On the Calibration of Nested Dichotomies for Large Multiclass Tasks

Tim Leathart, Eibe Frank, Bernhard Pfahringer and Geoffrey Holmes
Department of Computer Science, University of Waikato, New Zealand

Abstract

Nested dichotomies are used as a method of transforming a multiclass classification problem into a series of binary problems. A tree structure is induced that recursively splits the set of classes into subsets, and a binary classification model learns to discriminate between the two subsets of classes at each node. In this paper, we demonstrate that these nested dichotomies typically exhibit poor probability calibration, even when the base binary models are well calibrated. We also show that this problem is exacerbated when the binary models are poorly calibrated. We discuss the effectiveness of different calibration strategies and show that accuracy and log-loss can be significantly improved by calibrating both the internal base models and the full nested dichotomy structure, especially when the number of classes is high.

Introduction

As the amount of data collected online continues to grow, modern datasets utilised in machine learning are increasing in size. Not only do these datasets exhibit a large number of examples and features, but many also have a very high number of classes. It is not uncommon in some application areas to see datasets containing tens of thousands (Deng et al. 2009), or even millions of classes (Dekel and Shamir 2010; Agrawal et al. 2013).

An attractive option to handle datasets with such large label spaces is to induce a binary tree structure over the label space. At each split node k, the set of classes present, $\mathcal{C}_k$, is split into two disjoint subsets $\mathcal{C}_{k1}$ and $\mathcal{C}_{k2}$. Then, a binary classification model is trained to distinguish between these two subsets of classes. Many algorithms have been proposed that fit this general description, for example (Beygelzimer, Langford, and Ravikumar 2009; Bengio, Weston, and Grangier 2010; Choromanska and Langford 2015; Daumé et al. 2017). Usually, in these tree structures, a greedy inference approach is taken, i.e., test examples only take a single path from the root node to leaf nodes. This has the inherent drawback that a single mistake along the path to a leaf node results in an incorrect prediction.

In this paper, we consider methods with probabilistic classifiers at the internal nodes, called *nested dichotomies* in the literature (Frank and Kramer 2004). Utilising probabilistic

binary classifiers at the internal nodes to make routing decisions for test examples has several advantages over simply taking a hard 0/1 classification. For example, multiclass class probability estimates can be computed in a natural way by taking the product of binary probability estimates on the path from the root to the leaf node (Fox 1997). However, although hard classification decisions are avoided, small errors in the binary probability estimates can accumulate over this product, resulting in overall potentially poorer quality predictions. Datasets with more classes result in deeper trees, exacerbating this issue.

In this paper, we investigate approaches to reduce the impact of the accumulation of errors by utilising *probability calibration* techniques. Probability calibration is the task of transforming the probabilities output by a model to reflect their true empirical distribution; for the group of test examples that are predicted to belong to some class with probability 0.8, we expect about 80% of them to actually belong to that class if our model is well calibrated. Our main hypothesis is that the overall predictive performance of nested dichotomies can be improved by calibrating the individual binary models at internal nodes (referred to in this paper as *internal calibration*). However, we also observe that significant performance gains can be achieved by calibrating the entire nested dichotomy (referred to as *external calibration*), even if the internal models are well calibrated.

This paper is structured as follows. First, we briefly review nested dichotomies and probability calibration. We then discuss internal and external calibration, providing theoretical motivation and showing experimental results for each method. Finally, we conclude and discuss future research directions.

Nested Dichotomies

Nested dichotomies are used as a binary decomposition method for multiclass problems (Frank and Kramer 2004). In this paper, we only consider the case where a single nested dichotomy structure is built, although generally superior performance can be achieved by training an ensemble of nested dichotomies with different structures (Rodríguez, García-Osorio, and Maudes 2010). Ensembles of nested dichotomies have been shown to outperform binary decomposition methods like one-vs-all (Rifkin and Klautau 2004), one-vs-one (Friedman 1996) and error-correcting output
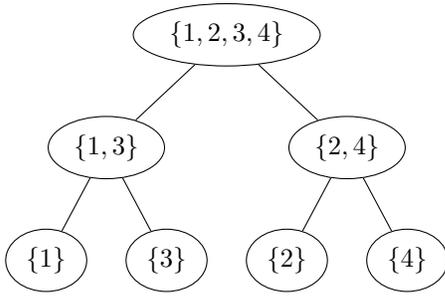
Figure 1: An example of nested a dichotomy for a four class problem.

codes (Dietterich and Bakiri 1995), on some classification problems (Frank and Kramer 2004).

The structure of a nested dichotomy can have a large impact on the predictive performance, training time and prediction time. To this end, several methods have been proposed for deciding the structure of nested dichotomies (Dong, Frank, and Kramer 2005; Leathart, Pfahringer, and Frank 2016; Melnikov and Hüllermeier 2018; Wever, Mohr, and Hüllermeier 2018; Leathart et al. 2018). In this paper, we focus on a simple method that randomly splits the set of classes into two at each internal node.

