Certification of Model Robustness in Active Class Selection

Mirko Bunse(✉) and Katharina Morik

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

Abstract. Active class selection provides machine learning practitioners with the freedom to actively choose the class proportions of their training data. While this freedom can improve the model performance and decrease the data acquisition cost, it also puts the practical value of the trained model into question: is this model really appropriate for the class proportions that are handled during deployment? What if the deployment class proportions are uncertain or change over time? We address these questions by *certifying* supervised models that are trained through active class selection. Specifically, our certificate declares a set of class proportions for which the certified model induces a training-to-deployment gap that is small with a high probability. This declaration is theoretically justified by PAC bounds. We apply our proposed certification method in astro-particle physics, where a simulation generates telescope recordings from actively chosen particle classes.

Keywords: Active class selection · Label shift · Model certification · Learning theory · Classification · Validation · Imbalanced learning.

1 Introduction

The increasing adoption of machine learning in practice motivates model performance reports [17, 2, 22] that are easily accessible by a diverse group of stakeholders. One particular concern of this trend is the robustness of trained models with regard to changing deployment conditions, like distribution shifts [26], input perturbations [12], or adversarial attacks [25, 30]. Ideally, robustness criteria are *certified* in the sense of being formally proven or thoroughly tested [12].

The framework of active class selection (ACS; see Fig. 1) [16, 13] presumes a *class-conditional* training data generator, e.g. an experiment or a simulation that produces feature vectors for arbitrarily chosen classes. As a consequence, the developer of a machine learning model must actively decide for the class proportions of the training data set. This freedom can benefit the learning process in terms of data acquisition cost and model performance. However, the active decision for class proportions is difficult if the class proportions that occur during deployment are not precisely known or are subject to changes. In astro-particle physics, for instance, the ratio between the signal and the background class is

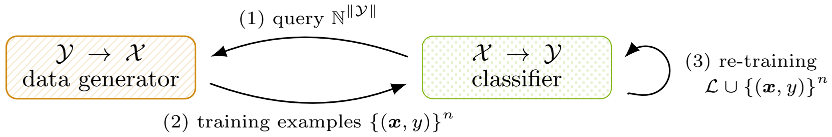
$\mathcal{Y} \to \mathcal{X}$
data generator
(1) query $\mathbb{N}^{\|\mathcal{Y}\|}$
(2) training examples $\{(\boldsymbol{x}, y)\}^n$
$\mathcal{X} \to \mathcal{Y}$
classifier
(3) re-training
$\mathcal{L} \cup \{(\boldsymbol{x}, y)\}^n$

Fig. 1. Active class selection optimizes class-conditioned data acquisition [5].

only roughly estimated as $1 : 10^3$ or even $1 : 10^4$ [4]. Other use cases of ACS are brain computer interaction [21, 29, 10] and gas sensor arrays [16].

Recently, we have studied ACS from an information-theoretic viewpoint [5]. This viewpoint suggests that, in ACS, the training class proportions should be chosen identically to the deployment class proportions, at least when the sample size is sufficiently large. However, we also know from imbalanced classification [7] that highly imbalanced class proportions, as in the astro-particle use case, are often far from optimal. Moreover, the deployment class proportions may be uncertain at training time or may be subject to change during deployment. Therefore, we make the following contributions in the present paper:

- We study ACS through the lens of learning theory. This view-point provides us with PAC bounds that are more nuanced than previous [5] results. Namely, our bounds are applicable also to extremely imbalanced domains and they account for finite data volumes.
- Through these bounds, we quantify the *domain gap* which results from the label shift between the ACS-generated training data and the data that is predicted during deployment.
- We refine these results in a *certificate* for binary classifiers under label shift. This certificate verifies the range of class proportions for which an ACS-trained classifier is accurate (i.e. has a domain-induced error smaller than some $\varepsilon > 0$) with a high probability (i.e. with probability at least $1 - \delta$). This certificate is specific, understandable, theoretically justified, and applicable to any learning method. Users specify ε and δ according to their demands.

In the following, we briefly review the ACS problem statement (Sec. 1.1) and previous work on the topic (Sec. 1.2). Sec. 2 presents our theoretical contributions, followed by their experimental verification in Sec. 3. We present additional related work in Sec. 4 and conclude with Sec. 5.

1.1 Active Class Selection Constitutes a Domain Gap

Following the terminology of domain adaptation [27, 20], we consider a *domain* as a probability density function over the labeled data space $\mathcal{X} \times \mathcal{Y}$. This density function stems from some particular data-generating process; a different process will induce a different domain. To this end, let $\mathcal{S}$ be the *source* domain where a machine learning model is trained and let $\mathcal{T}$ be the *target* domain where the trained model is required to be accurate. We are interested in the impact of deviations $\mathcal{S} \neq \mathcal{T}$ on the deployment performance.

Here, we assume that $\mathcal{S}$ and $\mathcal{T}$ *only* differ in their class proportions, to study the effect of ACS in isolation from other potential deviations between $\mathcal{S}$ and $\mathcal{T}$. Put differently, we let all data be generated by the same causal mechanism $Y \rightarrow X$, according to the factorization $\mathbb{P}(x, y) = \mathbb{P}(x \mid y) \cdot \mathbb{P}(y)$. More formally:

Definition 1 (Identical mechanism assumption [5] a.k.a. label shift or target shift [31]). *Assume that all data in the domains $\mathcal{S}$ and $\mathcal{T}$ is generated independently by the same class-conditioned mechanism, i.e.*

$$\mathbb{P}_{\mathcal{S}}(X = x \mid Y = y) \ = \ \mathbb{P}_{\mathcal{T}}(X = x \mid Y = y) \qquad \forall x \in \mathcal{X},\ \forall y \in \mathcal{Y}$$

1.2 A Qualitative Intuition from Information Theory

We have recently studied the domain gap $\mathcal{S} \neq \mathcal{T}$ in the limit of data acquisition, i.e. when the sample size $m \rightarrow \infty$ [5]. In this limit, the deployment proportions should also be reflected in the ACS-generated training data. However, we have also observed that certain deviations from this fixed-point are feasible without impairing the classifier; the range of these feasible deviations depends on the correlation between features and labels.

