The Hitchhiker's Guide to Prior-Shift Adaptation

Tomáš Šipka, Milan Šulc and Jiří Matas
Czech Technical University in Prague
sipka3@seznam.cz, milansulc01@gmail.com, matas@fel.cvut.cz

Abstract

In many computer vision classification tasks, class priors at test time often differ from priors on the training set. In the case of such prior shift, classifiers must be adapted correspondingly to maintain close to optimal performance. This paper analyzes methods for adaptation of probabilistic classifiers to new priors and for estimating new priors on an unlabeled test set. We propose a novel method to address a known issue of prior estimation methods based on confusion matrices, where inconsistent estimates of decision probabilities and confusion matrices lead to negative values in the estimated priors. Experiments on fine-grained image classification datasets provide insight into the best practice of prior shift estimation and classifier adaptation, and show that the proposed method achieves state-of-the-art results in prior adaptation. Applying the best practice to two tasks with naturally imbalanced priors, learning from web-crawled images and plant species classification, increased the recognition accuracy by 1.1% and 3.4% respectively.

1. Introduction

Let us consider probabilistic classifiers that estimate the posterior probability $p(Y|X)$, where X and Y are random variables describing an observation and its correct class label respectively. In the framework of empirical risk minimization, the classifier parameters are trained by minimizing the loss function on a training set, which is assumed to come from the same distribution as the expected test data. However, in many classification tasks, the class prior probabilities $p_{\mathcal{E}}(Y)$ on an evaluation set (test set) differ from $p_{\mathcal{T}}(Y)$ on the training set, while the class-conditional distributions remain unchanged, i.e. $p_{\mathcal{E}}(X|Y) = p_{\mathcal{T}}(X|Y)$. In case of this phenomenon, called *prior shift* or *label shift*, classifiers require adaptation to maintain close to optimal performance.

For example, let us assume symptoms X common to several diseases with fixed conditional probabilities $p(X|Y)$. If we use a previously trained classifier during an outbreak of a disease, the classification results should change according to the new prior probabilities $p_{\mathcal{E}}(Y)$. As another example, let us consider species classification, where specimens of a species have the same appearance model $p(X|Y)$, yet the species incidence $p_{\mathcal{E}}(Y)$ may change according to location, time of year, and other environmental factors.

Other domain adaptation scenarios considered in the literature include *covariate shift* [23, 19], also called *sample selection bias* – when the appearance model $p(X)$ changes, but the conditional output distribution $p(Y|X)$ is invariant; and *conditional shift* [24] – where $p(Y)$ remains the same, but $p(X|Y)$ changes. This paper focuses solely on the problem of *prior shift*, also denoted *label shift* or *target shift*.

Class priors often follow a long-tail (LT) distribution. Classification on the less frequent classes can be improved by specific imbalanced/long-tail data losses, such as Focal loss [13] or LDAM loss [3], and training methods, such as OLTR [15] or BLT [11]. While we experiment with prior shifts from and to LT distributions, we focus on test-time adaptation of pre-trained classifiers with outputs approximating posterior probabilities, trained by cross entropy loss minimization. The paper does not aim at improving long-tail training methods, which is a different task than prior shift adaptation, where the shift can happen between arbitrary class distribution, not only long-tailed. We consider the standard classification task, i.e. a Bayesian decision problem with a 0/1 loss. We thus aim at improving the accuracy of the pre-trained classifier after prior shift.

As the first contribution, the paper summarizes existing methods for adaptation to prior shift and experimentally answers questions about the best practice, such as:

- Should an existing classifier be adapted to new priors by re-weighting its predictions, or is it worth it to re-train the classifier with training sampling matching the shift?
- What is the best practice to estimate, in an unsupervised way, class priors on a test set?
- Is it better to directly estimate test priors, or shall the importance weights [14] be estimated?
- How does the estimate quality depend on the test set size?

As the second contribution, we propose a maximum like-

Code is available at https://github.com/sipkatom/The-Hitchhiker-s-Guide-to-Prior-Shift-Adaptation

lihood approach for correcting all estimates based on the inversion of confusion matrix [6, 18, 21], where inconsistent estimates of decision probabilities and confusion matrices could result in negative values. Vucetic and Obradovic [21], who use a bootstrapping framework, avoid such infeasible solutions by discarding corresponding bootstrap replicates. McLachlan [16] and Forman [6] mention clipping the estimate into the range 0–100%. To the best of our knowledge, this paper is the first to provide a well-grounded solution to this problem. The proposed method, (S)CM$^{\mathrm{L}}$, achieves state-of-the-art results, in most cases performing better than existing methods including the "Hard-To-Beat" EM with Bias-Corrected Calibration [1].

As the third contribution, we propose a Maximum A-Posteriori estimation method (S)CM$^{\mathrm{M}}$, extending the proposed CM-based likelihood maximization (S)CM$^{\mathrm{L}}$ by adding a hyper-prior. We show that a Dirichlet hyper-prior, as in [20], improves the estimation of dense distributions, and performs better than the existing MAP estimation [20].

2. Related Work

Several methods have been proposed to tackle adaptation to prior shift, either by adapting the predictions of a pre-trained classifier [5, 18, 20, 21] or by re-training the classifier with adjusted training sample weights [2, 14].

The new priors are commonly unknown, but can be estimated from an unlabeled set of observations. Vucetic and Obradovic [21] estimate the new priors $p_{\mathcal{E}}(Y)$ from the classifier's probabilities $p(D = i|Y = k)$ of predicting class i when the true class is k, i.e. from the confusion matrix (CM) of the classifier. Saerens et al. [18] propose an EM algorithm for Maximum Likelihood Estimation (MLE) of priors from predictions of posterior probabilities $p(Y|X)$, and experimentally show an improvement compared to no adaptation and compared to a Confusion Matrix based estimate. Du Plessis and Sugiyama [5] prove that the EM procedure [18] is equivalent to fixed-point-iteration minimization of the KL divergence between the new input density $p_{\mathcal{E}}(X)$ and the marginalization of the joint distribution $\sum_Y p_{\mathcal{E}}(Y)p(X|Y)$. Du Plessis and Sugiyama [5] also propose methods for direct divergence minimization in cases where the input density $p(x)$ can be directly modeled, e.g. with a kernel-based non-parametric estimate. Sulc and Matas [20] emphasize the importance of adapting to new priors in fine-grained image classification and propose a Maximum A Posteriori (MAP) estimation adding a Dirichlet hyper-prior.

Lipton et al. [14] propose a confusion matrix based estimate of the prior ratio $w(Y) = \frac{p_{\mathcal{E}}(Y)}{p_{\mathcal{T}}(Y)}$, forming a Black Box Shift Learning (BBSL) framework. Azizzadenesheli et al. [2] propose to increase the stability of the prior ratio estimation by regularizing the distribution shift, forming the Regularized Learning under Label Shifts (RLLS) framework. Both methods [2, 14] then re-train the classifier with sample weights according to the prior ratio.