As previously stated, a useful feature of nested dichotomies is the ability to produce multiclass probability estimates $\hat{\mathbf{p}}_i$ for a test instance $(\mathbf{x}_i, y_i)$ from the product of binary estimates on the path $\mathcal{P}_c$ to the leaf node corresponding to class c:

$$
\begin{aligned}
\hat{\mathbf{p}}_i^{(c)} &= p(y_i = c | \mathbf{x}_i) \\
&= \prod_{k \in \mathcal{P}_c} \big(\mathbb{I}(c \in \mathcal{C}_{k1}) p(c \in \mathcal{C}_{k1} | \mathbf{x}_i, y_i \in \mathcal{C}_k) + \\
&\qquad \mathbb{I}(c \in \mathcal{C}_{k2}) p(c \in \mathcal{C}_{k2} | \mathbf{x}_i, y_i \in \mathcal{C}_k) \big).
\end{aligned}
$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{C}_k$ is the set of classes present at node k and $\mathcal{C}_{k1}, \mathcal{C}_{k2} \subset \mathcal{C}_k$ are the sets of classes present at the left and right child of node k, respectively. If desired, one can still perform greedy inference by taking the most promising branch at each split point, but this is not guaranteed to find the best solution (Beygelzimer, Langford, and Ravikumar 2009). Having binary class probability estimates facilitates efficient tree search techniques (Kumar et al. 2013; Mena et al. 2015; Dembczyński et al. 2016) for better inference, as well as top-k prediction.

Probability Calibration

Probability calibration is the task of transforming the outputs of a classifier to accurate probabilities. It is useful in a range of settings, such as cost-sensitive classification and scenarios where the outputs of a model are used as inputs for another. It is also important in real world decision making systems to know when a prediction from a model is likely to be incorrect. For example, models used in an automated healthcare system should indicate when they are unsure of a prediction so a doctor can take control.

Some models, like logistic regression, tend to be well calibrated out-of-the-box, while other models like naïve Bayes and boosted decision trees usually exhibit poor calibration, despite high classification accuracy (Niculescu-Mizil and Caruana 2005). Some other models such as support vector machines cannot output probabilities at all, so calibration can be applied to produce a probability estimate. Calibration is typically applied as a post-processing step—a calibration model is trained to transform the output score from a model into a well calibrated probability. The calibration model should be trained on a held-out dataset that was not used for training of the base model to avoid overfitting.

Calibration Methods

The most commonly used calibration methods in practice are Platt scaling (Platt 1999) and isotonic regression (Zadrozny and Elkan 2001). Both of these methods are only directly applicable to binary problems, but standard multiclass transformation techniques can be used to apply them to multiclass problems (Zadrozny and Elkan 2002).

Platt Scaling is a technique that fits a sigmoid curve

$$
\boldsymbol{\sigma}(z_i) = \frac{1}{1 + e^{\alpha z_i + \beta}}
$$

from the output of a binary classifier z_i to the true labels. The parameters α and β are fitted using logistic regression. Platt scaling was originally proposed for scaling the output of SVMs, but has been shown to be an effective calibration technique for a range of models (Niculescu-Mizil and Caruana 2005). Usually, Platt scaling is applied to the log-odds (sometimes called logits) of the positive class, rather than the probability.

Matrix and Vector Scaling are simple extensions of Platt scaling for multiclass problems (Guo et al. 2017). In matrix scaling, instead of single parameters α and β, a matrix $\mathbf{W}$ and bias vector $\mathbf{b}$ is learned:

$$
\boldsymbol{\sigma}(\mathbf{z}_i) = \frac{1}{1 + e^{\mathbf{W}\mathbf{z}_i + \mathbf{b}}}
$$

where $\mathbf{z}_i$ is the vector of the log-odds of each class for instance i. Matrix scaling is equivalent to a standard multiple logistic regression model applied to the log-odds. It is expensive for datasets with many classes, as the weight matrix $\mathbf{W}$ grows quadratically with the number of classes. Vector scaling is proposed to overcome this. It is a variant where $\mathbf{W}$ is restricted to be a diagonal matrix to achieve scaling that is linear in the number of classes.

Isotonic Regression is a non-parametric technique for probability calibration (Zadrozny and Elkan 2001). It fits a piecewise constant function to minimise the mean squared error between the estimated class probabilities and the true labels. Isotonic regression is a more general method than Platt scaling because no assumptions are made about the function used to map classifier outputs, other than that the function is non-decreasing (isotonicity). Isotonic regression has been found to work well as a calibration model, but the flexibility of the fitted function means it can overfit on small samples.
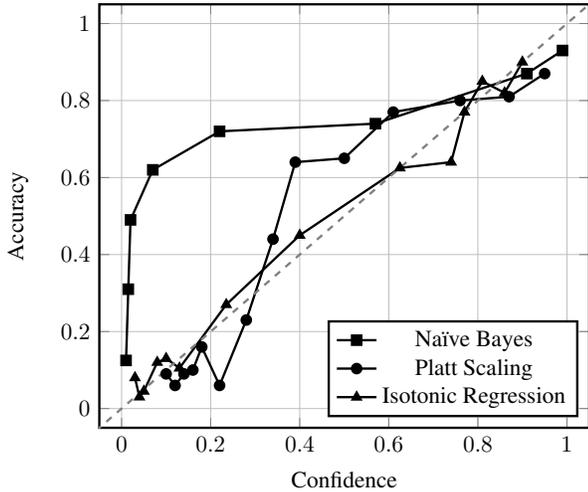
Figure 2: Reliability diagram for of the `credit` dataset from the UCI repository. Curves are shown for the predictions made by naïve Bayes, as well as calibrated probabilities through Platt scaling and isotonic regression.