Proposition 2 (Information-theoretical bound [5]). *The domain $\mathcal{S}$ misrepresents the prediction function $\mathbb{P}_{\mathcal{T}}(Y \mid X)$ by the KL divergence $d_{Y|X}$, which is bounded above by the KL divergence d_Y between $\mathbb{P}_{\mathcal{S}}(Y)$ and $\mathbb{P}_{\mathcal{T}}(Y)$:*

$$d_{Y|X} \ = \ d_Y \, - \, d_X \ \leq \ d_Y$$

Remarkably, the more data is being acquired by ACS, the less beneficial will any class proportion other than $\mathbb{P}_{\mathcal{T}}(Y)$ be. Beyond these qualitative insights, however, the information-theoretic perspective has not allowed us to provide quantitative bounds which precisely assess the impact of the sample size. In the following, we therefore employ a different perspective on ACS from PAC learning theory. This perspective accounts for the sample size m, an error margin $\varepsilon > 0$ and a desired probability $1 - \delta < 1$.

2 A Quantitative Perspective from Learning Theory

We start by recalling a standard i.i.d. bound from learning theory, which we then extend to the domain gap induced by ACS. The standard i.i.d. bound quantifies the probability that the estimation error of the loss $L_{\mathcal{S}}(h)$, induced by the finite amount m of data in a data set $\mathrm{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \,:\, 1 \leq i \leq m\} \sim \mathcal{S}^m$ relative to the training domain $\mathcal{S}$, is bounded above by some $\varepsilon > 0$:

Proposition 3 (I.i.d. bound [24]). *For any $\varepsilon > 0$ and any fixed $h \in \mathcal{H}$, it holds with probability at least $1 - \delta$, where $\delta = 2e^{-2m\varepsilon^2}$, that:*

$$|L_D(h) - L_{\mathcal{S}}(h)| \ \leq \ \varepsilon$$

Proof. We repeat the proof by Shalev-Shwartz & Ben-David [24, Sec. 4.2] here to extract Corollary 4 for later reference. Let $L_{\mathrm{D}}(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(y_i, h(x_i))$ be the empirical loss over a data set D and let $L_{\mathcal{S}}(h) = \mathbb{E}_{(x,y)\sim\mathbb{P}_{\mathcal{S}}}[\ell(y, h(x))]$ be the expected value of $L_{\mathrm{D}}(h)$ and every $\ell(y_i, h(x_i))$. Then, by letting $\theta_i = \ell(y_i, h(x_i))$ and $\mu = L_{\mathcal{S}}(h)$, we apply Hoeffding's inequality for $0 \leq \theta_i \leq 1$:

$$\mathbb{P}_{\mathrm{D}\sim\mathcal{S}^m}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right) \;\leq\; 2e^{-2m\varepsilon^2} \;=\; \delta \tag{1}$$

We see that the converse, i.e. $\left|\frac{1}{m}\sum_{i_1}^{m}\theta_i - \mu\right| \leq \varepsilon$, holds with probability at least $1-\delta$; taking Eq. 1 for granted would therefore already yield our claim (□). Instead, however, we take another step back and prove Eq. 1 from Hoeffding's Lemma, which states that for every $\lambda > 0$ and any random variable $X \in [a, b]$ with $\mathbb{E}[X] = 0$ it holds that:

$$\mathbb{E}[e^{\lambda X}] \;\leq\; e^{\frac{\lambda^2(b-a)^2}{8}} \tag{2}$$

Letting $X_i = \theta_i - \mu$ and $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$, we use *i)* monotonicity, *ii)* Markov's inequality, *iii)* independence, and *iv)* Eq. 2 with $a = 0$, $b = 1$, and $\lambda = 4m\varepsilon$:

$$\mathbb{P}[\bar{X} \geq \varepsilon] \overset{i)}{=} \mathbb{P}[e^{\lambda\bar{X}} \geq e^{\lambda\varepsilon}] \overset{ii)}{\leq} e^{-\lambda\varepsilon}\,\mathbb{E}[e^{\lambda\bar{X}}] \overset{iii)}{=} e^{-\lambda\varepsilon}\prod_{i=1}^{m}\mathbb{E}[e^{\lambda X_i/m}] \overset{iv)}{=} e^{-2m\varepsilon^2} \tag{3}$$

We apply Eq. 3 to $\bar{X}$ and $-\bar{X}$ to yield Eq. 1 via the union bound. □

Corollary 4 (Asymmetric i.i.d. bound). *For any $\varepsilon^{(l)}, \varepsilon^{(u)} > 0$ and any fixed $h \in \mathcal{H}$, each of the following bounds holds with probability at least $1 - \delta^{(i)}$ respectively, where $\delta^{(i)} = e^{-2m(\varepsilon^{(i)})^2}$ and $i \in \{l, u\}$:*

i) $L_D(h) - L_{\mathcal{S}}(h) \;\leq\; \varepsilon^{(l)}$

ii) $L_{\mathcal{S}}(h) - L_D(h) \;\leq\; \varepsilon^{(u)}$

Proof. The claim follows from applying Eq. 3 to $\bar{X}$ and $-\bar{X}$, just like in the proof of Proposition 3. This time, however, we use two different $\varepsilon^{(l)}$, $\varepsilon^{(u)}$ for the two sides of the bound and we do not combine them via the union bound. □

To study the ACS problem, we now replace the i.i.d. assumption above with the identical mechanism assumption from Def. 1. The result is Theorem 5, in which the factor 4 in δ, as compared to the factor 2 in the δ of Lemma 3, stems from the fact that either the upper bound or the lower bound might be violated, each time with at most the same probability. Fig. 2 illustrates the idea.

Theorem 5 (Identical mechanism bound). *For any $\varepsilon > 0$ and any fixed $h \in \mathcal{H}$, it holds with probability at least $1 - \delta$, where $\delta = 4e^{-2m\varepsilon^2}$, that:*

$$|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \;-\; \varepsilon \;\leq\; |L_{\mathcal{T}}(h) - L_D(h)| \;\leq\; |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \;+\; \varepsilon$$

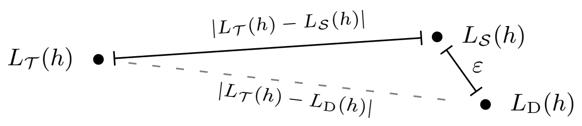
$|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)|$
$L_{\mathcal{T}}(h)$
$L_{\mathcal{S}}(h)$
ε
$|L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)|$
$L_{\mathrm{D}}(h)$

Fig. 2. Illustration of Theorems 5 and 6. Keeping $\delta > 0$ fixed, we can choose $\varepsilon \to 0$ as $m \to \infty$. What remains is the inter-domain gap $|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)|$.