Alexandri et al. [1] propose a bias-corrected version of temperature scaling for classifier calibration and use the EM algorithm for prior estimation on the calibrated predictions. They experimentally show that adapting a calibrated classifier to new priors estimated by EM outperforms the re-trained classifiers using BBSL [14] and RLLS [2], making EM with Bias-Corrected Calibration "Hard-To-Beat".

In the following subsections, we assess the existing methods and formulate them in a unified notation.

2.1. Classifier Adaptation

Let $\mathbb{X}$ be a feature space, K the number of classes and $\mathbf{f} : \mathbb{X} \to \Delta_{K-1}$ a classifier mapping observations $\mathbf{x} \in \mathbb{X}$ onto the probability simplex Δ_{K-1}. The classifier is trained to approximate class posteriors $f(\mathbf{x}) \approx p(Y|\mathbf{x})$, e.g. by cross-entropy minimization. The training set $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ is sampled from distribution $p_{\mathcal{T}}(X, Y)$. In the case of *prior shift*, the priors on evaluation set $\mathcal{E} = \{\mathbf{x}_i\}_{i=1}^M$ change to $p_{\mathcal{E}}(Y)$, while the appearance model $p_{\mathcal{T}}(X|Y) = p_{\mathcal{E}}(X|Y)$ remains the same.

We consider different cases of classifier adaptation, where the new priors $p_{\mathcal{E}}(Y)$ are either known or unknown, and the classifier:

1. is fixed and trained on a known training set $\mathcal{T}$,
2. will be trained on the training set $\mathcal{T}$ and we can change the training procedure,
3. is fixed and trained on an unknown training set $\mathcal{T}$.

2.2. Adaptation of a Fixed Classifier to New Priors

A probabilistic classifiers $f_{\mathcal{T}}(\mathbf{x}) \approx p_{\mathcal{T}}(Y|\mathbf{x})$ is simply adapted [5, 18, 20] to new a-priori probabilities $p_{\mathcal{E}}(Y)$ following the Bayes theorem:

$$p_{\mathcal{T}}(\mathbf{x}|Y) = p_{\mathcal{E}}(\mathbf{x}|Y) = \frac{p_{\mathcal{T}}(Y|\mathbf{x})p_{\mathcal{T}}(\mathbf{x})}{p_{\mathcal{T}}(Y)} = \frac{p_{\mathcal{E}}(Y|\mathbf{x})p_{\mathcal{E}}(\mathbf{x})}{p_{\mathcal{E}}(Y)} \tag{1}$$

The new predictive prior $p_{\mathcal{E}}(Y|\mathbf{x})$ is then:

$$p_{\mathcal{E}}(Y|\mathbf{x}) = p_{\mathcal{T}}(Y|\mathbf{x})\frac{p_{\mathcal{E}}(Y)p_{\mathcal{T}}(\mathbf{x})}{p_{\mathcal{T}}(Y)p_{\mathcal{E}}(\mathbf{x})} \propto p_{\mathcal{T}}(Y|\mathbf{x})\frac{p_{\mathcal{E}}(Y)}{p_{\mathcal{T}}(Y)} \tag{2}$$

We can proceed with estimating $p_{\mathcal{E}}(Y)$ with one of the methods from Sections 2.3 and 2.4. Alternatively, we can directly estimate the ratio of the priors $w(Y) = p_{\mathcal{E}}(Y)/p_{\mathcal{T}}(Y)$ as in Section 2.5.

2.3. Estimation of New Priors Based on Confusion Matrices

A standard procedure [16, 18] for prior estimation is based on a $K \times K$ confusion matrix (CM) in the format $\mathbf{C}_{d|y}$, where the value in the k-th column and i-th row is

the probability $p(D = i|Y = k)$ of classifier $\mathbf{f}$ deciding for class i when the true class is k. Assuming that the density $p_{\mathcal{T}}(X|Y) = p_{\mathcal{E}}(X|Y)$ remains unchanged, the confusion matrix $\mathbf{C}_{d|y}$ of a classifier does not change with prior shift [14]. Marginalizing over the joint density $p(D, Y)$:

$$\begin{aligned} p(D = i) &= \sum_{k=1}^{K} p(D = i|Y = k)p(Y = k) \\ p(D) &= \mathbf{C}_{d|y}p(Y) \end{aligned} \tag{3}$$

McLachnan [16] and Saerens et al. [18] simply compute the new priors $p_{\mathcal{E}}(Y)$ from Equation (3):

$$\hat{p}_{\mathcal{E}}(Y) = \hat{\mathbf{C}}_{d|y}^{-1}\hat{p}_{\mathcal{E}}(D), \tag{4}$$

using an estimate of $\mathbf{C}_{d|y}$ computed on a validation set and an estimate of $p(D)$ computed by counting the classifier decisions on the test set.

Let us also consider a *soft confusion matrix*[1] (SCM) $\mathbf{C}_{d|y}^{\text{soft}}$ estimated from the classifier's soft predictions $\mathbf{f}$ as

$$\hat{\mathbf{c}}_{:,k}^{\text{soft}} = \frac{1}{N_k} \sum_{\mathbf{x}_i : y_i = k} \mathbf{f}(\mathbf{x}_i), \tag{5}$$

where $\hat{\mathbf{c}}_{:,k}^{\text{soft}}$ denotes the k-th column of SCM. The probability $p_{\mathcal{E}}^{\text{soft}}(D)$ can be estimated by averaging predictions $f(\mathbf{x})$ over the test set. The new priors are then computed similarly to Equation (4).

2.4. Estimation of New Priors Based on Posterior Predictions

2.4.1 Maximum Likelihood and EM Algorithm

Saerens et al. [18] suggested to estimate priors $p_{\mathcal{E}}(Y)$ on the evaluation set $\mathcal{E} = \{\mathbf{x}_i\}_{i=1}^{M}$ by maximizing the likelihood:

$$L(\mathcal{E}) = \prod_{i=1}^{M} p_{\mathcal{E}}(\mathbf{x}_i) = \prod_{i=1}^{M} \sum_{k=1}^{K} p_{\mathcal{E}}(\mathbf{x}_i|Y = k)p_{\mathcal{E}}(Y = k) \tag{6}$$

They proposed an EM algorithm which iteratively recomputes the prior estimates.

Du Plessis and Sugiyama [5] showed that the EM algorithm can be derived from minimization of KL divergence between $p_{\mathcal{E}}(\mathbf{x})$ and its approximation. This leads to maximization of log-likelihood $l(\mathcal{E}) = \log L(\mathcal{E})$:

$$\begin{aligned} \hat{\mathbf{P}}^* = \arg\max_{\hat{\mathbf{P}}} \frac{1}{M} \sum_{i=1}^{M} \log \sum_{k=1}^{K} \hat{P}_k \frac{p_{\mathcal{T}}(Y = k|\mathbf{x}_i)}{p_{\mathcal{T}}(Y = k)} \\ s.t. \sum_{k=1}^{K} \hat{P}_k = 1; \quad \forall k : \hat{P}_k \geq 0, \end{aligned} \tag{7}$$

where $\hat{P}_k = \hat{p}_{\mathcal{E}}(Y = k)$.

