$$\frac{1}{N} \sum_{j=1}^{N} u_j L(f(x_j), y_j),$$

where $u_j \geq 1$ is equal to the number of times the object x_j from the initial sample $\mathcal{D}$ is selected in the oversampling procedure (we count x_j as well). Thus the binary classification problem in case of the oversampling can be interpreted as a classification problem with a weighted empirical loss: we optimize the weighted empirical risk when training a classifier and measure its accuracy using a non-weighted theoretical risk.

Therefore, we define some (fixed) weighting function

$$u : (\mathcal{X} \times \mathcal{Y}) \to (0, +\infty).$$

The weighted empirical risk is equal to

$$\mathbb{E}_{\mathcal{D}} u(x, y) L(f(x), y) = \frac{1}{N} \sum_{i=1}^{N} u(x_i, y_i) L(f(x_i), y_i),$$

so that the empirical classifier

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} u(x, y) L(f(x), y). \tag{1}$$

We would like to derive an upper bound for the excess risk

$$\varDelta(\mathcal{F}, \mathbb{P}) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbb{P}} L(f(x), y) - \mathbb{E}_{\mathcal{D}} u(x, y) L(f(x), y) \right),$$

which characterizes a generalization ability of the classifier. In particular, high values of the excess risk means that the function class $\mathcal{F}$ is "too complex" for the considered problem.

There exist theoretical results about the classification performance when the classifier is trained with the weighted loss. E.g. in [9] a bayesian framework for imbalanced classification with a weighted risk is proposed, [19] investigated the calibration of asymmetric surrogate losses, [16] considered the case of cost-sensitive learning with noisy labels. The case of weighted risk for the one-dimensional classification based on probability density functions estimates is considered in [14].

However, to the best of our knowledge, there is no studied upper bound for the excess risk with explicitly derived dependence on the class imbalance value π and the weighting scheme $u(\cdot)$ that quantifies their influence on the overall classification performance.

3 Generalization Bound

To derive explicit expressions we use an additional assumption, namely, we consider

$$u(x, y) = (1 + g_{+}(w)) \mathbb{I}_{\{y=+1\}} + (1 + g_{-}(w)) \, \mathbb{I}_{\{y=-1\}}$$

for some positive weighting functions $g_{+}(w)$ and $g_{-}(w)$ of the weight value $w \geq 0$. We can tune w to re-balance the proportion between classes and decrease $\varDelta(\mathcal{F}, \mathbb{P})$.

Theorem 1. *With probability* $1 - \delta$, $\delta > 0$ *for* $\mathcal{D} \sim \mathbb{P}^N$ *the excess risk* $\varDelta(\mathcal{F}, \mathbb{P})$ *is upper bounded by*

$$\overline{\varDelta}(w) = 3 \left(g_{+}(w)\pi + g_{-}(w)(1 - \pi) \right) + \mathcal{R}_N(\mathcal{F}) + \left(2 + g_{+}(w) + g_{-}(w) \right) \alpha_N, \quad (2)$$

where $\alpha_N = \sqrt{\frac{\log \delta^{-1}}{2N}}$.

Proof. Let

$$\mathcal{L} = \{(x, y) \to L(f(x), y) : \, f \in \mathcal{F}\}$$

be a composite loss class. For any $L \in \mathcal{L}$ we get that

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} L - \mathbb{E}_{\mathcal{D}} u L &= \mathbb{E}_{\mathbb{P}} L - \mathbb{E}_{\mathbb{P}} u L + \mathbb{E}_{\mathbb{P}} u L - \mathbb{E}_{\mathcal{D}} u L \\ &\leq \mathbb{E}_{\mathbb{P}} |(1 - u) L| + (\mathbb{E}_{\mathbb{P}} u L - \mathbb{E}_{\mathcal{D}} u L). \end{aligned} \quad (3)$$

Since any $L \in \mathcal{L}$ is bounded from above by 1 for the first term in (3) we obtain

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} |(1 - u) L| &\leq \mathbb{E}_{\mathbb{P}} g_{+}(w) \mathbb{I}_{\{y=+1\}} + \mathbb{E}_{\mathbb{P}} g_{-}(w) \mathbb{I}_{\{y=-1\}} \\ &= g_{+}(w)\pi + g_{-}(w)(1 - \pi). \end{aligned} \quad (4)$$

