Reliably Calibrated Isotonic Regression

Nyberg, Otto

Springer International Publishing AG
2021

Nyberg , O & Klami , A 2021 , Reliably Calibrated Isotonic Regression . in K Karlapalem , H Cheng , N Ramakrishnan , R K Agrawal , P K Reddy , J Srivastava & T Chakraborty (eds) , Advances in Knowledge Discovery and Data Mining . Lecture Notes in Artificial Intelligence , vol. 12712 , Springer International Publishing AG , pp. 578-589 , 25th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) , 11/05/2021 . https://doi.org/10.1007/978-3-030-75

http://hdl.handle.net/10138/343456
https://doi.org/10.1007/978-3-030-75762-5_46

acceptedVersion

Reliably Calibrated Isotonic Regression

Otto Nyberg[0000−0002−8481−9325] and Arto Klami[0000−0002−7950−1355]

University of Helsinki, Finland
https://www.helsinki.fi
{otto.nyberg,arto.klami}@helsinki.fi

Abstract. Using classifiers for decision making requires well-calibrated probabilities for estimation of expected utility. Furthermore, knowledge of the reliability is needed to quantify uncertainty. Outputs of most classifiers can be calibrated, typically by using isotonic regression that bins classifier outputs together to form empirical probability estimates. However, especially for highly imbalanced problems it produces bins with few samples resulting in probability estimates with very large uncertainty. We provide a formal method for quantifying the reliability of calibration and extend isotonic regression to provide reliable calibration with guarantees for width of credible intervals of the probability estimates. We demonstrate the method in calibrating purchase probabilities in e-commerce and achieve significant reduction in uncertainty without compromising accuracy.

Keywords: Isotonic regression · Calibration · E-commerce

1 Introduction

Even though classification is in academic contexts often studied in isolation, in real use the predictions are used for making decisions with associated costs and benefits. Evaluating task performance requires knowledge of the probability of each possible class, and for accurate evaluation we need *well-calibrated* probabilities [10, 22, 20]. A binary classifier is said to be well-calibrated if the empirical probability (relative frequency) of the positive class $p(y = 1|s(x) = x)$ converges to the output score $s(x)$ of the classifier at the limit of infinite data.

The definition extends naturally to multi-class problems, but for simplicity of notation we consider binary problems used e.g. in medical diagnosis [2], credit scoring [8] and e-commerce [19, 9].

Most classifiers do not directly produce well-calibrated probabilities. Many methods like deep neural networks with logistic outputs or many decision trees formally output probabilities, but they are often poorly calibrated in practice [14, 6], whereas other models like support vector machines output scores that are larger when the classifier is more certain of the result but do not even attempt to represent probabilities. The outputs of all such classifiers can be calibrated after training. Various algorithms have been proposed for this [16, 13, 12, 7] but the classical method of *isotonic regression (IR)* [22, 15] remains the most common

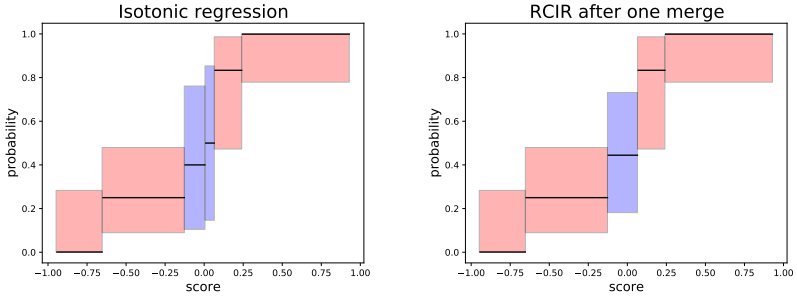

Fig. 1. Isotonic regression (left) calibrates classifier outputs by binning, but the probability estimates (horizontal lines) are unreliable as shown here using 95% credible intervals (boxes) on artificial data. Reliably calibrated isotonic regression merges bins further (right figure shows the result after one merge combining the two bins indicated by blue), retaining calibration and monotonicity while reducing uncertainty. In this example the maximum credible interval width decreased by 22% in one step, and on real data we can often achieve 70-90% reduction in uncertainty while retaining almost identical accuracy – see Figure 2 for an example.

approach; it works particularly well with large sample sizes and is robust over different classification problems. For an overview and empirical comparison of calibration methods in context of neural network classifiers, arguably the most relevant practical model family today, see [6].

