Certifiable Active Class Selection in Multi-Class Classification

Martin Senz[0000−0002−9377−3939], Mirko Bunse[0000−0002−5515−6278], and Katharina Morik[0000−0003−1153−5986]

TU Dortmund University, Artificial Intelligence Group, D-44227 Dortmund, Germany
{martin.senz, mirko.bunse, katharina.morik}@tu-dortmund.de

Abstract. Active class selection (ACS) requires the developer of a classifier to actively choose the class proportions of the training data. This freedom of choice puts the trust in the trained classifier at risk if the true class proportions, which occur during deployment, are subject to uncertainties. This issue has recently motivated a *certificate* for ACS-trained classifiers, which builds trust by proving that a classifier is sufficiently correct within a specific set of class proportions and with a high probability. However, this certificate was only developed in the context of binary classification. In this paper, we employ Hölder's inequality to extend the binary ACS certificate to multi-class settings. We demonstrate that our extension indeed provides correct and tight upper bounds of the classifier's error. We conclude with several directions for future work.

Keywords: Active class selection · Prior probability shift · Multi-class classification · Model certification · Learning theory · Validation.

1 Introduction

The proceeding deployment of machine learning models in real-world applications increases the importance of validating these models thoroughly. Ideally, the robustness of these models against distribution shifts [5] is *certified* in the sense of being formally proven or extensively tested [3].

In active class selection [4], a class-conditional data generator is repeatedly asked to produce feature vectors for arbitrarily chosen classes. In this setting, the developer of a classifier must actively decide for the class proportions in which the training data set is produced. While this freedom can reduce the data acquisition cost while improving classification performance, it also puts the trust in the trained classifier at risk: what if the class proportions, which occur during deployment, are not precisely known or are even subject to changes?

These uncertainties have motivated a *certificate* for ACS-trained classifiers, which declares a set of class proportions to which a classifier is safely applicable [2]. In particular, the certified classifier is required to exhibit an ACS-induced error of at most some $\epsilon > 0$, with a probability of at least $1 - \delta$. However, this certificate was only developed in the context of binary classification; a multi-class certificate has not yet been proposed, to the best of our knowledge.

In this paper, we close the gap between ACS model certification and multi-class classification. In the following, we recapitulate the theoretical background of binary ACS certification in Section 2 before we develop our multi-class ACS certificate in Section 3. We validate our claims empirically in Section 4 before we conclude with Section 5.

2 Theoretical Background

The term "domain", as used in domain adaption [6], describes a probability density function over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space. 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 $p_{\mathcal{S}} \neq p_{\mathcal{T}}$ [2]. Such deviations, also known as target shift [8] or as prior probability shift [5], occur due to the freedom of choosing any $p_{\mathcal{S}}$ for the acquisition of training data. We are interested in the impact of such deviations on the classification performance with respect to $\mathcal{T}$.

Recently, a PAC learning perspective [2] on this setting has provided us with Theorem 1. 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 $|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)|$, which is independent of the random draw of D from $\mathcal{S}$, 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. Since the inter-domain gap is constant with respect to the random draw of the training set D, it is also independent of ϵ, δ, and m.

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

$$|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| - \epsilon \;\leq\; |L_{\mathcal{T}}(h) - L_{\mathcal{D}}(h)| \;\leq\; |L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| + \epsilon.$$

This theorem can be used to certify a trained classification model h with N classes in terms of a set of safe class proportions $\mathcal{P} \subseteq [0,1]^N$. By "safe", we mean that, during the deployment of h on $\mathcal{T}$, the trained model induces, with a high probability, at most a small domain-induced error ϵ.

Definition 1 (Certified hypothesis [2]). *A hypothesis $h \in \mathcal{H}$ is certified for all class proportions in $\mathcal{P} \subseteq [0,1]^N$ if, with probability at least $1-\delta$ and $\epsilon, \delta > 0$,*

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

Let $\mathbf{p}_S$, $\mathbf{p}_{\mathcal{T}} \in [0,1]^N$ be vectors with components $[\mathbf{p}_\bullet]_i = \mathbb{P}_\bullet(Y = i)$, which express the probabilities of the class labels in the respective domains S and $\mathcal{T}$.