Thanks to McDiarmid'd concentration inequality [15], applied to the function class $\mathcal{L}_u = \{uL : L \in \mathcal{L}\}$, with probability $1-\delta$, $\delta > 0$ for $\mathcal{D} \sim \mathbb{P}^N$ we get the upper bound on the excess risk

$$\sup_{L\in\mathcal{L}}(\mathbb{E}_{\mathbb{P}}uL - \mathbb{E}_{\mathcal{D}}uL) \le 2\mathcal{R}_N(\mathcal{L}_u) + \max[(1+g_+(w)),(1+g_-(w))]\alpha_N \le$$
$$\le 2\mathcal{R}_N(\mathcal{L}_u) + (2+g_+(w)+g_-(w))\alpha_N. \tag{5}$$

Let us find a relation between $\mathcal{R}_N(\mathcal{L}_u)$ and $\mathcal{R}_N(\mathcal{L})$. We denote by z_i a pair $z_i = (x_i, y_i)$. By the definition (see [15]) the empirical Rademacher complexity

$$\begin{aligned}\hat{\mathcal{R}}_{\mathcal{D}}(\mathcal{L}_u) &= \frac{1}{N}\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i=1}^N \sigma_i u(z_i)L(z_i)\\ &\le \frac{1}{N}\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i=1}^N \sigma_i L(z_i) + \frac{g_+(w)}{N}\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i:y_i=+1}\sigma_i L(z_i)\\ &+ \frac{g_-(w)}{N}\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i:y_i=-1}\sigma_i L(z_i)\\ &\le \hat{\mathcal{R}}_{\mathcal{D}}(\mathcal{L}) + \frac{g_+(w)}{N}\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i:y_i=+1}\sigma_i L(z_i)\\ &+ \frac{g_-(w)}{N}\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i:y_i=-1}\sigma_i L(z_i).\end{aligned}$$

For the zero-one loss

$$\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i:y_i=+1}\sigma_i L(z_i) \le \#\{i : y_i = +1\},$$

and

$$\mathbb{E}_\sigma \sup_{L\in\mathcal{L}_u}\sum_{i:y_i=-1}\sigma_i L(z_i) \le \#\{i : y_i = -1\}.$$

The Rademacher complexity

$$\begin{aligned}\mathcal{R}_N(\mathcal{L}_u) &= \mathbb{E}_{\mathcal{D}\sim\mathbb{P}^N}\hat{\mathcal{R}}_{\mathcal{D}}(\mathcal{L}_u)\\ &\le \mathbb{E}_{\mathcal{D}\sim\mathbb{P}^N}\left[\hat{\mathcal{R}}_{\mathcal{D}}(\mathcal{L}) + \frac{g_+(w)}{N}\#\{i : y_i = +1\} + \frac{g_-(w)}{N}\#\{i : y_i = -1\}\right]\\ &= \mathcal{R}_N(\mathcal{L}) + g_+(w)\pi + g_-(w)(1-\pi).\end{aligned} \tag{6}$$

Using the fact that $\mathcal{R}_N(\mathcal{L}) = \frac{1}{2}\mathcal{R}_N(\mathcal{F})$, substituting inequalities (4), (5) and (6) into (3), we get that

$$\Delta(\mathcal{F},\mathbb{P}) \le 3\left(g_+(w)\pi + g_-(w)(1-\pi)\right) + \mathcal{R}_N(\mathcal{F}) + \left(2+g_+(w)+g_-(w)\right)\alpha_N.$$

By collecting the terms with w in $\overline{\Delta}(w)$ (2) we get

$$g_+(w)\left(3\pi + \alpha_N\right) + g_-(w)\left(3(1-\pi)+\alpha_N\right),$$

and so minimizing this quantity w.r.t. w we can make the upper bound $\overline{\Delta}(w)$ tighter. For example, in case we set

$$g_+(w) = w \;\; g_-(w) = 1/w,$$

the optimal weight

$$w^{opt} = \sqrt{\frac{3(1-\pi)+\alpha_N}{3\pi+\alpha_N}} \approx \sqrt{\frac{1-\pi}{\pi}}, \tag{7}$$

where $\alpha_N \approx 0$ for $N \gg 1$. For such optimal w^{opt} we get

$$\overline{\Delta}^{opt} = \overline{\Delta}(w^{opt}) = 6\sqrt{\pi(1-\pi)} + \mathcal{R}_N(\mathcal{F}) + \alpha_N\left(2 + \frac{1}{\sqrt{\pi(1-\pi)}}\right).$$

