Active Class Selection with Uncertain Deployment Class Proportions

Mirko Bunse(✉) and Katharina Morik

TU Dortmund University, Artificial Intelligence Group, 44221 Dortmund, Germany
{firstname.lastname}@tu-dortmund.de

Abstract. Active class selection strategies actively choose the class proportions of the data with which a classifier is trained. While this freedom of choice can improve the classification accuracy and reduce the data acquisition cost, it has also motivated theoretical studies that quantify the limited trustworthiness of the resulting classifier when the chosen class proportions differ from the class proportions that need to be handled during deployment. In this work, we build on these theoretic foundations to propose an active class selection strategy that allows machine learning practitioners to express their prior beliefs about the deployment class proportions. Unlike existing approaches, our strategy is justified by PAC learning bounds and naturally supports any degree of uncertainty with respect to these prior beliefs.

Keywords: Active class selection · Imbalanced binary classification · PAC learning theory.

1 Introduction

Active class selection (ACS) [11, 9] allows machine learning practitioners to actively choose the label proportions of their training data. This freedom of choice is due to a *class-conditional* data generator, e.g. an experiment or a simulation, which acquires feature vectors for arbitrarily chosen classes. Data generators of this kind appear in various use cases, such as astro-particle physics [4, 3], gas sensor arrays [11], and brain computer interaction [13].

Lomasky et al. [11] have put forward the idea that such a generator can be leveraged in a sequence of multiple acquisition steps, as sketched in Fig. 1. In each step, a classifier is trained and evaluated on all examples that have been acquired so far, starting from a small initial data set (i). Based on the classifier's performance, a data acquisition *strategy* is then allowed to choose the label proportions of the next acquisition step (ii). The class-conditional data generator realizes these proportions, i.e. it produces a batch of labeled data according to the choice of the strategy (iii). This batch adds to the training set from which the classifier will be trained in all subsequent iterations. The promise of such a sequential and informed data acquisition is that the classifier can benefit in terms of data acquisition cost and performance, as compared to being trained with some predetermined proportions of classes.

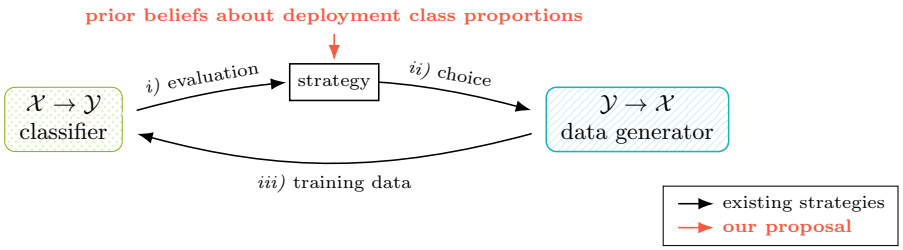

Fig. 1. Strategies for active class selection choose the label proportions of newly acquired training data. They are allowed to base their decisions on the performance of a classifier that is trained with all previously acquired data. We propose to incorporate prior beliefs, which can be uncertain, into the decision making.

Existing strategies [9, 11] for ACS do not account for the class proportions that a trained model needs to handle during deployment; they solely focus on the perceived *difficulty* of classes. One notable exception is a strategy that acquires training data precisely with those label proportions that are faced in the deployment stage; by design, this strategy requires the practitioner to know the deployment class proportions precisely in advance. However, what if we know the deployment class proportions not precisely, but with some degree of uncertainty? For instance, astro-particle physicists can estimate the ratio between their signal and their background class only roughly, as being approximately $1 : 10^3$ or even $1 : 10^4$ [2]. We are not aware of any ACS strategy that supports uncertain deployment class proportions out of the box.

Motivated by such uncertainties, we have recently proposed a theoretically justified *certificate* for ACS-trained models [4]. This certificate declares a range of deployment class proportions for which a given model is accurate (i.e. has an ACS-induced error smaller than some $\varepsilon > 0$) with a high probability (i.e. with probability at least $1 - \delta$). This declaration can help practitioners in assessing the practical value of an ACS-trained model. However, it has no immediate implication on how to acquire data—in terms of an ACS strategy—when the deployment class proportions are uncertain.