Furthermore, let $\boldsymbol{\ell}_h \in \mathbb{R}^N$ be a vector that represents the class-wise losses

$$[\boldsymbol{\ell}_h]_i \;=\; \ell_X(h,i) \;=\; \int_{\mathcal{X}} \mathbb{P}(X = \mathbf{x} \mid Y = i) \cdot \ell(h(\mathbf{x}), i)\, \mathrm{d}\mathbf{x}, \tag{1}$$

as according to some loss function ℓ. The total loss of the hypothesis h is then given by $L_\bullet(h) = \sum_{i \in \mathcal{Y}} [\mathbf{p}_\bullet]_i [\boldsymbol{\ell}_h]_i = \langle \mathbf{p}_\bullet, \boldsymbol{\ell}_h \rangle$. Consequently the inter-domain gap for classification problems can be expressed as

$$\begin{aligned} |L_\mathcal{T}(h) - L_S(h)| \;&=\; |\langle \mathbf{p}_\mathcal{T}, \boldsymbol{\ell}_h \rangle - \langle \mathbf{p}_\mathcal{S}, \boldsymbol{\ell}_h \rangle| \\ &=\; |\langle \mathbf{p}_\mathcal{T} - \mathbf{p}_\mathcal{S}, \boldsymbol{\ell}_h \rangle| \\ &=\; |\langle \mathbf{d}, \boldsymbol{\ell}_h \rangle|, \end{aligned} \tag{2}$$

where $\mathbf{d} = \mathbf{p}_\mathcal{T} - \mathbf{p}_\mathcal{S}$ is the difference between the class probabilities in the domains $\mathcal{S}$ and $\mathcal{T}$.

In order to certify classification models, it is necessary to calculate Eq. 2. However, the true class-wise losses $\boldsymbol{\ell}_h$ are unknown, and we can only estimate the empirical class-wise losses $\hat{\ell}_X(h,y) = \frac{1}{m_y} \sum_{i:y_i=y} \ell(y, h(\mathbf{x}_i))$ from a finite amount of labeled validation data. Therefore, our goal is to constrain Eq. 2 with the smallest upper bound, which holds with a high probability.

For binary classification problems, the inter-domain gap can be factorized into a product of two scalars, $\Delta p \cdot \Delta \ell_X$. Here, $\Delta p = |p_\mathcal{T} - p_\mathcal{S}| \in \mathbb{R}$ denotes the difference between class proportions and $\Delta \ell = |\ell_{Y=2}(h) - \ell_{Y=1}(h)| \in \mathbb{R}$ denotes the difference between class-wise losses. A smallest upper bound $\Delta \ell^*$, which holds with probability $1 - \delta$, can be found for the empirical estimate $\Delta \hat{\ell}$. Therefore, by Def. 1, binary classifiers can be certified as a function of ϵ and δ, where $\mathcal{P}$ is characterized by the range $[p_\mathcal{T}^{\min}, p_\mathcal{T}^{\max}]$ of class proportions [2].

3 Certification in Multi-Class Classification

To certify multi-class classifiers according to Def. 1, an estimation for the inter-domain gap with multiple classes must be found. For this purpose, we will make use of Hölders inequality [7], a fundamental inequality theorem for the study of L^p spaces. This inequality will help us in using PAC bounds for multi-class certification, similar to the certification of binary classifiers.

Theorem 2 (Hölder's inequality [7]). *Let (S, Σ, μ) be a measure space and let $p, q \in [1, \infty]$ with $1/p + 1/q = 1$, where $1/\infty = 0$. Then, for all measurable real- or complex-valued functions f and g on S,*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q. \tag{3}$$

With this inequality, the inter-domain gap from Eq. 2 can be transformed to

$$|\langle \mathbf{d}, \boldsymbol{\ell}_h \rangle| \;\leq\; \begin{cases} \|\mathbf{d}\|_1 \cdot \|\boldsymbol{\ell}_h\|_\infty, & \text{for } p = 1, q = \infty \\ \|\mathbf{d}\|_2 \cdot \|\boldsymbol{\ell}_h\|_2, & \text{for } p = 2, q = 2 \\ \|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1, & \text{for } p = \infty, q = 1 \end{cases} \tag{4}$$

In the following we restrict ourselves to the consideration of the Hölder conjugate $p = \infty, q = 1$. In principle, the other conjugate forms are also applicable. However, we will see that the infinity norm on $\mathbf{d}$ provides a simple and intuitive characterization of $\mathcal{P}$.