[1] Following the terminology of Lipton et al. [14].

Sulc and Matas [20] experimented with maximizing the log-likelihood function by projected gradient ascent:

$$\hat{P}_k^{s+1} = \pi\left(\hat{P}_k^s + \frac{\partial l(\mathcal{E})}{\partial \hat{P}_k}\right) \tag{8}$$

where $\pi(\cdot)$ denotes projection onto the probability simplex and

$$\frac{\partial l(\mathcal{E})}{\partial \hat{P}_k} = \sum_{i=1}^{M} \frac{\frac{p_{\mathcal{T}}(Y=k|\mathbf{x}_i)}{p_{\mathcal{T}}(Y=k)}}{\sum_{j=1}^{K} \hat{P}_k \frac{p_{\mathcal{T}}(Y=j|\mathbf{x}_i')}{p_{\mathcal{T}}(Y=j)}}. \tag{9}$$

The experimental results showed that the EM algorithm converged faster than gradient ascent, while achieving similar results.

2.4.2 Maximum Aposteriori Estimation

Sulc and Matas [20] proposed a maximum a-posteriori estimation:

$$\begin{aligned} \hat{\mathbf{P}}^* &= \arg\max_{\hat{\mathbf{P}}} p(\hat{\mathbf{P}}|\mathcal{E}) = \arg\max_{\hat{\mathbf{P}}} p(\hat{\mathbf{P}})p(\mathcal{E}|\hat{\mathbf{P}}) \\ &= \arg\max_{\hat{\mathbf{P}}} \log p(\hat{\mathbf{P}}) + \underbrace{\log p(\mathcal{E}|\hat{\mathbf{P}})}_{l(\mathcal{E})} \end{aligned} \tag{10}$$

where the distribution $p(\hat{\mathbf{P}})$ is a hyper-prior representing some additional knowledge about class distribution. Specifically, they used a symmetric Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$.

The solution to maximum a-posteriori can be found by projected gradient ascent, adding the derivative of $\log \text{Dir}(\boldsymbol{\alpha})$ into the $\pi(\cdot)$ function: in Equation (8):

$$\frac{\partial \log p(\mathbf{P})}{\partial P_k} = \frac{\partial \log \text{Dir}(\boldsymbol{\alpha})}{\partial P_k} = \frac{\alpha - 1}{P_k} \tag{11}$$

2.5. Estimation of Prior Ratio Based on Confusion Matrices

Lipton et al. [14] estimate the prior ratio $w(Y) = \frac{p_{\mathcal{E}}(Y)}{p_{\mathcal{T}}(Y)}$ using a confusion matrix in the format $\mathbf{C}_{d,y}$, i.e. with joint probability $p(D = i, Y = k)$, unlike the conditional probability used in Section 2.3. Since $p_{\mathcal{T}}(D|Y) = p_{\mathcal{E}}(D|Y)$:

$$\begin{aligned} p_{\mathcal{E}}(D = i) &= \sum_{k=1}^{K} p_{\mathcal{T}}(D = i|Y = k)p_{\mathcal{E}}(Y = k) \\ &= \sum_{k=1}^{K} p_{\mathcal{T}}(D = i, Y = k) \underbrace{\frac{p_{\mathcal{E}}(Y = k)}{p_{\mathcal{T}}(Y = k)}}_{w(Y=k)} \\ p_{\mathcal{E}}(D) = \mathbf{C}_{d,y}w(Y) &\implies \hat{w}(Y) = \hat{\mathbf{C}}_{d,y}^{-1}\hat{p}_{\mathcal{E}}(D) \end{aligned} \tag{12}$$

This estimation is called Black Box Shift Estimation (BBSE). A variant using a soft confusion matrix $\mathbf{C}_{d,y}^{\text{soft}}$ is denoted BBSE-S.

Note that even the estimation of prior ratio may suffer from inconsistent estimates of the confusion matrix and $p(D)$, as demonstrated in the Supplementary Material.

2.6. Classifier Calibration

In the aforementioned methods, we often treated the classifier outputs $f_{\mathcal{T}}(\mathbf{x})$ as estimates of posterior probability $p_{\mathcal{T}}(Y|(x))$. In practice, outputs of common probabilistic classifiers, such as Convolutional Neural Networks, tend to provide over-confident predictions due to over-fitting to the training set. Guo et al. [9] study confidence calibration in the context of neural networks, and compare several models for classifier calibration, of which a simple temperature scaling (TS) procedure performs the best in terms of the calibration error. With temperature scaling, the softmax logits $z(\mathbf{x})$ are divided by the temperature T:

$$p^{\text{TS}}(y=i|\mathbf{x}) = \frac{\exp\left(z_i(\mathbf{x})/T\right)}{\sum_j \exp\left(z_i(\mathbf{x})/T\right)} \tag{13}$$

While lowering the calibration error, Alexandri et al. [1] show that temperature scaling is not a suitable calibration for adaptation to prior shift, possibly because of large systematic biases in the calibrated probabilities. They propose Bias-Corrected Temperature Scaling (BCTS), adding a class-specific bias term:

$$p^{\text{BCTS}}(y=i|\mathbf{x}) = \frac{\exp(z_i(\mathbf{x})/T + b_i)}{\sum_j \exp(z_j(\mathbf{x})/T + b_j)} \tag{14}$$

Alexandri et al. [1] show that such bias-corrected classifier calibration improves prior-adaptation with the EM-algorithm [18] from Section 2.4.1, outperforming both BBSL [14] and RLLS [2].

2.7. Training Data Sampling Strategies for Prior Shift Adaptation

A possible alternative to adapting the predictions $f(\mathbf{x}) \approx p_{\mathcal{T}}(Y|\mathbf{x})$ following Equation (2) is to instead train a new classifier $f_{\mathcal{E}}(\mathbf{x}) \approx p_{\mathcal{E}}(Y|\mathbf{x})$ by changing the sampling strategy from the training set according to $p_{\mathcal{E}}(Y)$. A similar approach was used e.g. in the winning submission of iNaturalist 2017 [4], where the training data was highly imbalanced, while the validation and test data were rather balanced in terms of class priors. The classifier in [4] was first trained on the full training set and then fine-tuned on a balanced subset of the training set. Unlike [4], we propose to use all training examples, but sample the training data following $p_{\mathcal{E}}(Y)$.

Prior adaptation following Equation (2) will be compared to re-training the network with adapted sampling in the experiments in Section 4.1. The practical disadvantage of the later approach is clear: the necessity to re-train the classifier with every new prior distribution.