Thus we obtain an estimate on how the weighting influences the classification accuracy: e.g. in the imbalanced case (when $\pi \approx 0$ or $\pi \approx 1$) for $N \gg 1$ by selecting the weight optimally we reduce the generalization gap almost to zero, as $\overline{\Delta}^{opt} \approx 0$; at the same time not optimal weight can lead to overfitting.

As we already discussed, under some mild modeling assumptions the binary classification problem in case of the oversampling can be interpreted as the classification problem with the weighted loss. Therefore not correctly selected resampling amount has the same negative effect as not optimal weight value for the classification with the weighted loss function. If we know the class imbalance, we can use the optimal value w^{opt} either to set the weight in case we use the weighted classification scheme, or as a reference value when selecting the resampling amount in case we use the oversampling approach—this should help to reduce the number of steps of the exhaustive search, used in [6,20].

4 Empirical Results

Let us perform an empirical evaluation of the obtained estimate (7). We expect that for the optimal weight value w^{opt} the classifier achieves better accuracy on the test when being trained by minimizing the weighted empirical loss. We consider the following protocol of experiments:

1. Consider different values of the weight $w \in W_K = \{w_1, \ldots, w_K\}$;
2. Train a classifier $f_w(x)$ by minimizing a weighted empirical loss (1) for the particular weight value $w = w_i$;
3. Estimate accuracy on the test set and find the weight $w^* \in W_K$ for which accuracy is the highest;
4. Compare the best obtained weight with the theoretical weight calculated using the formula (7).

We generated artificial datasets as pairs of 2D Gaussian samples with various means and covariance matrices and sample sizes, where each Gaussian sample

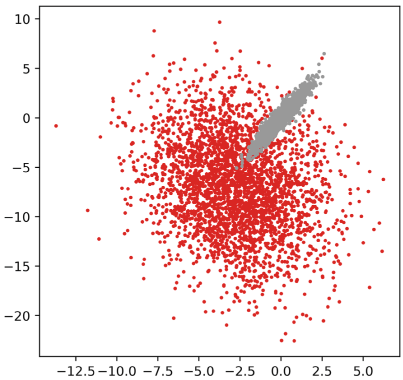


Fig. 1. Example of a toy dataset 1

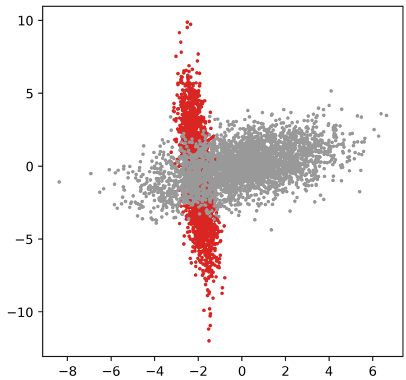


Fig. 2. Example of a toy dataset 2

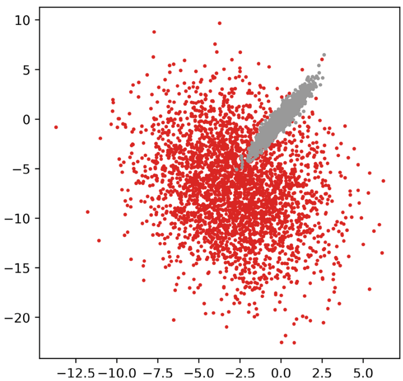


Fig. 3. Example of a toy dataset 3

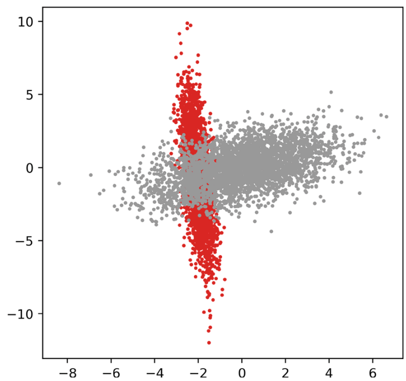


Fig. 4. Example of a toy dataset 4

corresponds to some class. Examples of artificial datasets 1, 2, 3 and 4 are shown in Figs. 1, 2, 3 and 4.