In the following, we therefore evolve the theoretical basis of our certificate towards a data acquisition strategy for ACS. This strategy uniquely combines the following qualities:

- our ACS strategy naturally supports uncertainty about the deployment class proportions, e.g. as expressed by a Beta prior for binary classification.
- our strategy is theoretically justified by PAC learning bounds.

Our experiments suggest that our strategy, even under high amounts of uncertainty, exhibits a performance that is comparable to the performance of an optimal strategy with privileged access to the class proportions of the test set. Other strategies, which are oblivious to the deployment class proportions, fall behind by a significant margin.

We summarize the theoretic foundations of ACS in Sec. 2 before we detail our strategy in Sec. 3. The experiments in Sec. 4 lead to our conclusion in Sec. 5.

2 Theoretical Background

The term "domain", as proposed by domain adaptation [14, 12], describes a probability density function over the data space $\mathcal{X} \times \mathcal{Y}$. In ACS, we assume that the *source domain* $\mathcal{S}$—where a machine learning model is trained—differs from the *target domain* $\mathcal{T}$—where the model is deployed—only in terms of the class proportions $p_{\mathcal{S}} \neq p_{\mathcal{T}}$. Such deviations occur due to the freedom of ACS strategies to choose any $p_{\mathcal{S}}$ for the acquisition of training data. We are interested in the impact of such deviations on the deployment performance, i.e. on the classification performance with respect to $\mathcal{T}$.

Recently, a PAC learning perspective on this setting has provided us with Theorem 1 [4]. This result quantifies the difference in loss values $L(h)$ between an ACS-generated training set D and the target domain $\mathcal{T}$. Only if this difference is small, we can expect to learn a classifier h from D that is accurate also with respect to $\mathcal{T}$, similar to standard PAC learning theory. The key insight of this theorem is that the relevant loss difference between D and $\mathcal{T}$ is continuously approaching the inter-domain gap $\Delta p \cdot \Delta \ell$ while the training set size m increases. In ACS, this increase happens naturally while more and more data is actively being acquired, so that the error of any ACS-trained classifier is increasingly dominated by this gap. Here, $\Delta p = |p_{\mathcal{T}} - p_{\mathcal{S}}|$ denotes the difference between class proportions and $\Delta \ell = |\ell_{Y=2}(h) - \ell_{Y=1}(h)|$ denotes the difference between class-wise losses. The latter of these terms is constant across domains $\mathcal{S}$ and $\mathcal{T}$. In turn, $\Delta p \cdot \Delta \ell$ is constant with respect to the random draw of the training set D and is therefore independent of ε, δ, and m; it reflects the interplay between the classifier h, the data distribution, and the loss function.

Theorem 1 (Identical mechanism bound; binary classification [4]). *For any $\varepsilon > 0$, any $h \in \mathcal{H}$, with probability at least $1 - \delta$, where $\delta = 4e^{-2m\varepsilon^2}$:*

$$\Delta p \cdot \Delta \ell \; - \; \varepsilon \quad \leq \quad |L_{\mathcal{T}}(h) - L_D(h)| \quad \leq \quad \Delta p \cdot \Delta \ell \; + \; \varepsilon$$

The true difference $\Delta \ell$ from Theorem 1 is unknown, but we can estimate an upper bound $\Delta \ell^*$ of this quantity from ACS-generated data. The details on this estimation are already presented in the scope of ACS model certification [4] and do not need to be repeated here. All we need to know to establish our ACS strategy is that $\Delta \ell^*$ is the smallest upper bound of $\Delta \ell$ that holds with probability at least $1 - \delta$. The probabilistic nature of this upper bound stems from the fact that $\Delta \ell^*$ is estimated from finite amounts of data.