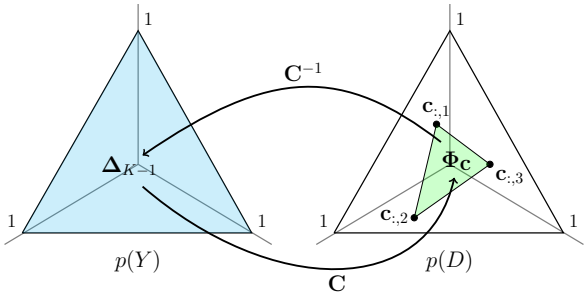

Figure 1: The convex set $\Phi_{\mathbf{C}} \subset \Delta_{K-1}$ of all possible values of $p(D)$ for a classifier with confusion matrix $\mathbf{C}$.

3. Proposed Methods

3.1. Maximum Likelihood Estimate Based on Confusion Matrices

As observed in the literature [6, 16, 21], Equation (4) can result in a vector outside of the Δ_{K-1} simplex, i.e. the estimate can contain negative values. We observe this phenomenon is common in practice.

Following Equation (3), the probability of classifier decisions $p(D)$ is a convex combination of columns in $\mathbf{C}_{d|y}$ as $p(Y) \in \Delta_{K-1}$. Since the columns of the confusion matrix are probability vectors, they define a convex set $\Phi_{\mathbf{C}}$ of feasible values $p(D)$ within the probability simplex Δ_{K-1}. In other words, a classifier with confusion matrix $\mathbf{C}$ will result in decisions from $p(D) \in \Phi_{\mathbf{C}}$. The class distribution $p(Y)$ determines the value of $p(D)$ within $\Phi_{\mathbf{C}}$. See Figure 1 for illustration. For the true distribution $p(D)$ and confusion matrix $\mathbf{C}_{d|y}$, Equation (3) holds. The problem occurs when we work with estimates of the true distribution $\hat{p}(D)$ and confusion matrix $\hat{\mathbf{C}}_{d|y}$. If the estimates computed from a limited sample are not consistent, there may be no prior probability $\hat{p}(Y)$ satisfying Equation (3): For example, having $_{d|y} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$, $\hat{p}(D) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, the unique solution to Equation (3) is $\hat{p}(Y) = \begin{bmatrix} \frac{4}{3} \\ -\frac{1}{3} \end{bmatrix}$.

We propose a novel procedure for prior estimation based on maximizing the likelihood of $p_{\mathcal{E}}(D)$, which handles inconsistent estimates of $\hat{p}(D)$ and confusion matrix $\hat{\mathbf{C}}_{d|y}$, as can work even with singular matrices $\hat{\mathbf{C}}_{d|y}$, as it does not use matrix inversion.

Let $\mathbf{n} = (n_1, \ldots, n_K)$ be the numbers of classifier's decisions for class $1, \ldots, K$ on test set $\mathcal{E}$ and let us denote $\mathbf{Q} = (q_1, \ldots, q_k) \coloneqq p_{\mathcal{E}}(D)$ the probabilities of classifier decisions on the test distribution $p_{\mathcal{E}}(X, Y)$. Assuming the independence of classifier decisions on the test set $\mathcal{E}$, the likelihood of $\mathbf{Q}$ follows by the multinomial distribution:

$$L(\mathbf{Q}) = p(\mathbf{n}|\mathbf{Q}) = \frac{(n_1 + \ldots + n_K)!}{n_1! \cdot \ldots \cdot n_K!} \cdot q_1^{n_1} \cdot \ldots \cdot q_K^{n_K} \tag{15}$$

Substituting Equation (3) into the likelihood function $L(\mathbf{Q})$, we can express the likelihood function of class priors $\mathbf{P}$:

$$L(\mathbf{P}) = p(\mathbf{n}|\mathbf{P}) = \frac{(n_1 + \ldots + n_K)!}{n_1! \cdot \ldots \cdot n_K!} \prod_{k=1}^{K} (\mathbf{c}_{k,:} \cdot \mathbf{P})^{n_k}, \tag{16}$$

where $\mathbf{c}_{k,:}$ is the k-th row of $\mathbf{C}_{d|y}$.

The log-likelihood is:

$$\ell(\mathbf{P}) = \log p(n|\mathbf{P}) = \sum_{k=1}^{K} n_k \log(\mathbf{c}_{k,:} \cdot \mathbf{P}) + \theta_{\mathbf{n}}, \tag{17}$$

where $\theta_{\mathbf{n}}$ is constant for a fixed $\mathbf{n}$.

We estimate the new class priors by maximizing the log-likelihood from Equation (17):

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}} \ell(\mathbf{P}) = \arg\max_{\mathbf{P}} \sum_{k=1}^{K} n_k \log \mathbf{c}_{k,:}\mathbf{P} \tag{18a}$$

$$\text{s.t.:} \quad \sum_{k=1}^{K} P_k = 1; \quad \forall k : P_k \geq 0 \tag{18b}$$

The convex objective can be iteratively maximized using projected gradient ascent:

$$\hat{\mathbf{P}}^{s+1} = \pi\left(\hat{\mathbf{P}}^{s} + \nabla\ell(\mathbf{P}^{s})\right) \tag{19}$$

where $\pi(\cdot)$ denotes projection onto the probability simplex [22] and the gradient is computed as:

$$\nabla\ell(\mathbf{P}) = \sum_{k=1}^{K} \frac{n_k}{\mathbf{c}_{k,:} \cdot \mathbf{P}} \mathbf{c}_{k,:} \tag{20}$$

3.2. Maximum A Posteriori Estimate Based on Confusion Matrices

Additional assumptions on the distribution $\mathbf{P}$ can be formulated as a hyper-prior $p(\mathbf{P})$. We can then extend the proposed procedure from Section 3.1 to formulate maximum a-posteriori (MAP) estimation:

$$\begin{aligned} \hat{\mathbf{P}}_{\mathbf{MAP}} &= \arg\max_{\mathbf{P}} p(\mathbf{P}|\mathbf{n}) = \arg\max_{\mathbf{P}} p(\mathbf{P})p(\mathbf{n}|\mathbf{P}) \\ &= \arg\max_{\mathbf{P}} \log p(\mathbf{P}) + \arg\max_{\mathbf{P}} \log p(\mathbf{n}|\mathbf{P}) \\ \text{s.t.:} \quad & \forall k : P_k \geq 0; \quad \sum_{k=1}^{K} P_k = 1 \end{aligned} \tag{21}$$

where $p(\mathbf{P})$ denotes a hyper-prior on $\mathbf{P}$ and $\log p(\mathbf{n}|\mathbf{P})$ is log-likelihood given by Equation (17).

Following [20] we use a symmetric Dirichlet hyper-prior $\mathrm{Dir}(\alpha)$, favouring dense distributions $\mathbf{P}$ with $\alpha > 1$, and a sparse distribution for $0 < \alpha < 1$.