Other related work. In this paper, efficiency is a concern as there are many models to be calibrated. For this reason, we opt for the simpler calibration methods mentioned above in our experiments. However, there are several more expressive (and expensive) calibration methods in the literature. Jiang et al. (2011) and Zhong and Kwok (2013) propose methods for creating a smooth spline from the piecewise constant function produced in isotonic regression. Naeini, Cooper, and Hauskrecht (2015) propose a method for performing Bayesian averaging over all possible binning schemes—schemes that split the probability space into several bins and calibrated probability value is established per bin. Leathart et al. (2017) split the feature space into regions using a decision tree and build a localised calibration model in each region.

Measuring Miscalibration

The level of probability calibration that a model exhibits is commonly measured by the negative log-likelihood (NLL):

$$\text{NLL} = -\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \log \hat{\mathbf{p}}_i$$

where n is the number of examples, $\mathbf{y}_i$ is the one-hot true label for an instance and $\hat{\mathbf{p}}_i$ is the estimated probability distribution. NLL heavily penalises probability estimates that are far from the true label. For this reason, models which optimise NLL in training tend to be well calibrated, although interestingly it has been shown recently that the kinds of architectures used in modern neural networks can also produce poorly calibrated models (Guo et al. 2017). NLL is also commonly used as a general measure of model accuracy.

Probability calibration for binary classification tasks can be visualised through reliability diagrams (DeGroot and Fienberg 1983). In reliability diagrams, the probability range

$[0, 1]$ is discretised into K bins $B_1, \ldots, B_k$. These bins are chosen such that they have equal width, or equal numbers of examples. The *confidence* of each bin is given as the average prediction of examples that fall inside the bin, while the *accuracy* of each bin is the empirical accuracy:

$$\text{acc}(B_k) = \frac{1}{|B_k|}\sum_{i \in B_k}\mathbb{I}(\hat{y}_i = y_i),$$

$$\text{conf}(B_k) = \frac{1}{|B_k|}\sum_{i \in B_k}\hat{p}_i$$

where y_i is the true binary label, $\hat{y}_i$ is the predicted label, and $\hat{p}_i$ is the predicted probability for an instance i. Intuitively, a well calibrated classifier should have comparable confidence and accuracy for each bin. The confidence and accuracy are plotted against each other for each bin, producing a straight diagonal line for a well calibrated classifier (Fig. 2).

Naeini, Cooper, and Hauskrecht (2015) applied the same idea to give a direct quantitative measure of calibration, called the *expected calibration error* (ECE):

$$\text{ECE} = \sum_{k=1}^{K}\frac{|B_k|}{n}\left|\text{acc}(B_k) - \text{conf}(B_k)\right|.$$

This is simply a weighted average of the residuals in a reliability diagram, weighted by the number of instances that fall inside each bin.

Internal Calibration

In this section, we investigate the effect of calibrating the internal models of nested dichotomies. Our hypothesis is that by improving the quality of the binary probability estimates, the final multiclass predictive performance will be improved. This is due to the fact that multiclass probability estimates are produced by computing the product of a series of binary probability estimates. If the binary probability estimates are not well calibrated, then these errors will accumulate throughout the calculation.

Theoretical Motivation

It seems reasonable that improving the calibration of internal models will result in superior probability estimates for the nested dichotomy, but can we theoretically quantify this improvement? It turns out that reducing the binary NLL of any internal model by some amount δ strictly reduces the multiclass NLL of the nested dichotomy, and depending on the depth of the internal model being calibrated, the reduction in multiclass NLL can be as high as δ.

Proposition 1. *The NLL of an instance under a nested dichotomy is equal to the sum of NLLs of the instance under the binary models on the path from the root node to the leaf node.*

Proof. The NLL of an instance is given by

$$\text{NLL} = -\mathbf{y}_i \log \hat{\mathbf{p}}_i = -\log \hat{\mathbf{p}}_i^{(c)} \tag{1}$$

where $\hat{\mathbf{p}}_i^{(c)}$ is the probability estimate for the true class c. Let $\mathcal{P}_c$ be the set of internal nodes on the path from the root

to the leaf corresponding to class c. Then, $\hat{\mathbf{p}}_i^{(c)}$ can be expressed as

$$\hat{\mathbf{p}}_i^{(c)} = \prod_{k \in \mathcal{P}_c} \tilde{y}_{ik} \hat{p}_{ik} + (1 - \tilde{y}_{ik})(1 - \hat{p}_{ik}) \quad (2)$$

where $\hat{p}_{ik} \in [0, 1]$ is the scalar prediction and $\tilde{y}_{ik} \in \{0, 1\}$ is the meta-label for instance i for the binary model at node k respectively. Because $\tilde{y}_{ik} \in \{0, 1\}$, it is equivalent to write

$$\hat{\mathbf{p}}_i^{(c)} = \prod_{k \in \mathcal{P}_c} \hat{p}_{ik}^{\tilde{y}_{ik}} (1 - \hat{p}_{ik})^{(1 - \tilde{y}_{ik})}. \quad (3)$$