In order to yield a certified hypothesis, as according to Def. 1, it must hold, with a probability of at least $1 - \delta$, that, for $\epsilon, \delta > 0$,

$$|L_{\mathcal{T}}(h) - L_{\mathcal{S}}(h)| \;\leq\; \|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1 \;\leq\; \epsilon \quad \forall\, \mathbf{p}_{\mathcal{T}} \in \mathcal{P}. \tag{5}$$

Like in the binary setting, only the empirical class-wise loss $\hat{\boldsymbol{\ell}}_h$ is given. Hence, a minimum upper bound $\|\boldsymbol{\ell}_h\|_1^*$ for the norm $\|\hat{\boldsymbol{\ell}}_h\|_1$, that is valid with a probability of at least $1 - \delta$, must be found. Each $\hat{\ell}(h, y)$ is associated with a positive corresponding error ϵ_y with $\delta_y = e^{-2m_y\epsilon_y^2}$. For a given probability budget of δ, we get the smallest upper bound $\|\boldsymbol{\ell}_h\|_1^* = \|\hat{\boldsymbol{\ell}}_h\|_1 + \sum_{y=1}^N \epsilon_y^*$ by minimizing $\sum_{y=1}^N \epsilon_y$ through the optimization problem

$$\min_{\epsilon_1,\ldots\epsilon_N \in \mathbb{R}} \sum_{y=1}^{N} \epsilon_y, \quad \text{s. t.} \begin{cases} \epsilon_1, \ldots \epsilon_N & \geq \tau \\ \delta - \sum_{y=1}^N \delta_y = e^{-2m_y\epsilon_y^2} & \geq 0 \end{cases}, \tag{6}$$

where strict inequalities are realized through non-strict inequalities with some sufficiently small $\tau > 0$.

Let us now describe the set $\mathcal{P}$ of safe class proportions. In extension to the requirement given in Def. 1, $\mathcal{P}$ is supposed to cover all class proportions that are valid according to the certificate. With the minimum upper bound $\|\boldsymbol{\ell}_h\|_1^*$, we can rearrange Eq. 5 to

$$\|\mathbf{d}\|_\infty \;\leq\; \frac{\epsilon}{\|\boldsymbol{\ell}_h\|_1^*} \quad \forall\, \mathbf{p}_{\mathcal{T}} \in \mathcal{P}. \tag{7}$$

By taking the infinity norm on $\mathbf{d}$, $\|\mathbf{d}\|_\infty$ reduces to the class i which has the largest absolute *label distribution shift* $|[\mathbf{p}_{\mathcal{T}}]_i - [\mathbf{p}_{\mathcal{S}}]_i| = \Delta p$. In analogy to the binary certification, the range of safe deployment proportions for a class i can be described by $[\,[\mathbf{p}_S]_i - \Delta p^*\,,\, [\mathbf{p}_{\mathcal{S}}]_i + \Delta p^*] = [p_{\mathcal{T},i}^{\min}, p_{\mathcal{T},i}^{\max}]$. Here, $\Delta p^* = \frac{\epsilon}{\|\boldsymbol{\ell}_h\|_1^*}$ is constant for all classes and represents the largest absolute shift that a class is allowed to have to satisfy Eq. 5 with probability at least $1 - \delta$. Therefore,

$$\mathcal{P} = \left\{ \mathbf{p} \in [0,1]^N \,:\, [\mathbf{p}]_i \in [p_{\mathcal{T},i}^{\min}, p_{\mathcal{T},i}^{\max}] \;\forall\, i \in \{1, \ldots, N\} \text{ and } \sum_{i=1}^{N} [\mathbf{p}]_i = 1 \right\} \tag{8}$$

defines the set of class proportions to which the certified classifier h is safely applicable.