Isotonic regression produces a function that maps the raw scores to well-calibrated probabilities, but the probabilities are calibrated only for the training data; each bin is associated with an empirical estimate for the probability but the algorithm provides no guarantees on the true value. We analyse isotonic regression from a perspective of reliability. We use Bayesian statistics to provide credible intervals for the true bin probability and observe that in many cases the credible intervals for IR outputs are extremely wide. This makes the method fragile and unreliable, even though the probability estimates are unbiased. The problem occurs in particular in imbalanced classification problems where the bins capturing the rare positive samples are the most uncertain, causing problems especially for the samples that are most interesting. As an example, in e-commerce the task is often to predict customer conversion with rates below 3%, and sometimes as small as 0.2% [3]. Quantifying the uncertainty is also important for other tasks such as medical diagnostics [20], credit scoring [8], and in modern portfolio management theory where variance is one of the fundamental building blocks [18].

To address the high uncertainty of IR, we extend isotonic regression in a manner that provides reliable estimates, coining the method *Reliably Calibrated Isotonic Regression (RCIR)*. The goal of the method is to provide well-calibrated probability estimates that additionally have low variance and probabilistic guarantees for maximum deviation, so that in downstream applications we can trust

the available estimates and handle the uncertainty as necessary. This is achieved by similar merging process that is at the core of IR, but that now uses uncertainty reduction measured by credible interval compression as criterion; see Figure 1 for illustration of the concept. The algorithm monotonically improves reliability by reducing output resolution (the number of bins). We show that significant improvements in credible interval width can be achieved with negligible loss in testing set accuracy.

We demonstrate the algorithm on real world data relating to behavioral modeling of online users. We show that for typical binary classification tasks with class-imbalance, regular isotonic regression produces bins with extremely wide credible intervals despite relatively large total sample size. Our algorithm compresses the credible intervals to a fraction of the original ones without compromising the accuracy of the classifier.

2 Background: Isotonic Regression

To provide sufficient background for the rest of the paper, we start by a recap of isotonic regression and its limitations.

Assume that we have a binary classifier $s(\mathbf{x}_i)$ that outputs scores z_i corresponding to the input features $\mathbf{x}_i$ for sample i. No assumptions on how the classifier has been trained is made, but higher scores z_i are assumed to more likely belong to the positive class. The goal of isotonic regression [22] is to calibrate the classifier so that the calibrated output, denoted by $g(z)$, matches the empirical ratio of the positive class among the training samples with the same output. That is, it seeks to find the function $g(z)$ that maps the arbitrary scores into actual probabilities of the positive class.

IR finds the function $g(\cdot)$ that minimizes the mean square error (MSE)

$$L(g) = \frac{1}{N} \sum_i \big(y_i - g(z_i)\big)^2 \tag{1}$$

under the constraint that g must be monotonic, i.e. $g(a) \geq g(b)$ for $a > b$. Here y_i is the dependent variable of sample i. The optimal solution is a piecewise constant function g that maps score ranges to some positive *values*. For a binary class variable $y \in \{0, 1\}$, these values are well-calibrated probabilities for the positive class [22]. The range of scores that map to a certain value is called a *bin*. At training time, a number of samples with similar scores are assigned to a bin, and the value (probability estimate) associated with the bin is the relative frequency of the positive class samples in that bin.

A global optimum of the objective (1) is obtained with the pair-adjacent violators algorithm [1] that starts by ordering all samples according to score. The algorithm then goes through all samples adjusting bin boundaries to smooth out any violations of the monotonicity, terminating when there are none left. Besides minimizing MSE, the result maximizes AUC-ROC [17]. While AUC-ROC is completely insensitive to calibration, it is an actual measure of refinement, i.e.

it measures how well positive samples are separated from negative ones. As MSE can be decomposed into refinement and reliability, we think AUC-ROC together with a calibration metric is a more comprehensive measure of goodness.