Plugging this into (1) yields

$$\text{NLL} = -\log \prod_{k \in \mathcal{P}_c} \hat{p}_{ik}^{\tilde{y}_{ik}} (1 - \hat{p}_{ik})^{(1 - \tilde{y}_{ik})} \quad (4)$$

$$= -\sum_{k \in \mathcal{P}_c} \log \left(\hat{p}_{ik}^{\tilde{y}_{ik}} (1 - \hat{p}_{ik})^{(1 - \tilde{y}_{ik})} \right) \quad (5)$$

$$= -\sum_{k \in \mathcal{P}_c} \tilde{y}_{ik} \log \hat{p}_{ik} + (1 - \tilde{y}_{ik}) \log(1 - \hat{p}_{ik}), \quad (6)$$

the sum of NLLs from the models $k \in \mathcal{P}_c$. $\qquad \square$

It directly follows that reducing the binary NLL for the model at internal node k by some amount δ results in a reduction of the multiclass NLL by δ for each class corresponding to the leaf nodes that are descendants of node k. This means that a calibration resulting in a binary NLL reduction of δ for some internal node k reduces the multiclass NLL by $\delta(n_k/n)$, where n_k is the number of examples that belong to classes whose corresponding leaf nodes are descendants of k.

The number of descendant leaf nodes for an internal node k in a balanced tree equals 2^l, where l is the length of the paths to the leaf nodes from k. Therefore, it is more important to ensure the models nearer the root node are well calibrated than the models nearer leaf nodes. However, calibration of entire layers of a nested dichotomy should have an effect that is independent of the particular layer being calibrated. Note that even though each layer has twice as many internal models as the layer before it, each internal model at the lower layer is trained on approximately half the amount of data as the models in the previous layer. Platt scaling and isotonic regression, the internal calibration models we consider in our experiments, both have linear complexity in the number of examples, so the time taken to train calibration models for each layer of a nested dichotomy should be comparable.

External Calibration

As well as calibrating each internal model, we also consider external calibration of the entire nested dichotomy. Even models like logistic regression are usually not perfectly calibrated in practice. We hypothesise that these minor miscalibrations accumulate as the nested dichotomy gets deeper, leading to overall more serious miscalibration in the overall model, which can be rectified by an external calibration model. Naturally, this effect is greater for problems with more classes, as the paths to leaf nodes will be longer.

Table 1: Datasets used in our experiments.

Name	Instances	Features	Classes
optdigits[1]	5,620	64	10
micromass[1]	571	1,301	20
letter[1]	20,000	16	26
devanagari[1]	92,000	1,000	46
RCV1[2]	15,564/518,571	47,236	53
sector[2]	6,412/3,207	55,197	105
ALOI[2]	97,200/10,800	128	1,000
ILSVR2010[3]	1,111,406/150,000	1,000	1,000
ODP-5K[4]	361,488/180,744	422,712	5,000

[1] UCI Repository (Lichman 2013)
[2] LIBSVM Repository (Chang and Lin 2011)
[3] ImageNet (Russakovsky et al. 2015)
[4] ODP (Bennett and Nguyen 2009)

As an illustrative investigation into the effect of nested dichotomy depth on their calibration, we built a nested dichotomy with logistic regression base learners for the ALOI dataset (see Tab. 1). Figure 3 shows reliability plots for this nested dichotomy that has been "cut-off" at incrementally increasing depths. A test example is considered to be classified correctly at depth d if its actual class is in the subset of classes $\mathcal{C}_k$ of the node k with highest probability and maximum depth d. Limited to a depth of one, the nested dichotomy is simply a single binary logistic regression model which exhibits good calibration. However, as the depth cutoff limit increases, it is clear that the nested dichotomy becomes increasingly *under-confident*, i.e., bins that have high accuracy often have low confidence (Fig. 3, top row). This corresponds to the reliability curve sitting above the diagonal line. The ECE increases approximately linearly with the depth of the tree.

This is adequately and efficiently compensated for by applying vector scaling (Fig. 3, bottom row). Vector scaling exhibits low complexity in the number of classes—only two parameters per class—making it suitable for problems with many classes typically handled by nested dichotomies. For externally calibrated nested dichotomies, the ECE initially increases linearly with the depth of the tree (although for $d > 1$, the ECE values are much lower than their uncalibrated counterparts). However, at $d = 5$, the ECE levels off and even begins to decrease slightly.