Proof. We employ Proposition 3 through the triangle inequality (see Fig. 2):

$$\begin{aligned} |L_{\mathrm{D}}(h) - L_{\mathcal{T}}(h)| &\leq |L_{\mathrm{D}}(h) - L_{\mathcal{S}}(h)| + |L_{\mathcal{S}}(h) - L_{\mathcal{T}}(h)| \\ &\leq \varepsilon + |L_{\mathcal{S}}(h) - L_{\mathcal{T}}(h)|, \end{aligned}$$

where the second inequality holds with probability at least $1 - 2e^{-2m\varepsilon^2}$.

Likewise, and with the same probability, we use the triangle inequality for the other side of the claim:

$$\begin{aligned} |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| &\leq |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| + |L_{\mathrm{D}}(h) - L_{\mathcal{S}}(h)| \\ \Leftrightarrow |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| &\geq |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| - |L_{\mathrm{D}}(h) - L_{\mathcal{S}}(h)| \\ &\geq |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| - \varepsilon \end{aligned}$$ □

The above bound addresses a single fixed hypothesis $h \in \mathcal{H}$, which suits our goal of certifying any given prediction model. For completeness, however, let us also mention that Theorem 5 can be extended to entire hypothesis classes $\mathcal{H}$. As an example, we obtain the following result for finite classes, i.e. for $|\mathcal{H}| < \infty$:

Theorem 6 (Identical mechanism bound; finite hypothesis class). *With probability at least $1-\delta$, where $\delta = 4|\mathcal{H}|e^{-2m\varepsilon^2}$, the upper and lower bounds from Theorem 5 hold for all $h \in \mathcal{H}$.*

Proof. The claim follows from the union bound, being detailed in Appendix 1.

The lower and upper bounds in Theorems 5 and 6 quantify how the total error $|L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)|$ approaches the inter-domain gap $|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)|$ in dependence of the interplay between ε, δ, m, L, and $\mathcal{H}$. It is therefore a quantitative and thus more nuanced equivalent of Proposition 2. The inter-domain gap is constant w.r.t. the random draw of the training sample $\mathrm{D} \sim \mathcal{S}^m$ and is therefore independent of ε, of δ, and of m. Consequently, it remains even with an infinite amount of training data. Depending on the choice of $\mathcal{H}$ and L, and in dependence of the data distribution, it may be large or negligible. In the following, we will therefore study this error in more detail.

2.1 Quantification of the Domain Gap

We begin by factorizing the total error $L(h)$ into label-dependent losses $\ell_X(h, y)$ that are marginalized over the entire feature space $\mathcal{X}$. These losses only depend on the hypothesis h and on the label y and are, under the identical mechanism assumption from Def. 1, identical among $\mathcal{S}$ and $\mathcal{T}$.

$$\begin{aligned} L(h) \;&=\; \int_{\mathcal{Y}} \int_{\mathcal{X}} \mathbb{P}(x, y) \ell(y, h(x)) \; dx \; dy \\ &=\; \int_{\mathcal{Y}} \mathbb{P}(y) \underbrace{\int_{\mathcal{X}} \mathbb{P}(x \mid y) \ell(y, h(x)) \; dx}_{= \, \ell_X(h,y)} \; dy \end{aligned}$$

Plugging $\ell_X(h, y)$ into the domain gap $|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)|$ from the Theorems 5 and 6 allows us to marginalize the label-dependent losses over the label space:

$$|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \;=\; \left| \int_{\mathcal{Y}} \mathbb{P}_{\mathcal{T}}(y) \, \ell_X(h, y) \; dy \;-\; \int_{\mathcal{Y}} \mathbb{P}_{\mathcal{S}}(y) \, \ell_X(h, y) \; dy \right|$$

For classification tasks, i.e. $\mathcal{Y} = \{1, 2, \ldots, N\}$ with $N \geq 2$, we define the vectors $\mathbf{p}_{\mathcal{S}}, \mathbf{p}_{\mathcal{T}} \in [0, 1]^N$ through $[\mathbf{p}_{\bullet}]_i = \mathbb{P}_{\bullet}(Y = i)$, i.e. through the label probabilities in the domains $\mathcal{S}$ and $\mathcal{T}$. Furthermore, we define a vector $\boldsymbol{\ell}_h \in \mathbb{R}_+^N$ of class-wise losses through $[\boldsymbol{\ell}_h]_i = \ell_X(h, i)$. The computation of the ACS-induced domain gap then simplifies to an absolute difference between scalar products $L_{\bullet}(h) = \sum_{i \in \mathcal{Y}} [\mathbf{p}_{\bullet}]_i [\boldsymbol{\ell}_h]_i = \langle \mathbf{p}_{\bullet}, \boldsymbol{\ell}_h \rangle$. Namely, for classification tasks:

$$|L_{\mathcal{T}}^{\text{clf}}(h) - L_{\mathcal{S}}^{\text{clf}}(h)| \;=\; \big| \langle \mathbf{p}_{\mathcal{T}}, \boldsymbol{\ell}_h \rangle - \langle \mathbf{p}_{\mathcal{S}}, \boldsymbol{\ell}_h \rangle \big| \tag{4}$$

Before we move on to a theorem about what Eq. 4 can mean in practice, let us build an intuition about the implications of this equation in a more simple setting: in binary classification.