We took real datasets from Penn Machine Learning Benchmark repository [18]: we selected diabetes, german, waveform-40, satimage, splice, spambase, hypothyroid, and mushroom, that have various types of data and features. Due to multiclass data, we took class 0 for waveform-40 and splice, class 1 for satimage and class 2 for diabetes as a positive, whereas other classes were combined into a negative one.

To obtain a specific balance between classes in experiments, we used undersampling of an excess class. In this way we can get learning samples $\mathcal{D}$ corresponding to different values of π. Using this method, we varied the positive class share to test the dependence of the results on π.

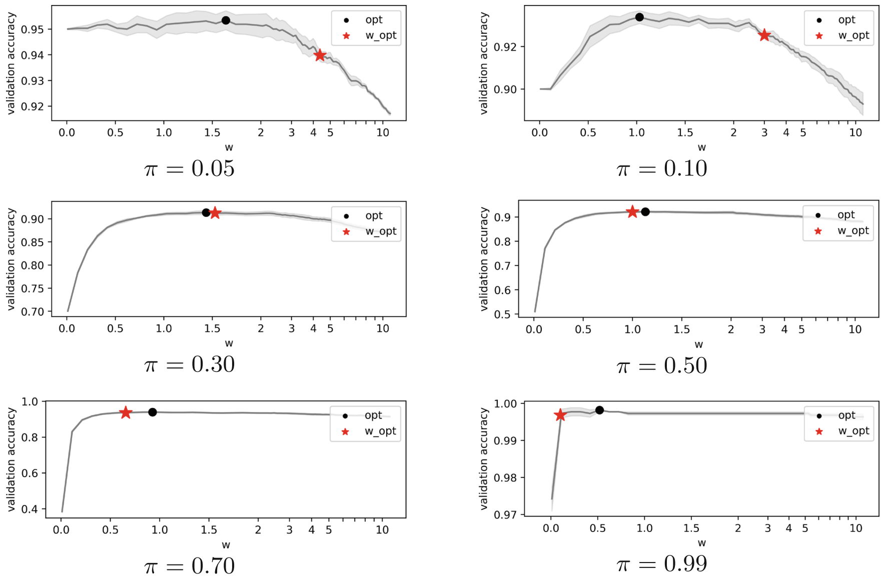


Fig. 5. w^* (black dot) vs. w^{opt} (red star) for the toy dataset 1 and different values of π

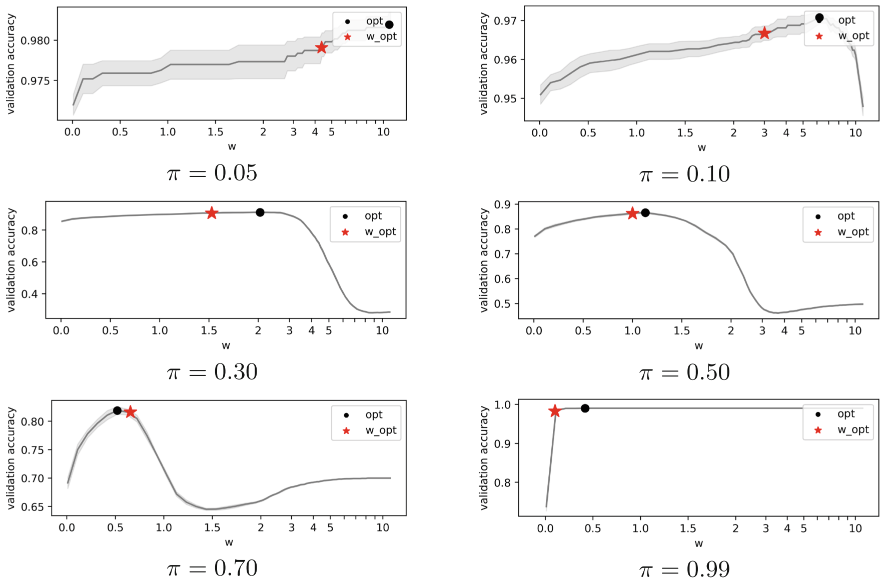
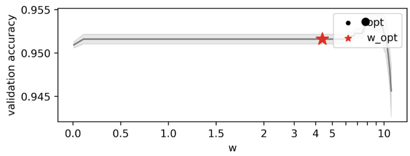
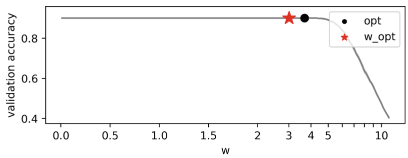