The solution to maximum a-posteriori can be found by projected gradient ascent, adding the hyper-prior derivative from Equation (11) into the $\pi(\cdot)$ function in Equation (19).

4. Experiments

In this section, we compare the existing and proposed methods for prior shift adaptation on existing long-tailed versions of standard image classification datasets: the CIFAR100-LT [3], Places365-LT [15] and ImageNet-LT [15]. Unlike Cao et al. [3], our experiments require a validation set. Therefore, our training set, denoted as CIFAR100-LT, is smaller than of the original CIFAR100-LT, keeping 50 samples from each class for the validation set. Using the same script as Cao et al. [3] to sample the training set, the resulting imbalance ratio of 112.5 slightly differs from the original ratio of 100. For Places365-LT and ImageNet-LT, we use the same training and validation splits as Liu et al. [15]. Networks trained on these long-tailed datasets are then evaluated on uniformly distributed test sets (UNI). We also provide experiments in the other direction, denoted as UNI→LT, where networks trained on the full CIFAR100 and Places365 datasets are evaluated on test sets subsampled from the full test sets following the prior distributions of CIFAR100-LT and Places365-LT. Additional experiments on subsets of CIFAR100 and Places365 with hand-picked class distributions are in the suppl. material.

To evaluate the methods on practical tasks with prior shift, we experiment with fine-grained plant classification on the PlantCLEF data [7, 8] and with learning to classify ImageNet [17] classes from a long-tailed noisy training dataset downloaded from the web, Webvision 1.0 [12].

In the experiments with ImageNet and Webvision, we trained a ResNet-18 [10] classifier with the standard input size 224x224 from scratch. In the experiments on Places365 and PlantCLEF, we finetuned ResNet-18 from an ImageNet-pretrained checkpoint. In the CIFAR100 experiments, we used a ResNet-32 adjusted to input size 32x32.

4.1. New Priors Are Known: Adapt or Re-Train?

Let us first examine the case when new class priors $p_{\mathcal{E}}$ are known, and compare:

1. Adapting the predictions of a previously trained classifier $f_{\mathcal{T}}(\mathbf{x}) \approx p_{\mathcal{T}}(Y|\mathbf{x})$, following Eq. (2).
2. Training a classifier $f_{\mathcal{E}}(\mathbf{x})$ with a sampler following the known new class priors $p_{\mathcal{E}}$.

When adapting predictions of classifier $f_{\mathcal{T}}$ following Eq. (2), the trained priors can be determined either as a proportion of class labels in the training set, $\hat{p}^{N}_{\mathcal{T}}(Y = k) = \frac{N_k}{N}$, or as the average of predictions $f(\mathbf{x})$ on the training set, $\hat{p}^{f}_{\mathcal{T}}(Y) = \frac{1}{N}\sum_{i=1}^{n} f(\mathbf{x_i})$.

The training- and adaptation- strategies are experimentally compared on the CIFAR100-LT*, Places365-LT and ImageNet-LT datasets in Table 1. The results show that adaptation of the classifier performs better than re-training the classifier with weighted sampling following $p_{\mathcal{E}}$. In most cases, the best results are achieved when trained priors are

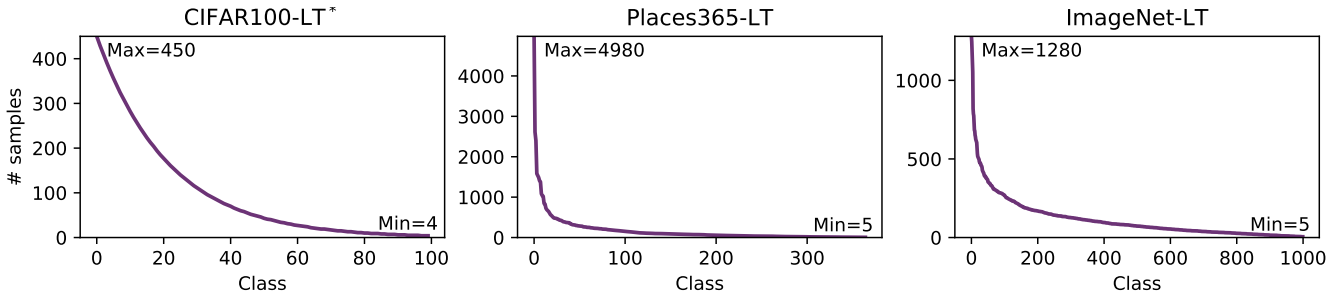

Figure 2: Long-tailed class distributions used in the CIFAR100-LT* [3], Places365-LT [15] and ImageNet-LT [15] datasets. Note that our CIFAR-100-LT* slightly differs from the original CIFAR-100-LT [3], which did not have a validation set.

Dataset		BCTS [1] calibrated	Standard training sampler NA	Standard training sampler $\frac{p_{\mathcal{E}}}{\hat{p}^N_{\mathcal{T}}}$	Standard training sampler $\frac{p_{\mathcal{E}}}{\hat{p}^f_{\mathcal{T}}}$	Sampler follows $p_{\mathcal{E}}$ NA
CIFAR100*	LT→UNI	✗	$31.66^{\pm 1.27}$	$33.99^{\pm 1.41}$	**$34.06^{\pm 1.35}$**	22.44
	LT→UNI	✓	$31.71^{\pm 1.29}$	$30.41^{\pm 1.57}$	**$34.54^{\pm 1.32}$**	22.44
	UNI→LT	✗	$63.83^{\pm 0.82}$	**$69.14^{\pm 0.61}$**	$69.13^{\pm 0.58}$	67.34
	UNI→LT	✓	$63.83^{\pm 0.82}$	$70.63^{\pm 0.75}$	**$70.65^{\pm 0.75}$**	67.34
Places365	LT→UNI	✗	$25.14^{\pm 0.14}$	$28.03^{\pm 6.09}$	**$32.99^{\pm 0.46}$**	24.98
	LT→UNI	✓	$25.16^{\pm 0.14}$	$27.36^{\pm 6.18}$	**$33.51^{\pm 0.25}$**	25.00
ImageNet	LT→UNI	✗	$34.30^{\pm 0.19}$	**$37.36^{\pm 0.07}$**	$37.34^{\pm 0.15}$	30.01
	LT→UNI	✓	$34.31^{\pm 0.19}$	$36.07^{\pm 0.30}$	**$37.45^{\pm 0.19}$**	30.01

Table 1: *"Adapt or Re-Train?"* Accuracy (± std. dev.) of classifiers adapted to new known priors $p_{\mathcal{E}}(Y)$ with different estimates of trained priors ($\hat{p}^N_{\mathcal{T}}$,$\hat{p}^f_{\mathcal{T}}$), compared to training a classifier with a sampling strategy following $p_{\mathcal{E}}(Y)$. NA denotes no adaptation of predictions. Results of classifier adaptation on CIFAR are averaged from 10 experiments, on Places365 and ImagNet from 5 experiments respectively. Re-training the classifier with a sampler following $p_{\mathcal{E}}(Y)$ was only experimented once for each dataset.

estimated from the predictions on the training set. We will thus estimate the trained priors by $\hat{p}_{\mathcal{T}}(Y) = \hat{p}^f_{\mathcal{T}}(Y)$.