Despite being frequently used, IR has limitations especially in problems with high class-imbalance. The resolution (the number of bins) of the output is determined solely by the monotonicity criterion with no user control, and for imbalanced setups some of the bins often contain very few samples [15]. For such bins there are no guarantees for the empirical estimate to accurately represent the true probability on test data. Many real-world classification problems are highly imbalanced so that the class of interest is smaller, and for such problems IR creates the smallest bins for the high probability samples. This is problematic because this happens precisely for the samples we care most about. For example, the conversion rate in online stores is typically below 3%, as is click-through rate of ads [3], and the goal is to identify users most likely to take action, yet IR provides the least reliable estimates exactly for those users.

3 Method

We first describe a formal procedure for inspecting isotonic regression in terms of reliability of the probability estimates, and then describe a practical algorithm for improving the reliability by merging bins with too high uncertainty.

3.1 Credible Intervals for Isotonic Regression

Isotonic regression assigns for each bin a single value, which in the case of binary classification corresponds to the empirical ratio of training samples in the positive class falling into this bin. This is equivalent to assuming a Bernoulli model

$$p(y = 1|s(x) = z_i) = \text{Bernoulli}(\theta_i)$$

and using maximum likelihood estimator for inferring the parameter θ_i.

We use Bayesian statistics to characterize the posterior distribution of the parameter using the conjugate prior $\text{Beta}(\alpha, \beta)$. Straightforward calculus (see, e.g. [5]) then provides the posterior

$$p(\theta_i|\text{data}) = \text{Beta}(k_i + \alpha, n_i - k_i + \beta) = \frac{\theta^{k_i+\alpha-1} \times (1-\theta)^{n_i-k_i+\beta-1}}{B(k_i + \alpha - 1, n_i - k_i + \beta - 1)},$$

where n_i indicates the number of samples falling into the ith bin, k_i is the number of samples in the positive class, and $B(a, b)$ is the beta-function. We use $\alpha = \beta = 1$ to indicate uniform prior, but other choices would be equally easy to implement. We characterize the posterior distribution using highest posterior density (HPD) credible intervals [11] $H(p(\cdot), c)$ corresponding to the smallest continuous range of parameter values capturing a given total mass $1 - c$ (where the confidence level c is often set to 0.05) of the posterior such that for all $\theta \in H(p(\cdot), c)$ we have $p(\theta|\text{data}) > p_0$ for some threshold p_0. This range is

typically not symmetric around the posterior mean, but captures the intuitive idea of credible alternatives better than central intervals – every point within the region has higher posterior density than any of the points outside it.

Since the beta distribution is unimodal, we can efficiently find the HPD interval by binary search in the log domain. Defining the credible interval as $[l, h]$, we start by making an initial guess for l and find h such that $p(\theta = h|\text{data}) = p(\theta = l|\text{data})$ and $l < h$. Here the denominator can be ignored and the search is fast. We then refine l (and consequently h) using binary search until the total probability mass within the region is sufficiently close to $1 - c$.

3.2 Reliably Calibrated Isotonic Regression

The only way of obtaining narrow credible intervals is to guarantee sufficiently many samples in each bin, but the required number depends on both n and k and hence we cannot set direct thresholds upfront. We can, however, set a threshold for the width of the HPD credible interval. This provides us the formal definition of reliably calibrated isotonic regression (RCIR):

Definition 1. Reliably calibrated isotonic regression: *Given a binary classifier that outputs scores z_i for N samples with true classes y_i, minimize the objective*

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - g(z_i))^2$$

under the constraints that $g(\cdot)$ is a monotonic function and for every distinct value of $g(\cdot)$ the highest posterior density credible interval at confidence level c for the estimated class probability $p(y = 1|g(z_i))$ is at most of width d.

A solution to the above problem provides well-calibrated probabilities in the same sense as regular isotonic regression, but additionally guarantees that the probability estimates provided for the bins are reliable also for future samples at a level chosen by the analyst. That is, for any given sample there are probabilistic guarantees for the estimate to be sufficiently close to the true value.