Experiments

In this section, we present experimental results of calibration of nested dichotomies using different base classifiers on a series of datasets. The datasets we used in our experiments are listed in Table 1, and were chosen to span a range of numbers of classes. Optdigits, letter and devanagari (Acharya, Pant, and Gyawali 2015) are character recognition datasets for digits, latin letters, and Devanagari script respectively. Micromass (Mahé et al. 2014) is for the identification of microorganisms from mass spectroscopy data. RCV1, sector and ODP-5K are text categorisation tasks, while ALOI and ILSVR2010 are object recognition tasks. We use the
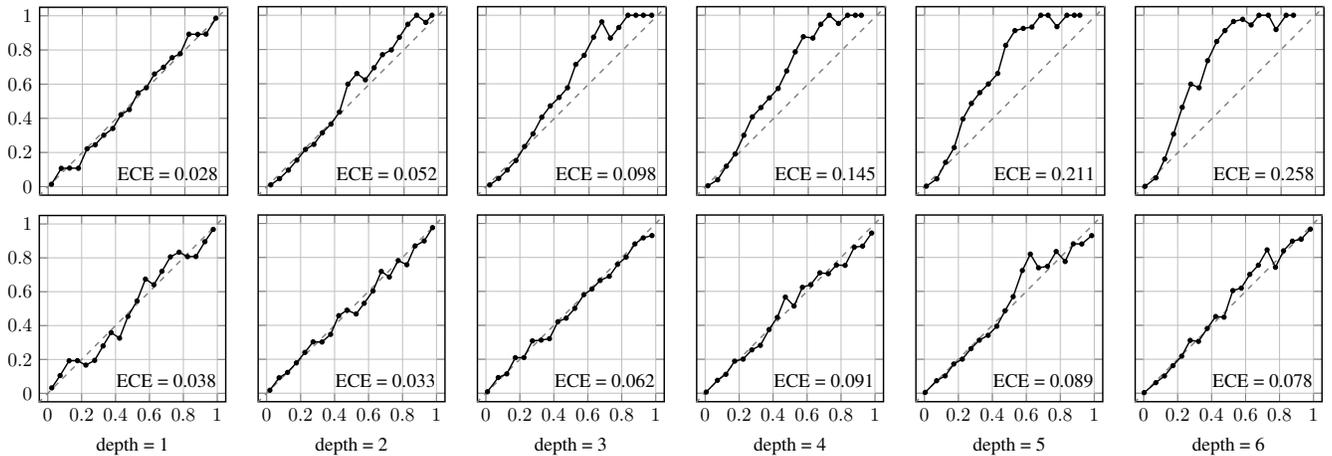
Figure 3: Reliability plots for a nested dichotomy, cut off at increasing depth. The nested dichotomy has logistic regression base learners, which individually, are well calibrated. Top: no external calibration. As the depth increases, the nested dichotomy becomes increasingly under-confident because of the effect of multiplying probabilities together. Bottom: externally calibrated using vector scaling.

Table 2: NLL of nested dichotomies with logistic regression, before and after external calibration is applied.

Dataset	Baseline	External VS
optdigits	0.302 (0.07)	**0.301 (0.08)**
micromass	5.918 (1.83)	**1.878 (0.51)**
letter	1.502 (0.06)	**1.435 (0.08)**
devanagari	2.430 (0.11)	**2.028 (0.05)**
RCV1	1.004 (0.02)	**0.584 (0.01)**
sector	2.858 (0.01)	**1.248 (0.02)**
ALOI	3.604 (0.02)	**3.050 (0.03)**
ILSVR2010	6.442 (0.01)	**5.780 (0.01)**
ODP-5K	5.792 (0.01)	**4.967 (0.01)**

Table 3: Classification accuracy of nested dichotomies with logistic regression, before and after external calibration.

Dataset	Baseline	External VS
optdigits	0.905 (0.02)	**0.906 (0.03)**
micromass	**0.804 (0.06)**	0.772 (0.05)
letter	0.512 (0.03)	**0.536 (0.03)**
devanagari	0.428 (0.02)	**0.428 (0.02)**
RCV1	0.814 (0.01)	**0.855 (0.00)**
sector	0.848 (0.01)	**0.867 (0.00)**
ALOI	0.274 (0.01)	**0.331 (0.01)**
ILSVR2010	**0.063 (0.00)**	0.053 (0.00)
ODP-5K	0.189 (0.00)	**0.228 (0.00)**

visual codewords representation for ILSVR2010, available from the ImageNet competition website.

In order to obtain performance estimates, we performed 10 times 10-fold cross-validation for the datasets from the UCI repository, while adopting the standard train/test splits for the larger datasets with a larger number of classes ($m > 50$). The number of instances stated in Table 1 for the larger datasets are split into number of training and test instances. Note that in each fold and run of 10 times 10-fold cross-validation, a different random nested dichotomy structure is constructed. In the case of the larger datasets, the average of 10 randomly constructed nested dichotomies is reported. Standard deviations are given in parentheses, and the best result per row appears in bold face. The original ODP dataset contains 105,000 classes—we took the subset of the most frequent 5,000 classes to create ODP-5K for the purposes of this investigation. We also reduce the dimensionality to 1,000 when evaluating the performance on boosted trees, by using a Gaussian random projection (Bingham and Mannila 2001).

We implemented vector scaling (Guo et al. 2017) and nested dichotomies in Python, and used the implementations of the base learners, isotonic regression and Platt scaling available in scikit-learn (Pedregosa et al. 2011).

Well Calibrated Base Learners

As shown in Figure 3, overall calibration of nested dichotomies can degrade as the depth of the tree increases, even if the base learners are well calibrated. To further investigate the effects of nested dichotomy depth on predictive performance, we performed experiments to determine the extent to which the classification accuracy and NLL are affected as well.

Tables 2 and 3 show the NLL and accuracy respectively of nested dichotomies with logistic regression, before and after external calibration is applied. Logistic regression models are known to be well-calibrated (Niculescu-Mizil and Caruana 2005). Vector scaling (Guo et al. 2017) is used as the external calibration model. We use a 10% stratified sample of the training data to train the external calibration model, and

Table 4: NLL of nested dichotomies with poorly calibrated base classifiers.