3 A Strategy for Uncertain Class Proportions

The goal of our strategy is to decrease the inter-domain gap $\Delta p \cdot \Delta \ell$ from Theorem 1 as much as possible, as according to a prior distribution $\hat{\mathbb{P}}$ of the deployment class proportions $p_{\mathcal{T}}$. This goal will allow any binary classification

algorithm to learn accurate predictions for the target domain, as according to the prior beliefs of a domain expert.

Formally, we assume a prior $\hat{\mathbb{P}} : [0,1] \to [0,1]$ of the positive class prevalence $p_{\mathcal{T}} \in [0,1]$ to be given. We incorporate $\hat{\mathbb{P}}$ by marginalizing the inter-domain gap over this prior, as according to Eq. 1. Since we do not know the true $\Delta\ell$, we are using the estimated upper bound $\Delta\ell^*$ instead. Consequently, the marginalization according to $\Delta\ell^*$ is an upper bound, with probability $1-\delta$, of the marginalization according to the true $\Delta\ell$.

$$\varepsilon^* \;=\; \int_0^1 \hat{\mathbb{P}}(p_{\mathcal{T}} = p) \cdot \underbrace{|p_{\mathcal{S}} - p|}_{=\,\Delta p} \cdot \Delta\ell^* \;\; \mathrm{d}\, p \tag{1}$$

In each ACS iteration, we are free to alter the class proportions $p_{\mathcal{S}}$ of the ACS-generated training set to some degree, depending on how much data we acquire in each batch and on how much data we already have acquired. In fact, we can understand $p_{\mathcal{S}} = \frac{m_2}{(m_1+m_2)}$ as a function of the class-wise numbers of samples m_1 and m_2. The upper bound $\Delta\ell^*$ also lends itself for being interpreted as a function of sample sizes: the more data is acquired in both classes, the tighter will our estimation of this quantity be. Ultimately, we consider ε^* to be a function of m_1 and m_2, so that we can minimize ε^* via an optimal choice of m_1 and m_2 in each data acquisition batch.

3.1 Minimizing the Marginalized Error

Our strategy decreases ε^* in the direction of its steepest descent, i.e. it takes a simple gradient step with respect to the acquisition vector $\boldsymbol{m} = (m_1, m_2)$. The gradient which defines the steepest descent is computed via the product rule:

$$\begin{aligned} \nabla_{\boldsymbol{m}}\, \varepsilon^* \;&=\; \nabla_{\boldsymbol{m}} f \cdot \Delta\ell^* \;+\; f \cdot \nabla_{\boldsymbol{m}} \Delta\ell^* \\ \text{where } \; f(\boldsymbol{m}) \;&=\; \int_0^1 \hat{\mathbb{P}}(p_{\mathcal{T}} = p) \cdot |p_{\mathcal{S}}(\boldsymbol{m}) - p| \;\; \mathrm{d}\, p \end{aligned} \tag{2}$$

We will come back to the function f shortly. For now, we plug $\Delta\ell^*$ and $\nabla_{\boldsymbol{m}}\Delta\ell^*$ into the equation above. These functions are defined by

$$\begin{aligned} \Delta\ell^*(\boldsymbol{m}) \;&=\; \hat{\ell}_{Y=2}(h) + \sqrt{\frac{\ln \delta_2}{-2m_2}} - \hat{\ell}_{Y=1}(h) + \sqrt{\frac{\ln \delta_1}{-2m_1}}, \\ [\nabla_{\boldsymbol{m}} \Delta\ell^*]_y \;&=\; \left(-\frac{\ln \delta_y}{m_y}\right)^{\frac{3}{2}} \cdot (2\sqrt{2} \ln \delta_y)^{-1}, \end{aligned} \tag{3}$$

where the δ_y are probabilities of violations of $\Delta\ell^*$ that occur from either one of the class-wise losses $\ell_{Y=y}(h)$ in $\Delta\ell$. In fact, finding a suitable assignment of δ_y values within a given probability budget $\delta = \delta_1 + \delta_2 - \delta_1\delta_2$ is the central difficulty in model certification; there, the sample size $\boldsymbol{m}$ is fixed, so that $\Delta\ell^*$ can be optimized over this assignment [4]. Here, we keep the δ_y fixed instead, to