We will next present a practical approximate algorithm for solving the problem, but before that we make a general remark that motivates our algorithm. As described in Section 2, IR both minimizes MSE of the predictions and maximizes the AUC-ROC; it directly optimizes the former with an algorithm guaranteed to find the optimal solution under the monotonicity constraint, and the latter follows from [4]. Since RCIR incorporates additional constraints for the same objective, IR directly provides a lower bound for MSE and an upper bound for AUC-ROC for any algorithm solving the problem of Definition 1. This only holds for training data; the additional constraints may regularize the solution so that these metrics improve on test data.

3.3 Greedy Optimization Algorithm

We are not aware of an efficient algorithm guaranteed to find the global optimum of the problem in Definition 1. We can, however, derive a practical algorithm

that solves it computationally efficiently and produces solutions that for practical problem instances are very close to optimal. The basic idea is to first solve the unconstrained IR problem, guaranteed to produce optimal solution without the additional constraints. We then refine the solution by merging bins with too wide confidence intervals in such a way that the monotonicity assumption is retained until all credible intervals are sufficiently narrow.

For each merge we greedily select the bin with the widest confidence interval. It can be merged either with the bin to its left or its right; both choices retain monotonicity. As we are already controlling for calibration, we wish to retain as much of the refinement as possible and hence chose the solution that maximizes AUC-ROC of the classifier after the merge. That is, we compute AUC-ROC for both alternative merges and choose the one that decreases the value less, corresponding to a better classifier.

The full algorithm is provided in Algorithm 1.[1] It is a deterministic algorithm that terminates in a finite number of steps since the number of bins is decreased by one at every step, and in practice it is computationally efficient due to merely selecting one bin at a time to be merged until all credible intervals are narrow enough. We do not have approximation bounds for the algorithm, but as illustrated in the empirical experiments the AUC-ROC of the training set of the final calibrated classifier is often extremely close (in our experiments always within 0.01%) to the AUC-ROC of the IR solution that provides an upper bound, and hence even if the solution is not optimal it cannot be far from it.

3.4 Parameter Choice

The algorithm has only two parameters, the maximum width d of credible interval for any of the bins and the associated confidence c. For c one can safely choose a standard value in the order of 0.05. The width parameter d, in turn, has a natural interpretation in terms of the decision problem since it corresponds to the maximum error one is willing to accept in probability estimates. We recommend setting this manually according to the task at hand, but would expect values between 0.1 and 0.2 to be acceptable in many problems.

In absence of domain knowledge for determining d, it can also be set automatically using cross-validation based on the intuition that good values generalize well for test samples. To find the optimal value we leave out a set of validation samples not used for training the classifier or for calibrating it, and evaluate AUC-ROC of the RCIR model on the validation set after every merge. We then choose the refinement level that corresponds to the best validation score.

4 Related Work

The value of calibration has clearly been recognized, exemplified e.g. by the observation that modern deep learning networks are typically poorly calibrated

[1] A Python implementation of the algorithm and the data used in the experiments are available at https://github.com/Trinli/calibration.

Algorithm 1: Greedy algorithm for reliably calibrated isotonic regression

Input : Real-valued scores z_i provided for N training samples with class labels $y_i \in \{0, 1\}$, a confidence level $c \in [0, 1]$, and maximum width of credible interval $d \in [0, 1]$

Output : A collection of bins B_k, each characterized by a range of input values $[a_k, b_k]$ mapped to the bin, and the associated class probabilities θ_k within highest posterior density credible interval $[l_k, h_k]$, such that $h_k - l_k \leq d$.