Example 7 (Binary classification). In binary classification, the situation from Eq. 4 simplifies to $\mathcal{Y} = \{1, 2\}$ with $\mathbb{P}_{\bullet}(Y = 1) = p_{\bullet}$ and $\mathbb{P}_{\bullet}(Y = 2) = 1 - p_{\bullet}$. Let $\Delta p = |p_{\mathcal{T}} - p_{\mathcal{S}}|$ be be the absolute difference of the binary class proportions between the two domains and let $\Delta \ell_X = |\ell_X(h, 2) - \ell_X(h, 1)|$ be the absolute difference between the class-wise losses. The difference $\Delta \ell_X$ is independent of the class proportions and can be defined over any loss function ℓ. Rearranging Eq. 4 for binary classification, we obtain

$$\begin{aligned} &|L_{\mathcal{T}}^{\text{bin}}(h) - L_{\mathcal{S}}^{\text{bin}}(h)| \\ &=\; \big| \big(p_{\mathcal{T}} \ell_X(h, 2) + (1 - p_{\mathcal{T}}) \ell_X(h, 1)\big) \;-\; \big(p_{\mathcal{S}} \ell_X(h, 2) + (1 - p_{\mathcal{S}}) \ell_X(h, 1)\big) \big| \\ &=\; \big| \big(p_{\mathcal{T}} - p_{\mathcal{S}}\big) \cdot \big(\ell_X(h, 2) - \ell_X(h, 1)\big) \big| \\ &=\; \Delta p \cdot \Delta \ell_X, \end{aligned} \tag{5}$$

from which we see that in binary classification, for any loss function, the domain gap induced by ACS is simply the product of the class proportion difference Δp and the (true) class-wise loss difference $\Delta \ell_X$. If one of these terms is zero, so is the inter-domain gap. If one of these terms is non-zero but fixed, the domain gap will grow linearly with the other term.

Example 8 (Binary classification with zero-one loss). Let us illustrate Eq. 5 a little further. The zero-one loss is defined by $\ell(y, h(x)) = 0$ if the prediction is correct, i.e. if $y = h(x)$, and $\ell(y, h(x)) = 1$ otherwise. Consequently, $\ell_X(h, 2)$ is the true rate of false positives and $\ell_X(h, 1)$ is the true rate of false negatives. The more similar these rates are, the smaller will the inter-domain gap be for any distribution of classes in the target domain. Supposing that balanced training sets tend to balance $\ell_X(h, 2)$ and $\ell_X(h, 1)$, we can argue that balanced training sets (supposedly) maximize the range of feasible target domains with respect to the zero-one loss.

Example 9 (Cost-sensitive learning). The situation is quite different if the binary zero-one loss is weighted by the class, i.e. $\ell(y, h(x)) = w_y$ for $y \neq h(x)$. Such a weighting is common in cost-sensitive and imbalanced classification [7]. Here, Eq. 5 illustrates how counteracting class imbalance with weights can increase the robustness of the model: balancing $\ell_X(h, 2)$ and $\ell_X(h, 1)$ will increase the range of target domains that are feasible under the class-based weighting.

For completeness, we extend a part of this intuition from binary classification to classification tasks with an arbitrary number of classes:

Theorem 10. *In classification, the inter-domain gap $|L_{\mathcal{T}}^{clf}(h) - L_{\mathcal{S}}^{clf}(h)|$ from Theorem 5 is equal to zero if one of the following conditions holds:*

i) $\mathbf{p}_{\mathcal{S}} = \mathbf{p}_{\mathcal{T}}$

ii) $\ell_X(h, i) = \ell_X(h, j) \;\; \forall\, i, j \in \mathcal{Y}$

Proof. Condition i) trivially yields the claim through Eq. 4. Condition ii) means that $\Delta\ell = |\ell_X(h, i) - \ell_X(h, j)| = 0$ for every binary sub-task in a one-vs-one decomposition of the label set $\mathcal{Y}$. The domain gap of each binary sub-task, and therefore the total domain gap, is then zero according to Eq. 5. □

Despite the fact that condition 10.ii) yields a domain gap of zero, one should not prematurely jump to the conclusion that a learning algorithm should enforce this condition necessarily. Recall that Theorem 10 addresses the domain gap, but not the deployment loss which we actually want to minimize; if enforcing condition 10.ii) results in a high source domain error, all domain robustness will not help to find an accurate target domain model. We therefore advise practitioners to carefully weigh out the source domain error with the domain robustness of the model, depending on the requirements of the use case at hand. Bayesian classifiers, which allow practitioners to mimic arbitrary $\mathbf{p}_{\mathcal{S}}$ even after training, can prove useful in this regard.

2.2 Certification of Domain Robustness for Binary Predictors

We certify the set of class proportions to which a fixed hypothesis h, trained on $\mathcal{S}$, is safely applicable. By "safely", we mean that during the deployment on $\mathcal{T}$, h induces only a small domain-induced error with a high probability.

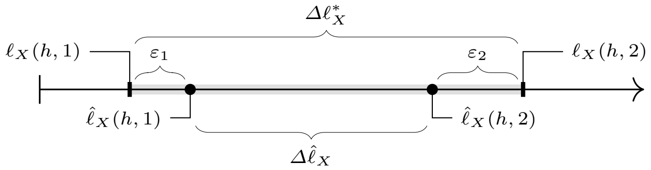

Fig. 3. Estimation of the minimum upper bound $\Delta\ell^*_X$ from data.

Definition 11 (Certified hypothesis). *A hypothesis $h \in \mathcal{H}$ is (ε, δ)-certified for all class proportions in the set $\mathcal{P} \subseteq [0,1]^N$ if with probability at least $1 - \delta$ and $\varepsilon, \delta > 0$:*

$$|L_\mathcal{T}(h) - L_\mathcal{S}(h)| \;\leq\; \varepsilon \quad \forall\, \mathbf{p}_\mathcal{T} \in \mathcal{P}$$

For simplicity, we limit our presentation to binary classification, i.e. $N = 2$ (see Example 7). In this case, $\mathcal{P}$ is simply a range $[\,p_\mathcal{T}^{\min},\, p_\mathcal{T}^{\max}]$ of class proportions. According to Eq. 5, this range is defined by the largest Δp^* for which

$$\Delta p \cdot \Delta\ell_X \;\leq\; \varepsilon \quad \forall\, \Delta p \;\leq\; \Delta p^*. \tag{6}$$

Keep in mind that $\Delta\ell_X$ is defined over the *true* class-wise losses. If we knew them, we could simply rearrange Eq. 6 to find the largest Δp for a given ε; the equation would then hold with probability one. However, we do not know the true class-wise losses; instead, we estimate an upper bound that is only exceeded by the true $\Delta\ell_X$ with a small probability of at most $\delta > 0$. Particularly, to maximize Δp^*, we find the *smallest* upper bound $\Delta\ell^*_X$ among all such upper bounds.