4.2. Prior Shift Estimation

4.2.1 Improving Estimates from Confusion Matrices

Table 2 compares accuracy after adaptation with new prior estimate based on confusion matrix (CM) inversion [18] and our proposed method from Section 3.1 (CM$^{\text{L}}$) . The proposed method handles inconsistent estimates $\hat{p}(D)$ and $\hat{\mathbf{C}}_{d|y}$ and consistently improves the results both using the confusion matrix (CM$^{\text{L}}$) and the soft confusion matrix (SCM$^{\text{L}}$). In all cases, the proposed SCM$^{\text{L}}$ method using soft confusion matrix achieves the best results.

4.2.2 Methods for MLE and MAP Prior Estimation

Existing methods for maximum likelihood and maximum a-posteriori prior estimation are compared against the methods proposed in Sections 3.1 and 3.2 respectively in Table 3. Note that the methods maximize a different likelihood function: The EM algorithm of Saerens et al. [18] maximizes the likelihood of observed classifier outputs $f(\mathbf{x}_i)$, while the proposed methods based on confusion matrix (CM^{L}) and soft confusion matrix (SCM^{L}) maximize the likelihood of classifiers decisions $\arg\max_k f(\mathbf{x}_i)$. The same difference in likelihood functions holds for the MAP approach of Sulc and Matas [20] and MAP estimate proposed in Section 3.2, but we use the same hyper-prior on $p_{\mathcal{E}}(Y)$ for all methods: $Dir(\alpha = 3)$.

From the maximum likelihood estimators, the proposed SCM$^{\text{L}}$ achieves the best results in most cases, with the exception of Places365 "UNI→LT", where the EM algorithm performed slightly better. Similarly, the Maximum A-Posteriori version of the proposed method, SCM$^{\text{M}}$ performs better than the existing MAP estimate [20] in most cases. As expected, the MAP estimation improves upon MLE on the dense test distributions, favoured by the Dirichlet hyper-prior.

4.3. Prior Ratio Estimation

Having an estimate of the trained priors, Table 4 compares prior ratio estimation with BBSE, BBSE-S [14] and RLLS [2] against the best performing prior estimation methods, CM$^{\text{L}}$ and SCM$^{\text{L}}$. The results indicate that it is

Dataset		BCTS [1] calibrated	NA	CM	CM^L	SCM	SCM^L	Oracle
CIFAR100*	LT→UNI	✗	$31.66^{\pm 1.27}$	$21.37^{\pm 2.68}$	$\mathbf{33.00^{\pm 1.56}}$	$26.64^{\pm 3.85}$	$\mathbf{33.47^{\pm 1.33}}$	$34.06^{\pm 1.35}$
	LT→UNI	✓	$31.67^{\pm 1.27}$	$19.11^{\pm 2.98}$	$\mathbf{32.41^{\pm 1.50}}$	$26.98^{\pm 3.76}$	$\mathbf{33.42^{\pm 1.51}}$	$34.40^{\pm 1.38}$
	UNI→LT	✗	$63.83^{\pm 0.82}$	$68.06^{\pm 0.92}$	$\mathbf{68.08^{\pm 0.75}}$	$68.10^{\pm 0.81}$	$\mathbf{68.24^{\pm 0.75}}$	$69.13^{\pm 0.58}$
	UNI→LT	✓	$63.83^{\pm 0.82}$	$69.08^{\pm 0.94}$	$\mathbf{69.10^{\pm 0.98}}$	$69.31^{\pm 0.94}$	$\mathbf{69.40^{\pm 0.77}}$	$70.65^{\pm 0.75}$
Places365	LT→UNI	✗	$25.14^{\pm 0.14}$	$17.45^{\pm 0.30}$	$\mathbf{27.77^{\pm 0.45}}$	$19.78^{\pm 2.21}$	$\mathbf{28.47^{\pm 0.14}}$	$32.99^{\pm 0.46}$
	LT→UNI	✓	$25.14^{\pm 0.14}$	$16.24^{\pm 1.39}$	$\mathbf{27.69^{\pm 0.51}}$	$18.88^{\pm 1.61}$	$\mathbf{27.83^{\pm 0.27}}$	$33.38^{\pm 0.31}$
	UNI→LT	✗	$58.17^{\pm 1.01}$	$81.16^{\pm 0.61}$	$\mathbf{81.64^{\pm 0.63}}$	$\mathbf{82.04^{\pm 0.15}}$	$82.04^{\pm 0.63}$	$88.14^{\pm 0.27}$
	UNI→LT	✓	$58.17^{\pm 1.01}$	$81.20^{\pm 0.61}$	$\mathbf{81.65^{\pm 0.61}}$	$82.04^{\pm 0.15}$	$\mathbf{82.07^{\pm 0.66}}$	$88.15^{\pm 0.30}$
ImageNet	LT→UNI	✗	$34.30^{\pm 0.19}$	$19.02^{\pm 0.26}$	$\mathbf{33.57^{\pm 0.33}}$	$23.94^{\pm 2.04}$	$\mathbf{35.91^{\pm 0.20}}$	$37.34^{\pm 0.15}$
	LT→UNI	✓	$34.30^{\pm 0.19}$	$17.28^{\pm 0.48}$	$\mathbf{32.34^{\pm 0.41}}$	$24.78^{\pm 3.06}$	$\mathbf{35.86^{\pm 0.17}}$	$37.39^{\pm 0.16}$

Table 2: *"Improve Estimates from Confusion Matrix."* Accuracy (± std. dev.) after adaptation with new prior estimate based on confusion matrix (CM) inversion [18] and our proposed method from Section 3.1 (CM^L). SCM denotes soft confusion matrix, NA denotes no adaptation, Oracle is adaptation with ground truth priors. Results on CIFAR are averaged from 10 experiments, results on Places and ImageNet are averaged from 5 experiments. Best results are displayed in bold.