values that are obtained with a certificate from previous ACS acquisitions. This change allows us to optimize $\Delta\ell^*$ over $\boldsymbol{m}$ to acquire new data and it guarantees that $\Delta\ell^*$ remains an upper bound of the true $\Delta\ell$ also in the next batch, at least with probability $1-\delta$. The class-wise estimates $\hat{\ell}_{Y=y}(h)$ in Eq. 3 are the average values of losses in the training data; they are also part of our certificate.

3.2 A Beta Prior for Binary Class Proportions

Now we turn to the function value and the gradient of the function f in Eq. 2. Plugging a parametric prior $\hat{\mathbb{P}}$ into this function can allow us to compute these terms efficiently, in closed forms. To this end, a Beta(α, β) prior is suitable for binary classification because the Beta distribution is a conjugate prior of the Bernoulli distribution, which in turn is a suitable model for the prevalence of binary class labels. As a matter of convenience, the parameters $\alpha > 0$ and $\beta > 0$ can be chosen such that the resulting distribution has some predetermined mean and standard deviation; we believe that domain experts can often express their prior beliefs in terms of these properties.

Plugging a Beta prior into the f function from Eq. 2 yields the following components, where I is the regularized incomplete Beta function:

$$\begin{aligned} f_{\alpha,\beta}(\boldsymbol{m}) &= \frac{2p_{\mathcal{S}}(\boldsymbol{m})^{\alpha}(1-p_{\mathcal{S}}(\boldsymbol{m}))^{\beta}}{(\alpha+\beta)B(\alpha,\beta)} + \Big(p_{\mathcal{S}}(\boldsymbol{m}) - \frac{\alpha}{\alpha+\beta}\Big)\Big(2I_{p_{\mathcal{S}}(\boldsymbol{m})}(\alpha,\beta) - 1\Big) \\ \nabla_{\boldsymbol{m}} f_{\alpha,\beta} &= \frac{2I_{p_{\mathcal{S}}(\boldsymbol{m})}(\alpha,\beta) - 1}{(m_1+m_2)^2} \cdot \begin{pmatrix} m_2 \\ -m_1 \end{pmatrix} \end{aligned} \tag{4}$$

Plugging Eq. 3 and 4 into Eq. 2 provides us with a gradient that we can compute analytically from a certificate with a δ_y assignment, from sample sizes m_1 and m_2 and from the prior parameters α and β. The negative gradient $-\nabla_{\boldsymbol{m}}\,\varepsilon^*$ of the marginalized error ε^* defines the class-wise numbers of samples that our strategy acquires in the next data acquisition batch.

With small data volumes or with highly imbalanced classes, our strategy is dominated by the $\Delta\ell^*$ component; small classes need additional data until this upper bound holds with some desired probability $1-\delta$. Constrastingly, when the total data volume is large, our strategy is dominated by the f component; to this end, a Beta prior favors class proportions that are close to its mean $\frac{\alpha}{\alpha+\beta}$. The turning point between these two behaviors is well-founded in the PAC learning theory that underlies the estimation of $\Delta\ell^*$.

4 Experiments

The first introduction of the ACS problem is already accompanied by the proposal of five heuristic ACS strategies [11]. In the following, we compare our own strategy from Sec. 3 to these five heuristics:

proportional: always sample according to $p_{\mathcal{T}}$, provided that these true proportions are already known at training time.

uniform: always sample all classes in the same amount.
inverse: sample according to the inverse accuracy of a classifier that is trained on earlier batches; the underlying assumption is that weak class-wise performances can be counteracted with over-sampling.
improvement: sample according to the class-wise improvement in accuracy that has occurred between the current iteration and the iteration before; this strategy assumes that stable performances, i.e. performances that did not change recently, will remain stable during future acquisitions.
redistriction: sample according to the class-wise number of training examples for which the prediction has changed between the current iteration and the iteration before; the assumption here is that stability can be promoted by over-sampling classes with volatile decision boundaries.