Based on this approach, a variant of the certificate can be derived by modifying $\mathbf{d}$ slightly. For this modification, the negative vector components of $\mathbf{d}$ are set to zero, so that a vector $\mathbf{d}_+$ is formed. This variant is motivated by the observation that, by applying the norm to $\mathbf{d}$, the negative loss components (falsely) contribute as positives to the estimation of the error. Accordingly, $\mathbf{d}_+$ addresses

only the positive error component and allows a more tighter estimate of the inter-domain gap. However, since with $\mathbf{d}_+$ only the positive error components are considered, the range of class proportions can no longer be expressed by $[p^{\min}_{\mathcal{T},i}, p^{\max}_{\mathcal{T},i}]$ and $\mathcal{P}_{\mathbf{d}_+}$ cannot be defined by Eq. 8. As a consequence, $\mathcal{P}_{\mathbf{d}_+}$ is more difficult to characterize than $\mathcal{P}$.

4 Experiments

In the following evaluation, we show that the introduced multi-class certificate indeed represents an upper bound of the inter-domain gap. Besides the correctness of the certificate, the accuracy and tightness of the estimated upper bound are inspected. Ideally, the certificates correspond to upper bounds that are both correct and tight. To this end, we randomly subsample the data to generate different deployment class proportions $\mathbf{p}_\mathcal{T}$ while keeping $\mathbb{P}(X = \mathbf{x} \mid Y = y)$ fixed. To facilitate visualizations in two dimensions, we limit our evaluation to data sets with three classes. The implementation of our configurable experiments is available online[1].

Correctness

The certificate is correct if $\hat{L}_\mathcal{S} + \epsilon \geq \hat{L}_\mathcal{T}$ holds, where ϵ is the predicted domain-induced error and $\hat{L}_\mathcal{T}$ is the empirical estimate of the target domain loss [2]. At this point, recognize that computing $\hat{L}_\mathcal{T}$ requires target domain data, which is typically *not* available in ACS. This unavailability raises the desire for an upper bound $\hat{L}_\mathcal{S} + \epsilon$ of $\hat{L}_\mathcal{T}$, which allows practitioners to assess, using only ACS-generated data from $\mathcal{S}$, whether their classifier is sufficiently accurate on $\mathcal{T}$. Our certificate is designed to provide this upper bound, and the purpose of our experiments is to validate this claim.

Our experiments cover a repeated three-fold cross validation on six data sets and two learning algorithms, to represent a broad range of scenarios. In total, we have generated 216 000 certificates under the zero-one loss with $\delta = 0.05$. Among these certificates, only one failed, by producing an $\hat{L}_\mathcal{S} + \epsilon$ that is larger than $\hat{L}_\mathcal{T}$. Due to the statistical nature of our certificates, $\delta = 0.05$ would have allowed for up to 10 800 failures. Therefore, the number of failures is much smaller than expected. This small number of failures results from the coarse bound estimation that Hölder's inequality provides.

Tightness

A fair comparison between our certificates and our empirical estimate $\hat{L}_\mathcal{T}$ requires us to take the estimation error $\epsilon_\mathcal{T}$ of the baseline, $\hat{L}_\mathcal{T}$, into account [2]. This necessity stems from the fact that $\hat{L}_\mathcal{T}$ is just an estimate from a finite amount of data. Having access to labeled target domain data would thus yield

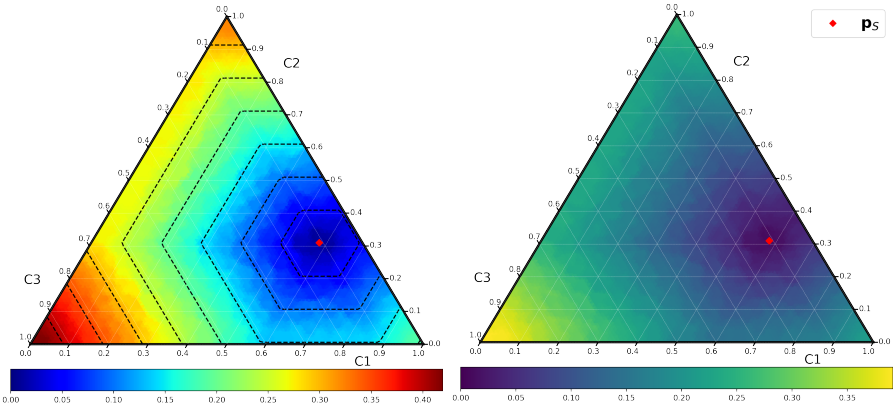
[1] https://github.com/martinsenz/MultiClassAcsCertificates