Dataset		BCTS [1] calibrated	NA	MLE			MAP			Oracle
				EM	CM^L	SCM^L	MAP	CM^M	SCM^M	
CIFAR100*	LT→UNI	✗	$31.66^{\pm 1.27}$	$32.81^{\pm 1.41}$	$33.00^{\pm 1.56}$	$\mathbf{33.47^{\pm 1.33}}$	$32.73^{\pm 1.42}$	$33.49^{\pm 1.45}$	$\mathbf{33.50^{\pm 1.40}}$	$34.06^{\pm 1.35}$
	LT→UNI	✓	$31.67^{\pm 1.27}$	$29.43^{\pm 1.59}$	$32.41^{\pm 1.50}$	$\mathbf{33.42^{\pm 1.51}}$	$24.46^{\pm 12.43}$	$33.99^{\pm 1.43}$	$\mathbf{34.13^{\pm 1.53}}$	$34.40^{\pm 1.38}$
	UNI→LT	✗	$63.83^{\pm 0.82}$	$67.23^{\pm 0.88}$	$68.08^{\pm 0.75}$	$\mathbf{68.24^{\pm 0.75}}$	$66.72^{\pm 0.91}$	$\mathbf{67.01^{\pm 0.91}}$	$67.00^{\pm 0.87}$	$69.13^{\pm 0.58}$
	UNI→LT	✓	$63.83^{\pm 0.82}$	$69.17^{\pm 0.91}$	$69.10^{\pm 0.98}$	$\mathbf{69.40^{\pm 0.77}}$	$68.30^{\pm 0.77}$	$\mathbf{68.42^{\pm 0.84}}$	$68.38^{\pm 0.73}$	$70.65^{\pm 0.75}$
Places365	LT→UNI	✗	$25.14^{\pm 0.14}$	$28.02^{\pm 0.92}$	$27.77^{\pm 0.45}$	$\mathbf{28.47^{\pm 0.14}}$	$25.22^{\pm 0.13}$	$\mathbf{28.02^{\pm 0.24}}$	$27.68^{\pm 0.13}$	$32.99^{\pm 0.46}$
	LT→UNI	✓	$25.14^{\pm 0.14}$	$\mathbf{28.09^{\pm 1.32}}$	$27.69^{\pm 0.51}$	$27.83^{\pm 0.27}$	$\mathbf{28.57^{\pm 0.24}}$	$27.92^{\pm 0.24}$	$27.41^{\pm 0.15}$	$33.38^{\pm 0.31}$
	UNI→LT	✗	$58.17^{\pm 1.01}$	$\mathbf{82.63^{\pm 0.31}}$	$81.64^{\pm 0.63}$	$82.04^{\pm 0.63}$	$\mathbf{76.97^{\pm 0.45}}$	$76.13^{\pm 0.56}$	$73.27^{\pm 0.46}$	$88.14^{\pm 0.27}$
	UNI→LT	✓	$58.17^{\pm 1.01}$	$\mathbf{82.63^{\pm 0.26}}$	$81.65^{\pm 0.61}$	$82.07^{\pm 0.66}$	$\mathbf{77.00^{\pm 0.41}}$	$76.16^{\pm 0.54}$	$73.30^{\pm 0.46}$	$88.15^{\pm 0.30}$
ImageNet	LT→UNI	✗	$34.30^{\pm 0.19}$	$34.63^{\pm 0.29}$	$33.57^{\pm 0.33}$	$\mathbf{35.91^{\pm 0.20}}$	$34.64^{\pm 0.20}$	$36.41^{\pm 0.17}$	$\mathbf{36.57^{\pm 0.16}}$	$37.34^{\pm 0.15}$
	LT→UNI	✓	$34.30^{\pm 0.19}$	$27.26^{\pm 2.25}$	$32.34^{\pm 0.41}$	$\mathbf{35.86^{\pm 0.17}}$	$20.65^{\pm 18.77}$	$36.18^{\pm 0.12}$	$\mathbf{36.80^{\pm 0.14}}$	$37.39^{\pm 0.16}$

Table 3: *"How to estimate new priors?"* Accuracy (± std. dev.) after adaptation to new priors estimated with different Maximum Likelihood and Maximum A Posteriori estimates. NA denotes no adaptation, Oracle is adaptation with ground truth priors. Best MLE and MAP results are underlined for $f_{\mathcal{T}}$ and calibrated $f_{\mathcal{T}}$. Results on CIFAR are averaged from 10 experiments, results on Places and ImageNet are averaged from 5 experiments. Best results are displayed in bold.

better to estimate the new priors than to directly estimate the prior ratio with BBSE or BBSE-S.

4.4. Dependence on the Number of Test Samples

Figure 3 displays the accuracy on the uniformly distributed sets after adaptation of classifiers trained on the CIFAR100-LT* and Places365-LT datasets with different prior estimation methods, as a function of the number of test examples used for prior estimation. While the proposed SCM^L method achieves slightly higher accuracy with more samples, the EM algorithm works slightly better with low number of samples. With extremely low number of samples, prior estimation should be omitted.

4.5. Applying the Best Practice

The lessons learned were applied to two tasks with naturally imbalanced priors:

A classifier trained on the imbalanced Webvision 1.0 dataset achieved 57.12% accuracy on the ImageNet validation set. ImageNet has a uniform class distribution. We calibrated the classifier with BCTS [1], estimated the trained prior, and adapted the predictions by Equation (2). The classification accuracy improved to 58.22%.

A classifier of 10,000 plant species trained on the PlantCLEF 2017 dataset (EOL+test) achieved 37.95% accuracy on the PlantCLEF 2018 test set of 2072 images. Because the number of samples was very low, we used the EM algorithm [1, 18] on the calibrated predictions. Adapting to the estimated priors increased the accuracy to 41.35%.

5. Conclusions

This paper reviews and compares existing methods for adaptation to prior shift and proposes a novel method to

Dataset		BCTS [1] calibrated	NA	SCM$^{\text{L}}$	RLLS	BBSE	BBSE-S	Oracle
CIFAR100*	LT→UNI	✗	$31.66^{\pm 1.27}$	**$33.47^{\pm 1.33}$**	$32.75^{\pm 1.40}$	$31.28^{\pm 1.58}$	$31.92^{\pm 1.62}$	$34.06^{\pm 1.35}$
	LT→UNI	✓	$31.67^{\pm 1.27}$	**$33.42^{\pm 1.51}$**	$32.62^{\pm 1.46}$	$26.47^{\pm 1.87}$	$29.06^{\pm 2.65}$	$34.40^{\pm 1.38}$
	UNI→LT	✗	$63.83^{\pm 0.82}$	**$68.24^{\pm 0.75}$**	$68.02^{\pm 0.77}$	$67.95^{\pm 0.96}$	$68.12^{\pm 0.90}$	$69.13^{\pm 0.58}$
	UNI→LT	✓	$63.83^{\pm 0.82}$	$69.40^{\pm 0.77}$	$69.05^{\pm 0.97}$	$69.30^{\pm 0.99}$	**$69.51^{\pm 0.98}$**	$70.65^{\pm 0.75}$
Places365	LT→UNI	✗	$25.14^{\pm 0.14}$	**$28.47^{\pm 0.14}$**	$26.94^{\pm 0.41}$	$24.79^{\pm 0.74}$	$25.55^{\pm 0.69}$	$32.99^{\pm 0.46}$
	LT→UNI	✓	$25.14^{\pm 0.14}$	**$27.83^{\pm 0.27}$**	$27.03^{\pm 0.40}$	$23.12^{\pm 0.79}$	$23.68^{\pm 0.75}$	$33.38^{\pm 0.31}$
	UNI→LT	✗	$58.17^{\pm 1.01}$	**$82.04^{\pm 0.63}$**	$82.04^{\pm 0.69}$	$80.66^{\pm 0.57}$	$81.69^{\pm 0.20}$	$88.14^{\pm 0.27}$
	UNI→LT	✓	$58.17^{\pm 1.01}$	**$82.07^{\pm 0.66}$**	$82.04^{\pm 0.69}$	$80.71^{\pm 0.56}$	$81.69^{\pm 0.20}$	$88.15^{\pm 0.30}$
ImageNet	LT→UNI	✗	$34.30^{\pm 0.19}$	**$35.91^{\pm 0.20}$**	$34.69^{\pm 0.14}$	$30.77^{\pm 0.31}$	$31.31^{\pm 0.90}$	$37.34^{\pm 0.15}$
	LT→UNI	✓	$34.30^{\pm 0.19}$	**$35.86^{\pm 0.17}$**	$34.31^{\pm 0.08}$	$26.89^{\pm 0.49}$	$28.05^{\pm 1.69}$	$37.39^{\pm 0.16}$