Base Model	Dataset	Baseline	External VS	Internal PS	Both PS	Internal IR	Both IR
Naïve Bayes	optdigits	4.252 (0.92)	0.841 (0.13)	0.852 (0.11)	0.807 (0.09)	0.714 (0.12)	**0.642 (0.09)**
	micromass	8.668 (1.72)	1.624 (0.42)	0.926 (0.08)	0.752 (0.12)	0.766 (0.12)	**0.712 (0.15)**
	letter	2.338 (0.08)	2.155 (0.08)	2.165 (0.06)	2.068 (0.07)	2.055 (0.07)	**1.953 (0.06)**
	devanagari	13.14 (0.59)	3.310 (0.16)	2.986 (0.05)	2.602 (0.02)	2.754 (0.07)	**2.444 (0.05)**
	RCV1	1.690 (0.19)	**0.866 (0.01)**	1.145 (0.07)	0.947 (0.03)	0.998 (0.04)	0.914 (0.02)
	sector	3.795 (0.36)	**1.408 (0.09)**	2.075 (0.20)	1.518 (0.10)	1.913 (0.21)	1.774 (0.09)
	ALOI	32.98 (0.43)	6.841 (0.02)	5.533 (0.02)	4.333 (0.03)	4.859 (0.03)	**4.137 (0.01)**
	ILSVR2010	32.39 (0.20)	6.819 (0.00)	6.162 (0.00)	6.125 (0.01)	6.176 (0.00)	**6.116 (0.00)**
	ODP-5K	8.498 (0.31)	**5.161 (0.05)**	6.103 (0.00)	5.518 (0.02)	6.051 (0.11)	5.364 (0.01)
Boosted Trees	optdigits	3.861 (0.57)	0.634 (0.07)	0.403 (0.04)	**0.298 (0.04)**	0.390 (0.03)	0.303 (0.04)
	micromass	10.01 (2.05)	2.518 (0.52)	1.263 (0.11)	1.005 (0.14)	1.235 (0.27)	**0.959 (0.19)**
	letter	4.869 (0.27)	0.924 (0.04)	0.563 (0.02)	**0.446 (0.03)**	0.557 (0.02)	0.449 (0.03)
	devanagari	3.424 (0.28)	1.030 (0.04)	2.266 (0.17)	**0.710 (0.02)**	1.977 (0.12)	0.735 (0.02)
	RCV1	1.964 (0.02)	1.029 (0.00)	0.932 (0.01)	**0.716 (0.01)**	0.862 (0.00)	0.745 (0.01)
	sector	3.631 (0.20)	2.910 (0.11)	2.672 (0.03)	**2.036 (0.03)**	2.597 (0.05)	2.208 (0.07)
	ALOI	4.443 (0.26)	2.512 (0.05)	4.889 (0.03)	**1.056 (0.02)**	4.284 (0.04)	1.173 (0.03)
	ILSVR2010	6.553 (0.10)	5.863 (0.00)	5.643 (0.00)	5.219 (0.00)	5.452 (0.00)	**5.201 (0.00)**
	ODP-5K	7.733 (0.04)	7.192 (0.00)	7.120 (0.00)	6.603 (0.00)	6.981 (0.00)	**6.576 (0.00)**

Table 5: Classification accuracy of nested dichotomies with poorly calibrated base classifiers.

Base Model	Dataset	Baseline	External VS	Internal PS	Both PS	Internal IR	Both IR
Naïve Bayes	optdigits	0.719 (0.05)	0.749 (0.04)	0.719 (0.05)	0.735 (0.04)	0.774 (0.04)	**0.795 (0.04)**
	micromass	0.749 (0.05)	0.724 (0.05)	0.770 (0.05)	0.762 (0.05)	**0.772 (0.05)**	0.756 (0.05)
	letter	0.329 (0.02)	0.364 (0.03)	0.318 (0.03)	0.365 (0.03)	0.376 (0.03)	**0.412 (0.03)**
	devanagari	0.202 (0.02)	0.224 (0.04)	0.167 (0.02)	0.269 (0.01)	0.265 (0.04)	**0.340 (0.01)**
	RCV1	0.644 (0.04)	**0.781 (0.00)**	0.691 (0.03)	0.756 (0.00)	0.734 (0.01)	0.765 (0.01)
	sector	0.337 (0.07)	**0.772 (0.01)**	0.633 (0.04)	0.737 (0.03)	0.692 (0.04)	0.690 (0.01)
	ALOI	0.024 (0.00)	0.029 (0.00)	0.019 (0.00)	0.124 (0.00)	0.094 (0.00)	**0.166 (0.01)**
	ILSVR2010	0.009 (0.00)	0.015 (0.00)	0.014 (0.00)	0.019 (0.00)	0.021 (0.00)	**0.026 (0.00)**
	ODP-5K	0.043 (0.01)	**0.210 (0.00)**	0.091 (0.00)	0.161 (0.00)	0.135 (0.01)	0.180 (0.00)
Boosted Trees	optdigits	0.888 (0.02)	0.883 (0.02)	**0.922 (0.01)**	0.917 (0.01)	0.921 (0.01)	0.915 (0.01)
	micromass	0.710 (0.06)	0.650 (0.06)	**0.741 (0.06)**	0.729 (0.06)	0.737 (0.06)	0.727 (0.05)
	letter	0.859 (0.01)	0.851 (0.01)	0.888 (0.01)	0.883 (0.01)	**0.890 (0.01)**	0.884 (0.01)
	devanagari	0.103 (0.06)	0.710 (0.01)	0.488 (0.11)	**0.793 (0.01)**	0.636 (0.04)	0.783 (0.01)
	RCV1	0.681 (0.06)	0.748 (0.01)	0.810 (0.00)	**0.814 (0.01)**	0.810 (0.00)	0.807 (0.00)
	sector	0.174 (0.08)	0.409 (0.02)	**0.603 (0.01)**	0.576 (0.01)	0.578 (0.01)	0.552 (0.01)
	ALOI	0.071 (0.03)	0.451 (0.01)	0.161 (0.01)	**0.743 (0.01)**	0.368 (0.01)	0.723 (0.00)
	ILSVR2010	0.019 (0.00)	0.040 (0.00)	0.093 (0.00)	**0.102 (0.00)**	0.097 (0.00)	0.100 (0.00)
	ODP-5K	0.032 (0.00)	0.038 (0.00)	0.055 (0.00)	0.068 (0.00)	0.062 (0.00)	**0.072 (0.00)**