An empirical estimate $\Delta\hat{\ell}_X$ of the true $\Delta\ell_X$ is given by the empirical class-wise losses $\hat{\ell}_X(h,y)$ observed in an ACS-generated validation sample D:

$$\Delta\hat{\ell}_X \;=\; \left|\hat{\ell}_X(h,1) - \hat{\ell}_X(h,2)\right|, \quad \text{where} \quad \hat{\ell}_X(h,y) \;=\; \tfrac{1}{m_y} \textstyle\sum_{i\,:\,y_i=y} \ell(y, h(x_i))$$

Here, each $\hat{\ell}_X(h,y)$ can be associated with maximum lower and upper errors $\varepsilon_y^{(l)}, \varepsilon_y^{(u)} > 0$ that are not exceeded with probabilities at least $1-\delta_y^{(l)}$ and $1-\delta_y^{(u)}$. By choosing $\varepsilon_y^{(l)}, \varepsilon_y^{(u)}$ for both classes, we can thus find all upper bounds of the true $\Delta\ell_X$ that hold with at least the desired probability $1 - \delta$.

Fig. 3 sketches our estimation of the *smallest* upper bound $\Delta\ell^*_X$. For simplicity, we assume that $\hat{\ell}_X(h,2) \geq \hat{\ell}_X(h,1)$. This assumption comes without loss of generality because we can otherwise simply switch the labels to make the assumption hold. Now, $\hat{\ell}_X(h,1)$ shrinks at most by ε_1 and $\hat{\ell}_X(h,2)$ grows at most by ε_2. Minimizing ε_1 and ε_2 simultaneously, within a user-specified probability budget δ, yields the desired minimum upper bound $\Delta\ell^*_X$ which the true $\Delta\ell_X$ only exceeds with probability at most $\delta = \delta_1 + \delta_2 - \delta_1\delta_2$. We find the values of

δ_1 and δ_2 through Corollary 4, letting

$$\begin{aligned} &\underbrace{-\,(\hat{\ell}_X(h,2) - \hat{\ell}_X(h,1) + \varepsilon_1)}_{=\,\varepsilon_2^{(l)}} \;\leq\; \ell_X(h,2) - \hat{\ell}_X(h,2) \;\leq\; \underbrace{\varepsilon_2}_{=\,\varepsilon_2^{(u)}} \\ \text{and}\quad &\underbrace{-\,(\hat{\ell}_X(h,2) - \hat{\ell}_X(h,1) + \varepsilon_2)}_{=\,\varepsilon_1^{(u)}} \;\leq\; \hat{\ell}_X(h,1) - \ell_X(h,1) \;\leq\; \underbrace{\varepsilon_1}_{=\,\varepsilon_1^{(l)}}, \end{aligned}$$

so that $\delta_y = \delta_y^{(l)} + \delta_y^{(u)} - \delta_y^{(l)}\delta_y^{(u)}$ and $\delta_y^{(i)} = e^{-2m_y(\varepsilon_y^{(i)})^2}$.

During the optimization, strict inequalities are realized through non-strict inequalities with some sufficiently small $\tau > 0$:

$$\min_{\varepsilon_1,\,\varepsilon_2 \in \mathbb{R}} \;\varepsilon_2 + \varepsilon_1, \quad \text{s.t.} \begin{cases} \varepsilon_1,\, \varepsilon_2 & \geq\; \tau \\ \delta - (\delta_1 + \delta_2 - \delta_1\delta_2) & \geq\; 0 \end{cases} \tag{7}$$

The minimizer $(\varepsilon_1^*, \varepsilon_2^*)$ of this optimization problem defines the smallest upper bound $\Delta\ell_X^* = (\hat{\ell}_X(h,2) + \varepsilon_2^*) - (\hat{\ell}_X(h,1) - \varepsilon_1^*)$ that is not exceeded by the true $\Delta\ell_X$ with probability at least $1 - \delta$. Choosing $\Delta p^* = {}^{\varepsilon}/_{\Delta\ell_X^*}$ will make Eq. 6 hold with the same probability, so that the range $[p_{\mathcal{S}} - \Delta p^*,\, p_{\mathcal{S}} + \Delta p^*]$ of binary deployment class proportions $p_{\mathcal{T}}$ is (ε, δ)-certified according to Def. 11.

If only small data volumes are available, it can happen that ϵ_1 must exceed $\hat{\ell}_X(h,1)$ to stay within the user-specified probability budget δ. This situation would mean that the lower bound $\ell_X(h,1) = \hat{\ell}_X(h,1) - \varepsilon_1$ is below zero, which does not reflect the basic loss property $\ell(h,y) \geq 0$. If the estimation of $\Delta\ell^*$ fails in this way, we fall back to a more simple, one-sided estimation. Namely, we only minimize the two upper bounds $\varepsilon_y^{(u)}$ that depend only on ε_2 and fix the two lower bounds to $\varepsilon_y^{(l)} = 0$. Doing so allows us to estimate a valid upper bound $\Delta\ell^*$ also for arbitrarily small data sets.

3 Experiments

In the following, we show that an (ε, δ) certified class proportion set $\mathcal{P}$ indeed characterizes an upper bound of the inter-domain gap. Our experiments even demonstrate that our certificate, being estimated only with source domain data, is very close to bounds that are obtained with labeled target domain data and are therefore not accessible in practice.

3.1 Binary (ϵ, δ) Certificates are Tight

We randomly subsample the data to generate different deployment class proportions $p_{\mathcal{T}}$ while keeping $\mathbb{P}(x|y)$ fixed, in accordance to Def. 1. We compare two ways of estimating the target domain loss:

a) Our baseline is an empirical estimate $\hat{L}_{\mathcal{T}}$ of the target domain loss that is computed with actual target domain data unavailable in practice.