Table 4: "Estimate test priors or directly the prior ratio?" Accuracy (± std. dev.) after adaptation with the priors estimated by SCM$^{\text{L}}$ or with the prior ratio estimated by BBSE [14] and RLLS [2] (without re-training). Results on CIFAR are averaged from 10 experiments, results on Places and ImageNet are averaged from 5 experiments. Best results are displayed in bold.

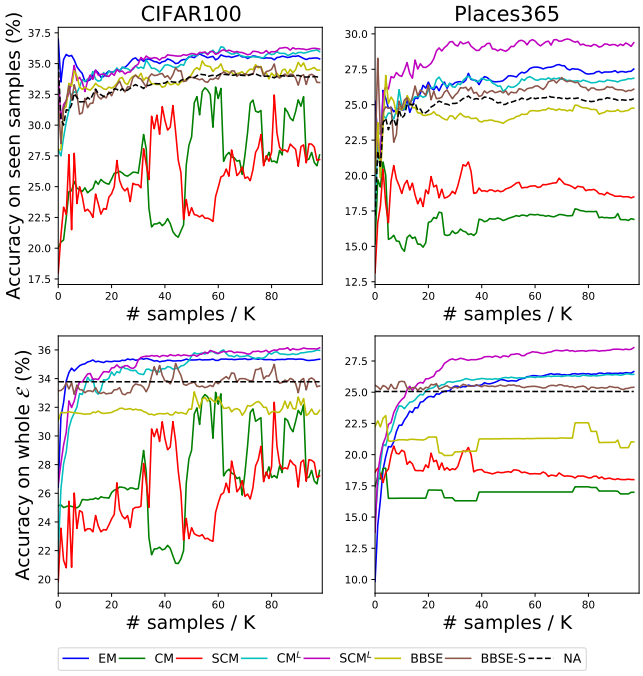

Figure 3: *"How many samples do I need?"* Accuracy after adapting CIFAR100-LT→UNI (left) and Places365-LT→UNI (right) using #samples for prior estimation.

deal with a known problem of existing methods based on confusion matrices [6, 18, 21], where inconsistent estimates of decision probabilities and confusion matrices can result in negative values in the estimated priors. The proposed method, (S)CM$^{\text{L}}$, deals with this problem by constrained maximization of the likelihood of classifier decisions on the new test set. It has further been extended into a Maximum A-Posteriori estimator (S)CM$^{\text{M}}$ by adding a hyper-prior on the new prior distribution.

Experimental analysis of the existing and proposed methods for prior shift adaptation suggests the following best practice:

- Adaptation of the original classifier typically performs better than re-training the classifier with sampling matching the shift, and is significantly computationally cheaper.
- The proposed method handles inconsistent estimates $\hat{p}(D)$ and $\hat{\mathbf{C}}_{d|y}$ and consistently improves the results both using the confusion matrix (CM$^{\text{L}}$) and the soft confusion matrix (SCM$^{\text{L}}$).
- From the compared maximum likelihood estimators, the proposed SCM$^{\text{L}}$ achieves the best results in most cases.
- The EM algorithm [18, 1] works better with a low number of samples. With extremely low number of samples, prior estimation should better be omitted at all.
- The proposed Maximum A-Posteriori approach, SCM$^{\text{M}}$, performs better than the existing MAP estimate [20].
- If trained priors can be estimated, it is better to estimate the test set priors than to directly estimate the prior ratio with BBSE or RLLS.
- Prior shift adaptation relies on a well-calibrated classifier, assumed in Eq. (2). In [1], BCTS improves prior shift adaptation of classifiers trained on uniform distribution, similarly to our UNI→LT experiments. With class-specific parameters, BCTS may overfit to errors on the validation set. For classifiers trained on LT datasets, we show BCTS is not a reliable calibration method, as it often decreases the final recognition accuracy.

Applying the best practice to two tasks with naturally imbalanced priors, learning from web-crawled images and plant classification, increased the accuracy by 1.1% and 3.4% respectively.

Acknowledgment This work was supported by Toyota Motor Europe, by CTU grant SGS20/171/OHK3/3T/13, and by OP VVV project CZ.02.1.01/0.0/0.0/16 019/0000765 Research Center for Informatics.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *ICML*, pages 222–232, 2020.

[2] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *ICLR*, 2019.

[3] Kaidi Cao et al. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

[4] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

[5] Marthinus Christoffel du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *CoRR*, abs/1206.4677, 2012.

[6] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, Oct 2008.

[7] Herve Goeau, Pierre Bonnet, and Alexis Joly. Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). 2017.

[8] Hervé Goëau, Pierre Bonnet, and Alexis Joly. Overview of expertlifeclef 2018: how far automated identification systems are from the best experts? 2018.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Jedrzej Kozerawski, Victor Fragoso, Nikolaos Karianakis, Gaurav Mittal, Matthew Turk, and Mei Chen. Blt: Balancing long-tailed datasets with adversarially-perturbed images. In *ACCV*, 2020.

[12] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[13] Tsung-Yi Lin et al. Focal loss for dense object detection. In *CVPR*, 2017.

[14] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors, 2018.

[15] Ziwei Liu et al. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[16] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc, Hoboken, NJ, USA, 1992-03-27.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[18] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.*, 14(1):21–41, Jan. 2002.

[19] Masashi Sugiyama et al. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2008.

[20] Milan Sulc and Jiri Matas. Improving cnn classifiers by estimating test-time priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[21] Slobodan Vucetic and Zoran Obradovic. Classification on data with biased class distribution. In *European Conference on Machine Learning*, pages 527–538. Springer, 2001.

[22] Weiran Wang and Miguel Á. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *ArXiv*, abs/1309.1541, 2013.

[23] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML*, 2004.

[24] Kun Zhang et al. Domain adaptation under target and conditional shift. In *ICML*, 2013.