the remaining 90% to build the nested dichotomy including the base classifiers.

Discussion. Tables 2 and 3 show that, for all datasets, a reduction in NLL is observed after applying external calibration with vector scaling (External VS). For some of the datasets with fewer classes (optdigits and letter), the reduction is modest, but the larger datasets see substantial improvements. Interestingly, for some datasets a large improvement in classification accuracy is also observed, especially for the datasets with more classes. Also, the classification accuracy degrades for ILSVR2010 and micromass,

despite a large improvement in NLL.

Poorly Calibrated Base Learners

Tables 4 and 5 show the NLL and classification accuracy respectively of nested dichotomies trained with poorly calibrated base learners, when different calibration strategies are applied. Specifically, we considered nested dichotomies with naïve Bayes and boosted decision trees as the base learners. The calibration schemes compared are internal Platt scaling (Internal PS), internal isotonic regression (Internal IR) and external vector scaling (External VS), as well

as each internal calibration scheme in conjunction with external vector scaling (Both PS and Both IR, respectively). Three-fold cross validation is used to produce the training data for the internal calibration models, rather than splitting the training data. This is to ensure that each internal calibration model has a reasonable amount of data points to train on, given that internal nodes near the leaves often have few training data available. When external calibration is performed, 10% of the data is held out to train the external calibration model. Note that this means 10% less data is available to train the nested dichotomy and (if applicable) perform internal calibration. Gaussian naïve Bayes is applied for optdigits, micromass, letter and devanagari, and multinomial naïve Bayes is used for RCV1, sector, ALOI, ILSVR2010 and ODP-5K as they have sparse features. We use 50 decision trees with AdaBoost (Freund and Schapire 1996), limiting the depth of the trees to three.

Discussion. Tables 4 and 5 show that applying internal calibration is very beneficial in terms of both NLL and classification accuracy. There is no combination of base learner and dataset for which the baseline gives the best results, and there are very few cases where the baseline does not perform the worst out of every scheme. When naïve Bayes is used as the base learner, applying internal calibration with isotonic regression always gives better results than the baseline, and when an ensemble of boosted trees is used as the base learner, applying internal Platt scaling always outperforms the uncalibrated case. It is well known that these calibration methods are well-suited to the respective base learners (Niculescu-Mizil and Caruana 2005), and this appears to also apply when they are used in a nested dichotomy.

External calibration also has a positive effect on both NLL and classification accuracy in most cases compared to the baseline. However, the best results are usually obtained when both internal and external calibration are applied together. For naïve Bayes, the smaller datasets as well as the two object recognition datasets (ALOI and ILSVR2010) generally see the best performance for both NLL and classification accuracy when applying internal isotonic regression in conjunction with external calibration. Interestingly, the best results for the three text categorisation datasets (RCV1, sector and ODP-5K) were obtained through external calibration only.

Performing both internal Platt scaling and external calibration also gives the best NLL performance for nested dichotomies with boosted trees in most cases, although the improvement compared to isotonic regression is usually small. However, with boosted trees, performance in terms of classification accuracy is less consistent, often being greater when only internal calibration is applied.

Conclusion

In this paper, we show that the predictive performance of nested dichotomies can be substantially improved by applying calibration techniques. Calibrating the internal models increases the likelihood that the path to the leaf node corresponding to the true class is assigned high probability, while external calibration can correct for the systematic under-

confidence exhibited by nested dichotomies. Both of these techniques have been empirically shown to provide large performance gains in terms of accuracy and NLL for a range of datasets when applied individually. Additionally, when both internal and external calibration are applied together, the performance often improves further. Improvements are especially noticeable when the number of classes is high.

Future work in this domain includes evaluating alternative external calibration methods. In our experiments, we applied vector scaling as it is an efficient and scalable solution for large multiclass tasks. However, when computational resources are available, it is possible that employing a more complex method such as matrix scaling, or isotonic regression with one-vs-rest, could provide superior results. It would also be interesting to investigate whether such calibration measures are as effective for other methods of constructing nested dichotomies than random subset selection (Dong, Frank, and Kramer 2005; Leathart, Pfahringer, and Frank 2016; Wever, Mohr, and Hüllermeier 2018; Melnikov and Hüllermeier 2018; Leathart et al. 2018). We expect that the calibration techniques discussed in this paper will transfer to such methods. Lastly, it would be of value to evaluate the effect of selective calibration of layers in nested dichotomies. It may be the case that reductions in training time can be achieved by only calibrating the worst models in the tree with little impact on predictive performance. It may also be beneficial to apply Platt scaling to internal models with few training data points instead of isotonic regression, as isotonic regression is known to overfit on smaller samples.