Initialization: Run pair-adjacent violators algorithm to produce initial bins B_k and estimate the HPD intervals $[l_k, h_k]$ for all k

while $h_k - l_k > d$ *for any bin* k **do**
  Choose $k = \arg\max h_k - l_k$
  Consider merging B_k with B_{k-1} to produce new B_j:
  begin
    Set new input range to $[a_{k-1}, b_k]$
    Compute the new HPD interval $[l_j, h_j]$
    Compute AUC-ROC for a classifier that replaces b_{k-1} and b_k with b_j
  end
  Consider merging B_k with B_{k+1} to produce new B_j:
  begin
    Set new input range to $[a_k, b_{k+1}]$
    Compute the new HPD interval $[l_j, h_j]$
    Compute AUC-ROC for a classifier that replaces B_k and B_{k+1} with B_j
  end
  Choose the better of the two alternatives based on AUC-ROC
  Replace B_K and its chosen neighbor with B_J
end

[6], but arguably has not been receiving sufficient attention [20]. Nevertheless, in recent years a few calibration methods have been proposed to improve on the classical choices of isotonic regression, Platt scaling [16], and histogram binning. We briefly discuss these alternatives here, but note that none of these works address our main goal, the reliability of the estimates. Instead, they are motivated simply as more accurate calibration methods that explicitly optimize for calibration accuracy on the training data. Building on our work for IR, it could be possible to estimate credible intervals also for some of these other methods and develop generalizations that would optimize also for that.

Bayesian Binning by Quantiles (BBQ) [13] builds on histogram binning, forming an ensemble that combines binning models with different bin widths using Bayesian weighting for different models. Even though it achieves good calibration metrics, it has a tendency to underestimate the highest scoring samples in imbalanced cases due to using bins of constant width. The Ensemble of Near-Isotonic Regression (ENIR) [12] is also an ensemble with Bayesian weighting and builds on a variant of isotonic regression that allows small violations to the

monotonicity constraint and typically achieves better empirical performance. As ensembles, both BBQ and ENIR are computationally less efficient (though this is often not a practical issue as training the classifier itself dominates the total computation) and more difficult to interpret and analyse theoretically. Importantly, neither of these optimizes for criteria that would account for reliability of the estimates and they provide no guarantees for it.

The recent scaling-binning calibration method [7] combines Platt scaling and histogram binning to achieve sample efficiency of the former with theoretical properties of the latter, achieving good calibration performance for deep neural networks and focusing in particular on sample complexity. Even though the method is motivated in part via variance reduction, they do not quantify the reliability of the probability estimates either theoretically or empirically, explicitly leaving finite sample guarantees for future work.

5 Experiments

We illustrate the method on data collected on e-commerce sites, where the classification problem concerns predicting a particular decision of the user, matching our eventual use case. For the purpose of this manuscript the specific data sets are largely irrelevant, as is the underlying classifier since we only access its outputs.

We conduct two separate experiments. First we demonstrate the method on a typical imbalanced binary classification problem, showing how we can obtain increasingly tighter bounds for the credible intervals with minimal loss in classification accuracy. The second experiment demonstrates the performance of the automatic selection of d explained in Section 3.4, removing the only tunable parameter in Algorithm 1.

5.1 Experiment 1

For this experiment only users that had more than two pageloads were included resulting in a generous 4.8% positive rate even though the conversion rate for the entire site is below 1.0%. We used a sample of 200,000 visitors not used to train the base classifier and randomly selected 1/3 of the data for testing.

Figure 2 illustrates how the method works, by depicting the bins and their credible intervals for both standard IR and the proposed RCIR for three choices of d. The reliability of the estimates is dramatically improved compared to regular IR, yet the refinement is sufficient as we still retain the important bins with high probability. With excessively small d (bottom right) the calibration further improves, but too much resolution is lost.

Table 1 quantifies the results using MSE and AUC-ROC evaluated on test data. The main observations are that standard IR is extremely unreliable having maximum credible interval width above $d = 0.8$, and that RCIR can dramatically improve reliability without compromising accuracy. In particular, for all $d \in [0.1, 0.3]$ AUC-ROC decreases by less than 0.004% and MSE increases by less