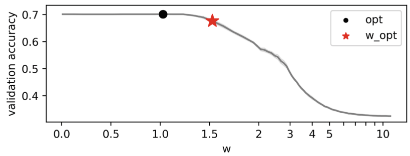
Fig. 6. w^* (black dot) vs. w^{opt} (red star) for the toy dataset 2 and different values of π



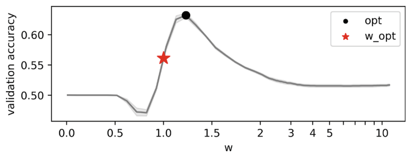
$\pi = 0.05$



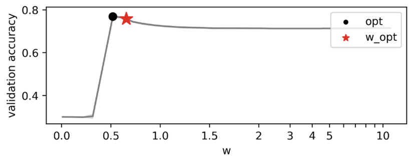
$\pi = 0.10$



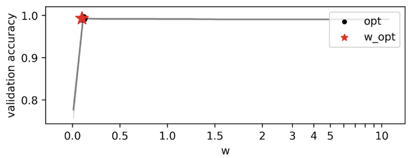
$\pi = 0.30$



$\pi = 0.50$

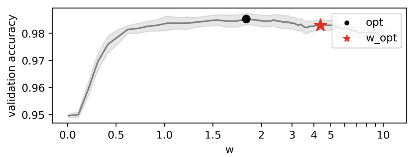


$\pi = 0.70$

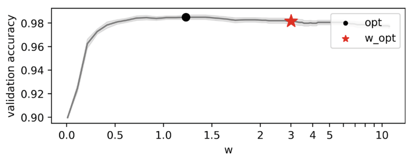


$\pi = 0.99$

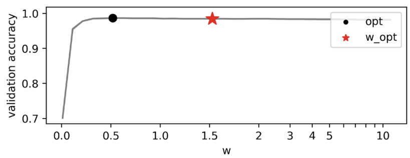
Fig. 7. w^* (black dot) vs. w^{opt} (red star) for the toy dataset 3 and different values of π



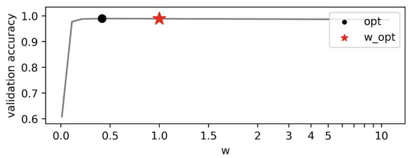
$\pi = 0.05$



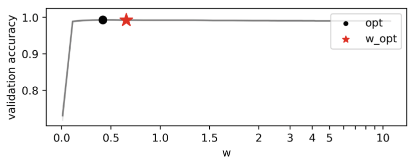
$\pi = 0.10$



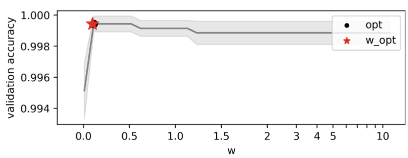
$\pi = 0.30$



$\pi = 0.50$



$\pi = 0.70$



$\pi = 0.99$

Fig. 8. w^* (black dot) vs. w^{opt} (red star) for the toy dataset 4 and different values of π