Our theoretical analysis of the ACS problem [5, 4] reveals that the *proportional* strategy is actually more than a heuristic; this strategy is indeed *optimal* in the limit of data acquisition. However, it requires precise knowledge of $p_{\mathcal{T}}$, which practitioners might not be able to provide. Contrastingly, all other strategies are entirely oblivious to the deployment proportions; they solely focus on different notions of class-wise difficulties.

This shortcoming is also shared by an ACS strategy that aggregates utility scores of pseudo-instances [9]. For now, we have excluded this approach from our comparison, due to this property. For future work, however, we expect that the method can overcome this limitation with a recent update of its utility function [8]. This update supports a prior of $p_{\mathcal{T}}$, which is in line with our idea of incorporating prior beliefs in ACS. Embedding the update in the original pseudo-instance strategy, however, might not be trivial.

4.1 Methodology

We have parameterized the Beta prior of our strategy with a predetermined mean and standard deviation, both set to the true value of $p_{\mathcal{T}}$. Accordingly, the mean of the prior is well aligned with the true class proportions of the deployment data; the uncertainty, however, is as large as possible.

In accordance to a reliable evaluation methodology [7], we present pairwise differences between ACS strategies in terms of their statistical significance. A comprehensive way of plotting such differences is through critical difference diagrams [6, 1], which compare multiple strategies over multiple data sets in a statistically sound way. We employ accuracy as the underlying performance metric and we conduct multiple trials to obtain an average performance value for each combination of strategy and data set. These average performances are then summarized through critical difference diagrams.

We define the trials via five repetitions of a three-fold cross validation. From the *imbalanced-learn*[1] package [10], we retrieve 13 data sets that have at least 150 minority class samples (to facilitate sampling) and at most 100 features (to

[1] `https://imbalanced-learn.org/stable/datasets/`

facilitate learning). We ensure comparability between all strategies by employing the same classifier in all experiments, a logistic regression with default meta-parameters. The data acquisition happens in up to 8 batches, each of which acquires 50 new training examples. However, not all strategies reach the last batch on all data sets; we stop each trial as soon as the strategy exhausts one of the classes. We opted for this early stopping criterion to focus on "realistic" acquisitions that happen due to free choices and not due to the fact that our experiment only simulates class-dependent data acquisition with finite pools of data. For the same reason, and due to weak performances on imbalanced data, we did not evaluate the *uniform* strategy here. Due to the early stopping, it becomes increasingly harder to detect significant differences; while the batches three and four can be evaluated on all data sets, only 9 data sets remain for batch eight. The implementation of our configurable experiments is available online[2].

4.2 Results and Discussion

Fig. 2 presents the critical difference diagrams, as according to our evaluation methodology. We see that our method, with access to an uncertain prior of $p_{\mathcal{T}}$, performs as well as the privileged strategy that knows $p_{\mathcal{T}}$ precisely. Moreover, our method outperforms all existing strategies which are oblivious to $p_{\mathcal{T}}$.

Fig. 3 traces this success back to the acquisition behavior that each strategy exhibits. Our own strategy quickly approaches the true proportions $p_{\mathcal{T}}$ of classes, due to the perfect alignment between the mean of the prior and $p_{\mathcal{T}}$. For the particular case of a Beta prior, this behavior is a reason for concern: if the mean of this prior was not well aligned with $p_{\mathcal{T}}$, we might have acquired data in mistaken class proportions; only if the mean of the Beta prior is sufficiently accurate, we can expect the competitive behavior that Fig. 2 suggests. Future research down this lane, e.g. with other types of prior distributions, is needed.

Fig. 3 further reveals two explanations for the poor performances of the existing strategies: first, all of these strategies exhibit a central tendency of staying close to the class proportions of the initial training set; second, each of these strategies prefers class proportions of an increasingly large variability. Both of these behaviors are due to the sole focus of these strategies on the perceived difficulties of classes, which can differ considerably between the data sets.