Table 1: Feasible class proportions Δp^*, according to $\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$ certificates, which are computed for a zero-one loss with $\epsilon = 0.1$ and $\delta = 0.05$.

data set	classifier	$L_S(h)$	$\mathbf{p}_S^\top$	Δp^*
optdigits	DecisionTree	0.100225	$[0.70, 0.20, 0.10]$	0.174888
optdigits	LogisticRegression	0.0955851	$[0.70, 0.20, 0.10]$	0.19131
satimage	DecisionTree	0.10647	$[0.58, 0.31, 0.11]$	0.2285
satimage	LogisticRegression	0.106242	$[0.58, 0.31, 0.11]$	0.217247
pendigits	DecisionTree	0.0467701	$[0.70, 0.20, 0.10]$	0.34621
pendigits	LogisticRegression	0.160971	$[0.70, 0.20, 0.10]$	0.137843
eye movements	DecisionTree	0.488188	$[0.35, 0.26, 0.40]$	0.0652661
eye movements	LogisticRegression	0.515892	$[0.35, 0.26, 0.40]$	0.061098
shuttle	DecisionTree	0.00521672	$[0.15, 0.79, 0.06]$	1.29197
shuttle	LogisticRegression	0.0573444	$[0.15, 0.79, 0.06]$	0.251336
connect4	DecisionTree	0.297242	$[0.65, 0.1, 0.25]$	0.0700096
connect4	LogisticRegression	0.343249	$[0.65, 0.1, 0.25]$	0.0491499

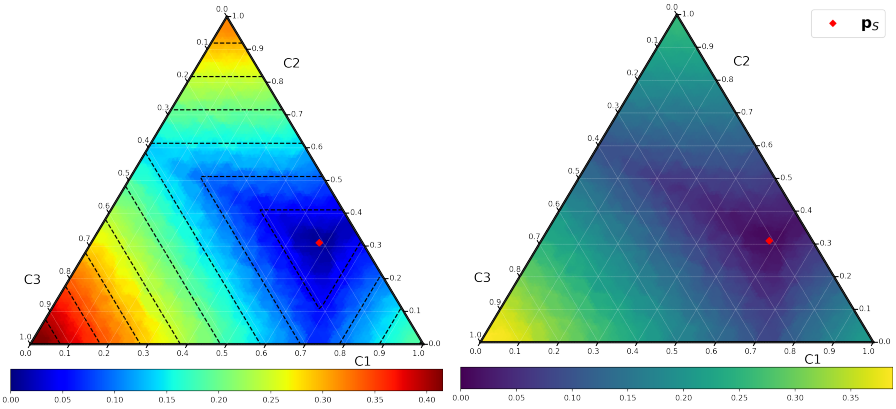
Table 2: MAD and quartiles of the absolute difference between $\hat{L}_S+\epsilon$ and $\hat{L}_\mathcal{T}+\epsilon_\mathcal{T}$.

data set	method	MAD	Q_1	Q_2	Q_3
optdigits	$\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.2023 ± 0.0954	0.1258	0.2015	0.2744
optdigits	$\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.18 ± 0.1006	0.1009	0.1607	0.2471
satimage	$\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.1809 ± 0.087	0.1218	0.1741	0.2324
satimage	$\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.1661 ± 0.0908	0.0999	0.1513	0.22
pendigits	$\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.1999 ± 0.1445	0.1034	0.168	0.2357
pendigits	$\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.1703 ± 0.1452	0.0751	0.1319	0.2034
eye movements	$\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.5433 ± 0.245	0.3639	0.5239	0.7331
eye movements	$\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.5207 ± 0.2643	0.3058	0.5006	0.7369
shuttle	$\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.0879 ± 0.0836	0.0315	0.0531	0.1203
shuttle	$\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.071 ± 0.0792	0.0223	0.0424	0.0825
connect4	$\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.5094 ± 0.221	0.3419	0.5167	0.6541
connect4	$\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$	0.4331 ± 0.2515	0.2274	0.405	0.613



Certified domain-induced error ϵ Tightness $= |\hat{L}_S + \epsilon - L_{\mathcal{T}}|$

Fig. 1: The certified error (left) and its tightness (right), according to $\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$ on the satimage data set, using a logistic regression and the zero-one loss.