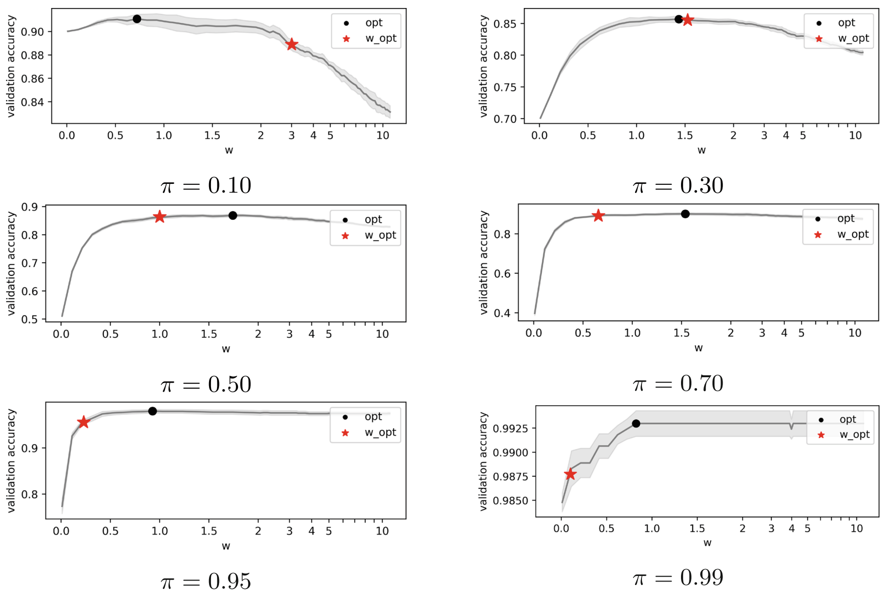


Fig. 9. w^* (black dot) vs. w^{opt} (red star) for the real dataset waveform-40 and different values of π

To measure the performance of the method, we conducted 5-fold cross-validation of a Logistic Regression classifier [10]. We provide examples of typical results on different datasets and for different positive class shares π: in Figs. 5, 6, 7, 8 there are results for toy datasets, and in Figs. 9 and 10 there are results for two real datasets—waveform-40 and hypothyroid). In particular, we show how the average validation accuracy depends on the weight w; we indicate empirically optimal values of w^* by black dots, and indicate theoretically optimal values of w^{opt} by red stars. We can observe that for most of the cases estimates w^* and w^{opt} agree rather well. Moreover, although the estimate w^{opt} is obtained under general conditions from the rather loose bound (2), still the classifier with $w = w^{opt}$ provides often quite good accuracy even if there is a big difference between w^* and w^{opt}.

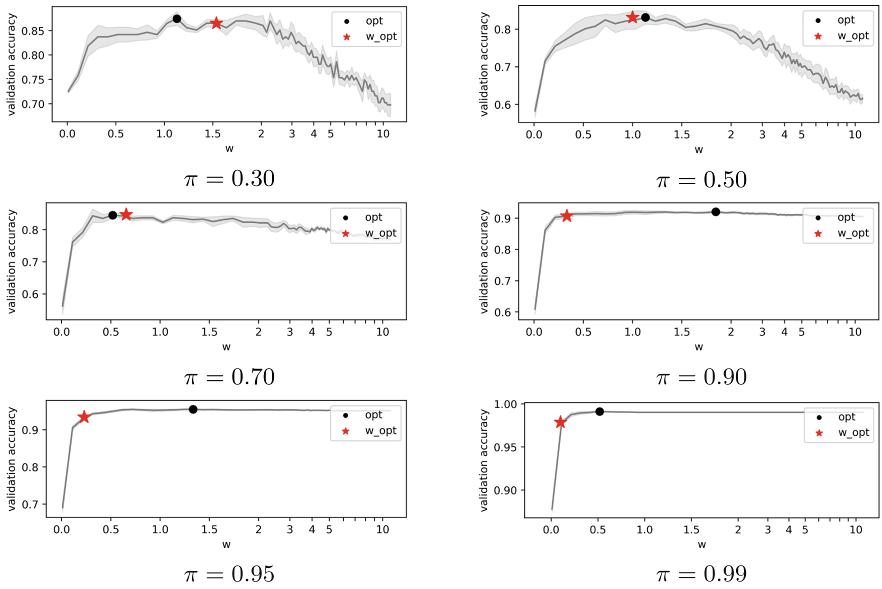


Fig. 10. w^* (black dot) vs. w^{opt} (red star) for the real dataset hypothyroid and different values of π

5 Conclusion

We considered the binary classification problem in the imbalanced setting. We showed that the oversampling approach under somewhat realistic assumptions can be interpreted as the weighted classification. We derived the generation bound for the weighted classification and discussed what connection the bound has with the selection of the resampling amount. We proposed the algorithm based on the derived optimal weighting. Results of the computational experiments demonstrated usefulness of the proposed approach.