5 Conclusion and Outlook

In contrast to existing ACS strategies, which either assume precise knowledge about the deployment class proportions or no knowledge at all, we have advocated the incorporation of a prior distribution that expresses beliefs about the class proportions with any degree of (un)certainty. Our ACS strategy is well-founded on PAC learning bounds which we have recently proposed for ACS [4]. Experiments suggest that our strategy performs as well as the fully certain case, which, however, is harder to specify than an uncertain prior.

[2] `https://github.com/mirkobunse/AcsCertificates.jl`

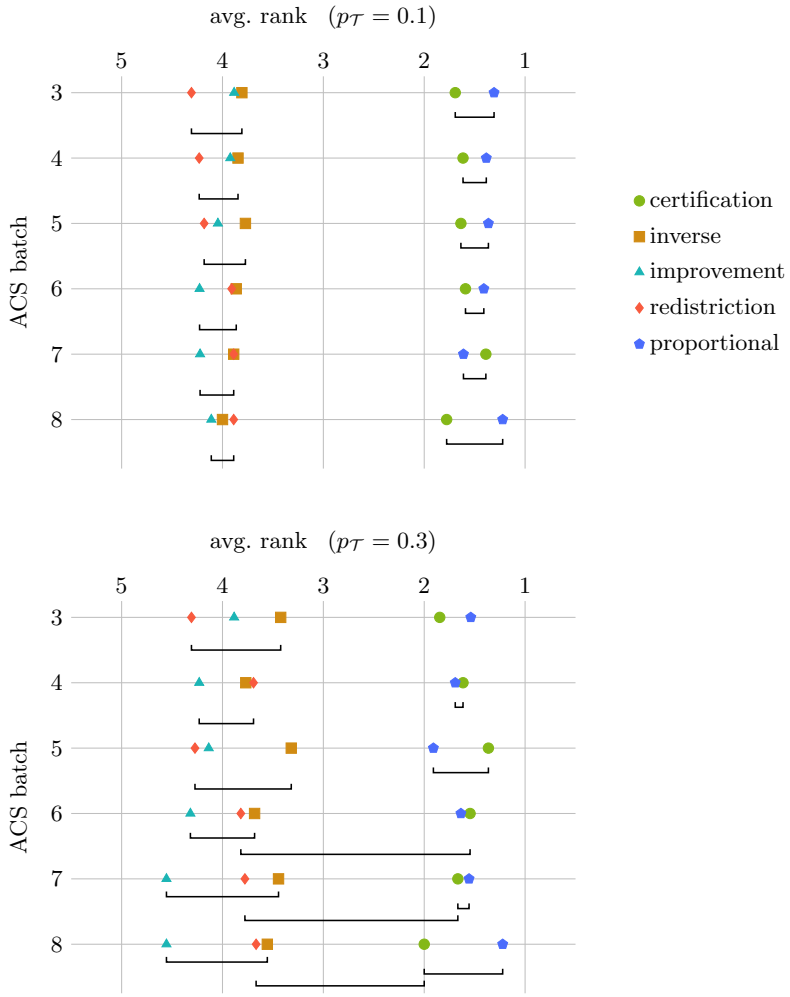

Fig. 2. Critical difference diagrams evaluate our ACS strategy (●) against existing ACS strategies [11], one of which has privileged access to the true class proportions $p_{\mathcal{T}}$ (⬟). The two plots present different values of $p_{\mathcal{T}}$. Each position on the vertical axes corresponds to one critical difference diagram for one batch in the ACS data acquisition loop. Horizontal positions correspond to the average ranks of strategies across multiple data sets, as according to the average accuracy in multiple trials; lower ranks are better. Horizontal connections between two or more strategies indicate that a Wilcoxon signed-rank test is not able to detect significant differences between these methods from the performances they exhibit.