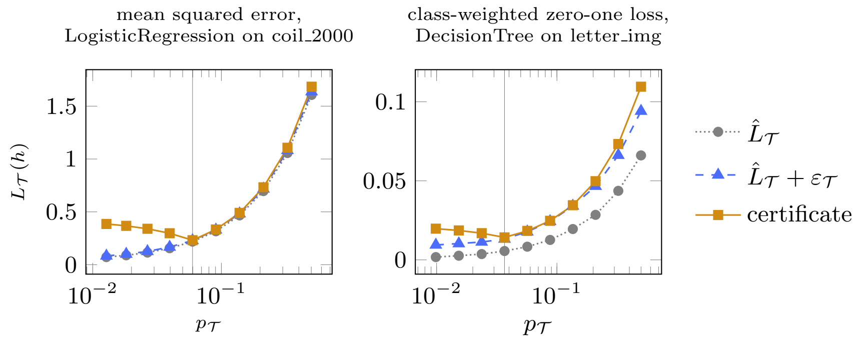

Fig. 4. The target domain loss $L_{\mathcal{T}}(h)$ is upper-bounded by our (ε, δ) certificate and a baseline $\hat{L}_{\mathcal{T}} + \varepsilon_{\mathcal{T}}$ with privileged access to target domain data. Each of the above plots displays a different combination of loss function, learning method, and data set. The class proportions $p_{\mathcal{T}}$ of the target domain are varied over the x axis with a thin vertical line indicating the source domain proportions $p_{\mathcal{S}}$.

b) We predict the target domain loss $L_{\mathcal{T}}$ from an (ε, δ) certificate by adding the domain gap parameter ε to the empirical source domain loss $\hat{L}_{\mathcal{S}}$. We always choose the certificates such that they cover the class proportion difference $\Delta p = |p_{\mathcal{T}} - p_{\mathcal{S}}|$; in fact, we consider ε as a function of Δp in this experiment.

The certificate is *correct* if $\hat{L}_{\mathcal{S}} + \varepsilon \geq \hat{L}_{\mathcal{T}}$ holds, i.e. if ε indeed characterizes an upper bound of the inter-domain gap. If the two values are close to each other, i.e. if $\hat{L}_{\mathcal{S}} + \varepsilon \approx \hat{L}_{\mathcal{T}}$, we speak of a *tight* upper bound.

Correctness: Our experiments cover a repeated three-fold cross validation on eight imbalanced data sets, eight loss functions, and three learning algorithms, to represent a broad range of scenarios. Of all 9000 certificates, only 4.5% fail the test of ensuring $\hat{L}_{\mathcal{S}} + \varepsilon \geq \hat{L}_{\mathcal{T}}$. Since we have used $\delta = 0.05$ in these experiments, this amount of failures is actually foreseen by the statistical nature of our certificate: if it holds with probability at least $1 - \delta$, it is allowed to fail in 5% of all tests. This margin is almost completely used but not exceeded. Consequently, our certificate is correct in the sense of indeed characterizing an upper bound ε of the inter-domain gap with probability at least $1 - \delta$.

Tightness: A *fair* comparison between our certificate and our baseline $\hat{L}_{\mathcal{T}}$ requires us to take the estimation error $\varepsilon_{\mathcal{T}}$ of the baseline into account. This necessity stems from the fact that $\hat{L}_{\mathcal{T}}$ is also just an estimate from a finite amount of data. Having access to labeled target domain data will thus yield an upper bound $\hat{L}_{\mathcal{T}} + \varepsilon_{\mathcal{T}}$ of the true target domain error $L_{\mathcal{T}}$, according to Proposition 3; this upper bound is then compared to our certificate, which has only seen the source domain data.

Fig. 4 presents this comparison for two of our experiments. For most target domains $p_{\mathcal{T}}$, the two predictions (■ and ▲) are almost indistinguishable from each other. This observation means that the certificate, which is based only on source domain data, is as accurate as estimating the target domain loss with a privileged access to labeled target domain data. Over all 9000 certificates, we find a mean absolute difference between the two predictions of merely 0.049; in fact, all supplementary plots look highly similar to those displayed in Fig. 4, despite covering many other data sets, loss functions, and learning methods. The margin to the left of each vertical line appears because our certificate covers an absolute inter-domain gap rather than a signed value.

3.2 Binary (ϵ, δ) Certificates in Astro-Particle Physics

The field of astro-particle physics studies the physical properties of cosmic particle accelerators such as active galactic nuclei and supernova remnants. Some of these accelerators produce gamma radiation, which physicists measure through imaging air Cherenkov telescopes (IACTs). Since IACTs also record non-gamma particles, it is necessary to separate the relevant gamma recordings from the non-gamma background. This task is commonly approached with classification models trained on simulated data [4]. The accurate physical simulations that are used for training produce telescope readings (feature vectors) from user-chosen particles (labels). The default approach to this ACS problem is to simulate a training set with balanced classes and to alter the decision threshold of the model after its training.

We apply our certification scheme to the FACT telescope [1], an IACT for which a big data set is publicly available. In particular, we reproduce the default analysis pipeline, fix δ to a small value (0.01 or 0.1), and select ε such that the resulting (ε, δ) certificate covers the anticipated class proportion difference $\mathit{\Delta} p = |p_{\mathcal{T}} - p_{\mathcal{S}}|$ between the simulated and the observed domain. For both δ values, we obtain similar ε values under the zero-one loss, namely $\varepsilon_{(\delta=0.01)} = 0.0315$ and $\varepsilon_{(\delta=0.1)} = 0.0313$. We conclude that the ACS-induced zero-one loss of the FACT pipeline is at most 3.15% with probability at least 99%, and at most 3.13% with probability at least 90%. The pipeline is trustworthy within these specific ranges and improvements to these certified values can be achieved by improving the performance of the pipeline. See Appendix 2 for additional details.

4 Related Work

Most of the previous work on ACS has focused on the empirical evaluation of heuristic data acquisition strategies [16, 13, 6, 21, 29, 10, 15, 28]. A recent theoretical contribution by us [5] is only valid for infinite data and lacks applicability to imbalanced domains; both of these issues motivate our present paper.

Model certification, in the broad sense of performance reports [17, 2, 22] and formal proofs of robustness [11, 26, 25, 30], has motivated us to study model robustness in ACS. Our certificate is only a single component in the more comprehensive reports that are conceived in the literature; yet, the certification of feasible class proportions is a trust-critical issue when the training class proportions are chosen arbitrarily. A related lane of research is concerned with the certification of learning algorithms [19, 18] instead of trained models.