Acknowledgments. E. Burnaev would like to thank Ivan Nazarov for fruitfull discussions on the problem statement and the derivation of the generalization bounds, Nikita Klyuchnikov for his help with computational experiments. E. Burnaev was supported by the Russian Fund of Basic Research under Grants 20-01-00203 and 21-51-12005 NNIO_a.

References

1. Artemov, A., Burnaev, E.: Ensembles of detectors for online detection of transient changes. In: Proceedings of the SPIE, vol. 9875, p. 9875–9875 - 5 (2015)
2. Artemov, A., Burnaev, E.: Detecting performance degradation of software-intensive systems in the presence of trends and long-range dependence. In: IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 29–36 (2016)
3. Artemov, A., Burnaev, E., Lokot, A.: Nonparametric decomposition of quasi-periodic time series for change-point detection. In: Proceedings of the SPIE, vol. 9875, p. 9875–9875 - 5 (2015)
4. Burnaev, E.: On construction of early warning systems for predictive maintenance in aerospace industry. J. Commun. Technol. Electron. **64**(12), 1473–1484 (2019)
5. Burnaev, E.: Rare failure prediction via event matching for aerospace applications. In: Proceedings of the 3rd International Conference on Circuits, System and Simulation (ICCSS-2019), pp. 214–220 (2019)
6. Burnaev, E., Erofeev, P., Papanov, A.: Influence of resampling on accuracy of imbalanced classification. In: Proceedings of the SPIE, vol. 9875, p. 9875–9875 - 5 (2015)
7. Burnaev, E., Erofeev, P., Smolyakov, D.: Model selection for anomaly detection. In: Proceedings of the SPIE, vol. 9875, p. 9875–9875 - 6 (2015)
8. Burnaev, E., Smolyakov, D.: One-class SVM with privileged information and its application to malware detection. In: IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 273–280 (2016)
9. Dupret, G., Koda, M.: Bootstrap re-sampling for unbalanced data in supervised learning. Eur. J. Oper. Res. **134**(1), 141–156 (2001)
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Heidelberg (2009)
11. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
12. Ignatiev, V., Trekin, A., Lobachev, V., Potapov, G., Burnaev, E.: Targeted change detection in remote sensing images. In: Proceedings of the SPIE (2019)
13. Kolos, M., Marin, A., Artemov, A., Burnaev, E.: Procedural synthesis of remote sensing images for robust change detection with neural networks. In: Lu, H., Tang, H., Wang, Z. (eds.) ISNN 2019. LNCS, vol. 11555, pp. 371–387. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22808-8_37
14. Maiboroda, R., Markovich, N.: Estimation of heavy-tailed probability density function with application to web data. Comput. Stat. **19**, 569–592 (2004)
15. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT Press, Cambridge (2018)
16. Natarajan, N., et al.: Cost-sensitive learning with noisy labels. JMLR **18**(1), 5666–5698 (2018)
17. Novikov, G., Trekin, A., Potapov, G., Ignatiev, V., Burnaev, E.: Satellite imagery analysis for operational damage assessment in emergency situations. In: Abramowicz, W., Paschke, A. (eds.) BIS 2018. LNBIP, vol. 320, pp. 347–358. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93931-5_25
18. Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J., Moore, J.H.: PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData Min. **10**(1), 36 (2017)
19. Scott, C.: Calibrated asymmetric surrogate losses. Electron. J. Stat. **6**, 958–992 (2012)

20. Smolyakov, D., Korotin, A., Erofeev, P., Papanov, A., Burnaev, E.: Meta-learning for resampling recommendation systems. In: Proceedings of the SPIE 11041, 11th International Conference on Machine Vision (ICMV 2018), p. 110411S (2019)
21. Smolyakov, D., Sviridenko, N., Burikov, E., Burnaev, E.: Anomaly pattern recognition with privileged information for sensor fault detection. In: Pancioni, L., Schwenker, F., Trentin, E. (eds.) ANNPR 2018. LNCS (LNAI), vol. 11081, pp. 320–332. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99978-4_25
22. Smolyakov, D., Sviridenko, N., Ishimtsev, V., Burikov, E., Burnaev, E.: Learning ensembles of anomaly detectors on synthetic data. In: Lu, H., Tang, H., Wang, Z. (eds.) ISNN 2019. LNCS, vol. 11555, pp. 292–306. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22808-8_30