Certified domain-induced error ϵ Tightness $= |\hat{L}_S + \epsilon - L_{\mathcal{T}}|$

Fig. 2: The certified error (left) and its tightness (right), according to the variant $\|\mathbf{d}_+\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$ on the satimage data set, using a logistic regression and the zero-one loss.

an upper bound $\hat{L}_{\mathcal{T}} + \epsilon_{\mathcal{T}}$ of the true target domain error $L_{\mathcal{T}}$. We speak of a *tight* bound, if $\hat{L}_S + \epsilon \approx \hat{L}_{\mathcal{T}} + \epsilon_{\mathcal{T}}$.

For example, the prediction of the domain induced error ϵ, as according to our certificate, can be inspected in Fig. 1. The prediction by our $\mathbf{d}_+$ certificate variant is shown in Fig. 2. As we can see, the upper bound is very tight for the area around $\mathbf{p}_S$. With increasing distance from $\mathbf{p}_S$, the estimation of the upper bound becomes larger, and hence, the upper bound becomes coarser. As it is expected, the variant using the $\mathbf{d}_+$ vector provides a finer bound of the inter-domain gap in some regions. Tab. 2 summarizes the absolute deviations between $\hat{L}_S + \epsilon$ and $\hat{L}_{\mathcal{T}} + \epsilon_{\mathcal{T}}$ in terms of mean absolute deviation (MAD) and quartiles (Q_1, Q_2, Q_3).

5 Conclusion and Outlook

Using Hölder's inequality and considering PAC bounds, we have proposed an upper bound $\|\mathbf{d}\|_\infty \cdot \|\boldsymbol{\ell}_h\|_1$, from which certificates of model robustness in multi-class ACS can be issued. Our experiments demonstrate that this certification is correct within a probability budget δ. Moreover, safe class proportions can easily be described by the maximum allowable absolute deviation Δp^*. Thus, the certification of a multi-class ACS classifier is straightforward for the practitioner to interpret and intuitive to understand, regardless of the number of classes considered in the classification problem.

By decomposing the inter-domain gap into positive and negative error components, it is possible to find estimates that bound the domain gap even more precisely. An example is the presented $\mathbf{d}_+$ certification variant, which considers only the positive error components. In order to obtain even more precise estimates, it is further conceivable to also take the negative error components (correctly) into account. However, as already indicated by the $\mathbf{d}_+$ variant, the complexity of describing the set $\mathcal{P}$ of valid class proportions increases with the expression strength of the upper bound.

In future work, we plan to evaluate more precise estimates of this kind, as well as the other bounds that are provided by Hölder's inequality in Eq. 4. We also plan to use our multi-class certificates as a basis for theoretically justified data acquisition strategies for multi-class ACS, similar to the binary acquisition strategy that is based on binary certificates [1].

References

1. Bunse, M., Morik, K.: Active class selection with uncertain deployment class proportions. In: Workshop on Interactive Adaptive Learning. p. 70 (2021)
2. Bunse, M., Morik, K.: Certification of model robustness in active class selection. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 266–281. Springer (2021)
3. Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., Yi, X.: A survey of safety and trustworthiness of deep neural networks: Verification, testing,

adversarial attack and defence, and interpretability. Computer Science Review **37** (2020)
4. Lomasky, R., Brodley, C.E., Aernecke, M., Walt, D., Friedl, M.: Active class selection. In: European Conference on Machine Learning. pp. 640–647. Springer (2007)
5. Moreno-Torres, J.G., Raeder, T., Alaíz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognition **45**(1), 521–530 (2012)
6. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010)
7. Yang, W.H.: On generalized Hölder inequality. Nonlinear Analysis: Theory, Methods & Applications **16**(5), 489–498 (1991)
8. Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain adaptation under target and conditional shift. In: International Conference on Machine Learning. JMLR Workshop and Conference Proceedings, vol. 28, pp. 819–827 (2013)