Domain adaptation [27,20] assumes data from $\mathcal{T}$ with which a source domain model can be transferred to the target domain. If the data from $\mathcal{T}$ are unlabeled, it becomes necessary to employ additional assumptions about the differences between $\mathcal{S}$ and $\mathcal{T}$. For instance, our identical mechanism assumption from Def. 1 has also been introduced as the *target shift* assumption [31]. In ACS, we are free to choose the shift between $\mathbb{P}_{\mathcal{S}}(Y)$ and $\mathbb{P}_{\mathcal{T}}(Y)$ as small as permitted by our knowledge about $\mathcal{T}$, instead of having to adapt to $\mathcal{T}$. We conceive combinations of ACS and domain adaptation for future work.

Imbalanced learning [7] handles majority and minority classes differently from each other, so that the resulting classifier is not impaired by the disproportion between these classes. For instance, over-sampling the minority class with synthetic instances [8,3] will achieve more balanced training sets in which the minority class is not "overlooked" by the learning algorithm. In ACS, we can generate *actual* instances instead of synthetic ones; still, the idea of over-sampling can guide us in selecting the class proportions for an imbalanced target domain $\mathcal{T}$. Conversely, our certificate can guide imbalanced learners in choosing the amount of over-sampling or under-sampling to apply: the certified class proportion range $[\, p_{\mathcal{T}}^{\min}, p_{\mathcal{T}}^{\max}]$ should ideally cover the class proportions $p_{\mathcal{T}}$ that are expected during deployment; otherwise the sampling scheme can introduce a domain gap larger than ε, which impairs the target domain performance.

Cost-sensitive learning is often discussed as a means to tackle imbalanced learning (e.g. Chap. 4 in [7]) because many applications associate a disproportionally high cost with mis-classifications in the minority class. Our certificate supports these settings, without loss of generality, via class-wise loss weights.

Active learning [23] assumes that an oracle $\mathcal{X} \rightarrow \mathcal{Y}$ (e.g. a human expert) can label feature vectors after their acquisition. This assumption is fundamentally different from ACS, where a data generator $\mathcal{Y} \rightarrow \mathcal{X}$ produces feature vectors from labels. Still, some acquisition heuristics for ACS borrow from active learning strategies by aggregating the scores of pseudo-instances [13,15].

Quantification Learning [9] estimates class prevalences in the target domain, which can help in assessing the amount of label shift that is to be expected.

5 Conclusion

Motivated by a limited trust in active class selection, we have developed an (ε, δ) certificate for classifiers, which declares a set of class proportions to which the certified model can be safely applied. "Safely" means that the inter-domain gap induced by active class selection (or any other reason for a shift in the class proportions) is at most ε with probability at least $1 - \delta$. Our experiments show that the certificate is correct and bounds the true domain gap tightly.

So far, we have assumed that the loss function is decomposable over $\mathcal{X} \times \mathcal{Y}$, like the (weighted) zero-one loss, the hinge loss, and the mean squared error are. Future work should extend these results to loss functions that do not have this property, like the F_β and AUROC scores. We are also looking forward to extensions of our certificate towards multi-class settings and regression.

Appendix 1: Proof of the Identical Mechanism Bound

We draw a training set D of size m, where each individual example is drawn from $\mathcal{X} \times \mathcal{Y}$, according to $\mathbb{P}_{\mathcal{S}}$. Consequently, the full training set is drawn from $(\mathcal{X} \times \mathcal{Y})^m$, according to the probability density $\mathbb{P}_{\mathcal{S}}^m$. We are now interested in the probability that $\mathbb{P}_{\mathcal{S}}^m$ assigns to the event that all $h \in \mathcal{H}$ admit to the identical mechanism bound:

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}^m\Big(\big\{\mathrm{D} : \forall h \in \mathcal{H},\ \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ - \ \varepsilon \ &\leq |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| \\ &\leq |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ + \ \varepsilon\big\}\Big)\end{aligned}$$

We estimate the above probability from the probability of the converse event; if the above probability is p, then the following must be $1 - p$:

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}^m\Big(\big\{\mathrm{D} : \exists h \in \mathcal{H},\ \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ - \ \varepsilon \ &> \ |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| \\ \wedge \ \ |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| \ &> \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ + \ \varepsilon\big\}\Big)\end{aligned}$$

We now apply the union bound twice. This bound states that $\mathbb{P}(A \wedge B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for any two events A and B:

$$\begin{aligned}\ldots \quad &\leq \ \mathbb{P}_{\mathcal{S}}^m\Big(\big\{\mathrm{D} : \exists h \in \mathcal{H},\ \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ - \ \varepsilon \ > \ \ |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)|\big\}\Big) \\ &\quad + \mathbb{P}_{\mathcal{S}}^m\Big(\big\{\mathrm{D} : \exists h \in \mathcal{H},\ \ |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| \ > \ \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ + \ \varepsilon\big\}\Big) \\ &\leq \ \sum_{h \in \mathcal{H}} \mathbb{P}_{\mathcal{S}}^m\Big(\big\{\mathrm{D} : |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ - \ \varepsilon \ > \ \ |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)|\big\}\Big) \\ &\quad + \ \sum_{h \in \mathcal{H}} \mathbb{P}_{\mathcal{S}}^m\Big(\big\{\mathrm{D} : |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| \ > \ \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ + \ \varepsilon\big\}\Big)\end{aligned}$$

We have thus reduced the probability of the identical mechanism bound w.r.t. an entire hypothesis class $\mathcal{H}$ to a sum of probabilities for single hypotheses $h \in \mathcal{H}$. The single-hypothesis case has already been proven in Sec. 1. Let us restate this result here to clarify the connection: Each of the following statements describes a violation of the Sec. 1 bound, each having a probability of at most $2e^{-2m\varepsilon^2}$:

- $|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ - \ \varepsilon \ \ > \ \ |L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)|$
- $|L_{\mathcal{T}}(h) - L_{\mathrm{D}}(h)| \ \ > \ \ |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \ + \ \varepsilon$

These two events, together with their probabilities, can be plugged into the above transformation, which proves the claim:

$$\ldots \quad \leq \ \sum_{h \in \mathcal{H}} 2e^{-2m\varepsilon^2} + \sum_{h \in \mathcal{H}} 2e^{-2m\varepsilon^2} \ \ = \ \ 4|\mathcal{H}|e^{-2m\varepsilon^2} \qquad \square$$

Appendix 2: Experimental Details and Reproducibility

We provide an implementation of our proposed (ε, δ) certificate with the supplementary material of this paper. This material also contains a 56-page supplement of plots that can be reproduced with this implementation. All supplements are hosted at `https://github.com/mirkobunse/AcsCertificates.jl`.