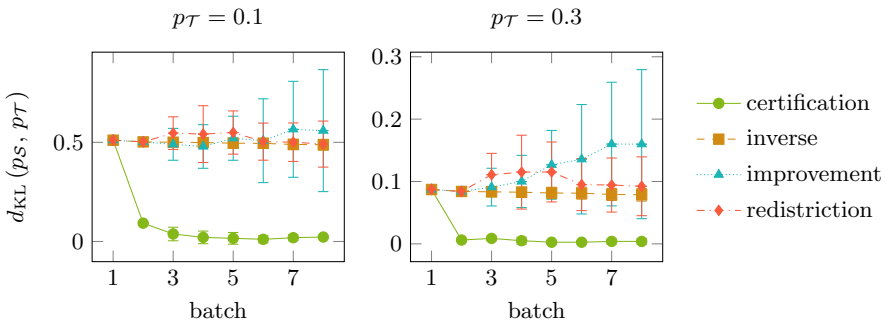

Fig. 3. Our ACS strategy (●) quickly approaches the true proportions $p_\mathcal{T}$ of classes in terms of the Kullback-Leibler divergence d_{KL}. Due to the uncertainty of the prior, however, this divergence always remains above zero. The standard deviation of d_{KL}, as displayed by the error bars, increases considerably with the other strategies.

Future work on ACS should focus on strategies that support multi-class classification and regression. We identify the PAL-ACS framework [9] with a recent update of its utility function [8] as a promising candidate in this direction.

Acknowledgments

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Data Analysis", project C3, and by the Federal Ministry of Education and Research of Germany as part of the competence center for machine learning ML2R (01IS18038 A/B).

References

1. Benavoli, A., Corani, G., Mangili, F.: Should we really use post-hoc tests based on mean-ranks? J. Mach. Learn. Res. **17**(1), 152–161 (2016)
2. Bockermann, C., Brügge, K., Buss, J., Egorov, A., Morik, K., Rhode, W., Ruhe, T.: Online analysis of high-volume data streams in astroparticle physics. In: Europ. Conf. on Mach. Learn. and Knowledge Discovery in Databases. Springer (2015)
3. Bunse, M., Bockermann, C., Buss, J., Morik, K., Rhode, W., Ruhe, T.: Smart control of Monte Carlo simulations for astroparticle physics. In: Astronomical Data Analysis Software and Systems. pp. 417–420 (2017)
4. Bunse, M., Morik, K.: Certification of model robustness in active class selection. In: Europ. Conf. on Mach. Learn. and Knowledge Discovery in Databases. Springer (2021)
5. Bunse, M., Weichert, D., Kister, A., Morik, K.: Optimal probabilistic classification in active class selection. In: Int. Conf. on Data Mining. IEEE (2020)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**(1), 1–30 (2006)

7. Kottke, D., Calma, A., Huseljic, D., Krempl, G., Sick, B.: Challenges of reliable, realistic and comparable active learning evaluation. In: Workshop and Tutorial on Interactive Adaptive Learn. pp. 2–14 (2017)
8. Kottke, D., Herde, M., Sandrock, C., Huseljic, D., Krempl, G., Sick, B.: Toward optimal probabilistic active learning using a Bayesian approach. Mach. Learn. **110**(6), 1199–1231 (2021)
9. Kottke, D., Krempl, G., Stecklina, M., von Rekowski, C.S., Sabsch, T., Minh, T.P., Deliano, M., et al.: Probabilistic active learning for active class selection. In: NeurIPS Workshop on the Future of Interactive Learn. Mach. (2016)
10. Lemaitre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18** (2017)
11. Lomasky, R., Brodley, C.E., Aernecke, M., Walt, D., Friedl, M.A.: Active class selection. In: Europ. Conf. on Mach. Learn. and Knowledge Discovery in Databases. Springer (2007)
12. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10) (2010)
13. Parsons, T.D., Reinebold, J.L.: Adaptive virtual environments for neuropsychological assessment in serious games. IEEE Trans. Consumer Electron. **58**(2) (2012)
14. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312** (2018)