References

Acharya, S.; Pant, A. K.; and Gyawali, P. K. 2015. Deep learning based large scale handwritten devanagari character recognition. In *Proceedings of the International Conference on Software, Knowledge, Information Management and Applications*, 1–6. IEEE.

Agrawal, R.; Gupta, A.; Prabhu, Y.; and Varma, M. 2013. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the International World Wide Web Conference*, 13–24.

Bengio, S.; Weston, J.; and Grangier, D. 2010. Label embedding trees for large multi-class tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, 163–171.

Bennett, P. N., and Nguyen, N. 2009. Refined experts: improving classification in large taxonomies. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11–18. ACM.

Beygelzimer, A.; Langford, J.; and Ravikumar, P. 2009. Error-correcting tournaments. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 247–262. Springer.

Bingham, E., and Mannila, H. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Pro-*

ceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 245–250. ACM.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27.

Choromanska, A. E., and Langford, J. 2015. Logarithmic time online multiclass prediction. In *Proceedings of the Conference on Neural Information Processing Systems*, 55–63.

Daumé, III, H.; Karampatziakis, N.; Langford, J.; and Mineiro, P. 2017. Logarithmic time one-against-some. In *Proceedings of the International Conference on Machine Learning*, 923–932. PMLR.

DeGroot, M. H., and Fienberg, S. E. 1983. The comparison and evaluation of forecasters. *The Statistician* 12–22.

Dekel, O., and Shamir, O. 2010. Multiclass-multilabel classification with more classes than examples. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 137–144. PMLR.

Dembczyński, K.; Kotłowski, W.; Waegeman, W.; Busa-Fekete, R.; and Hüllermeier, E. 2016. Consistency of probabilistic classifier trees. In *Proceedings of the Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 511–526. Springer.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Dietterich, T. G., and Bakiri, G. 1995. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research* 2:263–286.

Dong, L.; Frank, E.; and Kramer, S. 2005. Ensembles of balanced nested dichotomies for multi-class problems. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer. 84–95.

Fox, J. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage.

Frank, E., and Kramer, S. 2004. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the International Conference on Machine Learning*, 39–46. ACM.

Freund, Y., and Schapire, R. E. 1996. Game theory, on-line prediction and boosting. In *Proceedings of the Conference on Computational Learning Theory*, 325–332.

Friedman, J. H. 1996. Another approach to polychotomous classification. *Technical Report, Statistics Department, Stanford University*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, 1321–1330. PMLR.

Jiang, X.; Osl, M.; Kim, J.; and Ohno-Machado, L. 2011. Smooth isotonic regression: A new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings* 2011:16.

Kumar, A.; Vembu, S.; Menon, A. K.; and Elkan, C. 2013. Beam search algorithms for multilabel learning. *Machine Learning* 92(1):65–89.

Leathart, T.; Frank, E.; Pfahringer, B.; and Holmes, G. 2017. Probability calibration trees. In *Proceedings of the Asian Conference on Machine Learning*, 145–160. PMLR.

Leathart, T.; Frank, E.; Pfahringer, B.; and Holmes, G. 2018. Ensembles of nested dichotomies with multiple subset selection. *arXiv preprint arxiv:1809.02740*.

Leathart, T.; Pfahringer, B.; and Frank, E. 2016. Building ensembles of adaptive nested dichotomies with random-pair selection. In *Proceedings of the Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 179–194. Springer.

Lichman, M. 2013. UCI machine learning repository.

Mahé, P.; Arsac, M.; Chatellier, S.; Monnin, V.; Perrot, N.; Mailler, S.; Girard, V.; Ramjeet, M.; Surre, J.; Lacroix, B.; et al. 2014. Automatic identification of mixed bacterial species fingerprints in a maldi-tof mass-spectrum. *Bioinformatics* 30(9):1280–1286.

Melnikov, V., and Hüllermeier, E. 2018. On the effectiveness of heuristics for learning nested dichotomies: an empirical analysis. *Machine Learning* 107(8-10):1–24.

Mena, D.; Montañés, E.; Quevedo, J. R.; and Del Coz, J. J. 2015. Using A* for inference in probabilistic classifier chains. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Naeini, M.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2901–2907.

Niculescu-Mizil, A., and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the International Conference on Machine Learning*, 625–632. ACM.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Platt, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 10(3):61–74.

Rifkin, R., and Klautau, A. 2004. In defense of one-vs-all classification. *Journal of Machine Learning Research* 5:101–141.

Rodríguez, J. J.; García-Osorio, C.; and Maudes, J. 2010. Forests of nested dichotomies. *Pattern Recognition Letters* 31(2):125–132.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.

Wever, M.; Mohr, F.; and Hüllermeier, E. 2018. Ensembles of evolved nested dichotomies for classification. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 561–568. ACM.

Zadrozny, B., and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the International Conference on Machine Learning*, 609–616. ACM.

Zadrozny, B., and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699. ACM.

Zhong, W., and Kwok, J. T. 2013. Accurate probability calibration for multiple classifiers. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1939–1945.