The experiment in Sec. 3.1 verifies that (ε, δ) certificates are indeed correct and tight. However, by choosing ε as a function of Δp, we have "turned the certificate around"; in a usual application, a user would rather fix the ε value and look for a certified range Δp of feasible class proportions. Therefore, we provide the certified Δp values for all experiments in the supplementary material. Tab. 1 provides an excerpt of these values in which the certified target domain ranges $[p_\mathcal{S} - \Delta p^*,\, p_\mathcal{S} + \Delta p^*]$ induce a domain gap of at most $\varepsilon = 0.01$ with a probability of at least $1 - \delta = 0.95$. Since the domain gap is at most 0.01, we can expect a target domain loss of at most $L_\mathcal{S}(h) + 0.01$.

Table 1. Feasible class proportions Δp^*, according to (ε, δ) certificates that are computed for a class-weighted zero-one loss with $\varepsilon = 0.01$ and $\delta = 0.05$.

data	classifier	$L_\mathcal{S}(h)$	$p_\mathcal{S}$	Δp^*
coil_2000	LogisticRegression	0.0722	0.0597	0.0109
coil_2000	DecisionTree	0.0778	0.0597	0.0107
letter_img	LogisticRegression	0.0179	0.0367	0.0463
letter_img	DecisionTree	0.0139	0.0367	0.0504
optical_digits	LogisticRegression	0.0406	0.0986	0.0437
optical_digits	DecisionTree	0.0463	0.0986	0.0309
pen_digits	LogisticRegression	0.038	0.096	0.044
pen_digits	DecisionTree	0.0216	0.096	0.0695
protein_homo	LogisticRegression	0.0056	0.0089	0.036
protein_homo	DecisionTree	0.006	0.0089	0.0291
satimage	LogisticRegression	0.1205	0.0973	0.0118
satimage	DecisionTree	0.0763	0.0973	0.018

Tab. 2 presents the results of our astro-particle experiment. The significance of detection [14] is a domain-specific score which measures the effectiveness of the telescope. While higher values are better, 25σ are a usual value for accurate prediction models on the given data set. The fact that all ε values are close to each other stems from the large amount of source domain data (24000 examples) we use to certify the model.

Table 2. The parameters of an (ε, δ) certificate that covers the extreme class proportions $p_\mathcal{T} = 10^{-4}$ in astro-particle physics.

significance of detection $[\sigma]$	$L_\mathcal{S}(h)$	δ	ϵ_δ
25.067 ± 0.268	0.058 ± 0.015	0.01	0.0315
		0.025	0.0314
		0.05	0.0314
		0.1	0.0313

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. Anderhub, H., Backes, M., Biland, A., Boccone, V., Braun, I., Bretz, T., Buß, J., et al.: Design and operation of FACT–the first G-APD Cherenkov telescope. J. Inst. **8**(06) (2013)
2. Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., et al.: Factsheets: Increasing trust in AI services through supplier's declarations of conformity. IBM J. Res. Dev. **63**(4/5) (2019)
3. Bellinger, C., Sharma, S., Japkowicz, N., Zaïane, O.R.: Framework for extreme imbalance classification: SWIM - sampling with the majority class. Knowl. Inf. Syst. **62**(3) (2020)
4. 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)
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. Cakmak, M., Thomaz, A.L.: Designing robot learners that ask good questions. In: Int. Conf. on Human-Robot Interaction. ACM (2012)
7. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer (2018)
8. Fernández, A., García, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61** (2018)
9. González, P., Castaño, A., Chawla, N.V., del Coz, J.J.: A review on quantification learning. ACM Comput. Surv. **50**(5) (2017)
10. Hossain, I., Khosravi, A., Nahavandi, S.: Weighted informative inverse active class selection for motor imagery brain computer interface. In: Canadian Conf. on Electr. and Comp. Eng. IEEE (2017)
11. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In: Adv. in Neural Inf. Process. Syst. MIT Press (2006)
12. Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., et al.: A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Comput. Sci. Rev. **37** (2020)
13. 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)
14. Li, T.P., Ma, Y.Q.: Analysis methods for results in gamma-ray astronomy. Astrophysical J. **272** (1983)
15. Liu, S., Ding, W., Gao, F., Stepinski, T.F.: Adaptive selective learning for automatic identification of sub-kilometer craters. Neurocomputing **92** (2012)

16. 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)
17. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Conf. on Fairness, Accountability, and Transparency. ACM (2019)
18. Morik, K., Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Russell, S., Yeung, K.: Trustworthy AI, chap. 2. Springer (2021)
19. Morik, K., Kotthaus, H., Heppe, L., Heinrich, D., Fischer, R., Mücke, S., Pauly, A., Jakobs, M., Piatkowski, N.: Yes we care! – Certification for machine learning methods through the care label framework (2021), https://arxiv.org/abs/2105.10197
20. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10) (2010)
21. Parsons, T.D., Reinebold, J.L.: Adaptive virtual environments for neuropsychological assessment in serious games. IEEE Trans. Consumer Electron. **58**(2) (2012)
22. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Conf. on Fairness, Accountability, and Transparency. ACM (2020)
23. Settles, B.: Active Learning. Morgan & Claypool (2012)
24. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning - From Theory to Algorithms. Cambridge University Press (2014)
25. Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.T.: Fast and effective robustness certification. In: Adv. in Neural Inf. Process. Syst. (2018)
26. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. In: Adv. in Neural Inf. Process. Syst. (2020)
27. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312** (2018)
28. Weiss, G.M., Provost, F.J.: Learning when training data are costly: The effect of class distribution on tree induction. J. Artif. Intell. Res. **19** (2003)
29. Wu, D., Lance, B.J., Parsons, T.D.: Collaborative filtering for brain-computer interaction using transfer learning and active class selection **8**(2) (2013)
30. Zhang, D., Ye, M., Gong, C., Zhu, Z., Liu, Q.: Black-box certification with randomized smoothing: A functional optimization based framework. In: Adv. in Neural Inf. Process. Syst. (2020)
31. Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain adaptation under target and conditional shift. In: Int. Conf. on Mach. Learn. (2013)