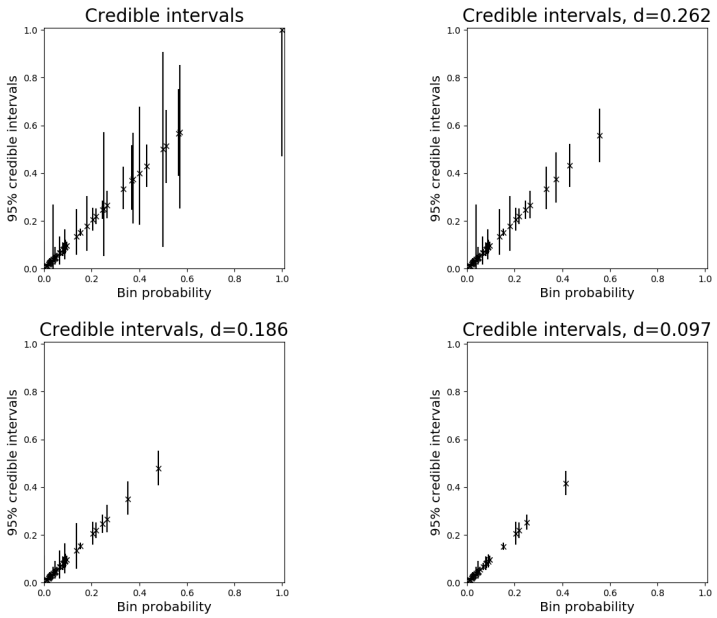

Fig. 2. The top left plot shows the result of isotonic regression with 95% credible intervals for the 45 bins produced. The remaining plots show the results of reliably calibrated isotonic regression for varying levels of reliability d. For these plots we used maximum threholds of 0.3, 0.2 and 0.1 (with $c = 0.05$), and report the actual maximum width in the plot titles.

than 0.03%. Differences this small are irrelevant in practice, yet we achieve over an 80% drop in credible interval width while retaining more than half of the bins. This indicates that the greedy algorithm is able to find a solution that is almost indistinguishable as a classifier from the original one while providing a formal reliability guarantee. The accuracy starts to drop noticeably only when enforcing extremely narrow confidence intervals ($d = 0.01$).

5.2 Experiment 2

To evaluate the automatic procedure for determining d, we use three different data sets. The first is the one used in experiment 1, whereas the other two have slightly higher positive class rates. In all cases the prediction task was the same, i.e. predicting the conversion probability of a visitor.

Table 2 reports the average results over 30 random splits into training, validation, and test sets (1/3 for each). We again see that standard IR is unreliable in all cases, with the largest confidence intervals exceeding 0.65 in all cases. For binary classification intervals this wide implies the bin provides essentially no information on true probability. For RCIR the automatic procedure is able to push the maximum width down to 0.13 or below, achieving always at least an

Table 1. Goodness of fit for different algorithms demonstrating that tighter credible interval requirements (reliably calibrated isotonic regression, RCIR, with different values for d) result naturally in lower number of bins, but do not decrease the overall performance of the classifier as measured by AUC-ROC and MSE.

Model	d	Bins	AUC-ROC	MSE
Uncalibrated classifier	-	-	0.714878	-
Isotonic regression	(0.811400)	45	0.713166	0.0435589
RCIR	.30	38	0.713167	0.0435541
RCIR	.20	34	0.713138	0.0435429
RCIR	.10	28	0.713149	0.0435715
RCIR	.05	23	0.712874	0.0436705
RCIR	.01	9	0.671561	0.0447910

Table 2. Average results of 30 repeated experiments for every data set with randomization. Compared to regular isotonic regression (IR) the proposed reliably calibrated isotonic regression (RCIR) results in nearly identical AUC-ROC and MSE metrics, but reduces the maximum credible interval width more than 80%.

	Data set 1	**Data set 2**	**Data set 3**
Samples [#]	200,000	79,155	200,000
Positive rate [%]	4.799	8.040	7.482
Max credible interval (IR)	0.6586	0.7113	0.6925
AUC-ROC (IR)	0.715180	0.773431	0.760706
MSE (IR)	0.0432544	0.0634838	0.0635181
Bin merges [#] (RCIR)	7.63	11.07	12.47
Max credible interval (RCIR)	0.1266	0.08200	0.04504
AUC-ROC (RCIR)	0.715179	0.773388	0.760740
MSE (RCIR)	0.0433455	0.0644942	0.0678055

80% reduction in credible interval width. Again this happens without sacrificing accuracy. For AUC-ROC the reduction is always below 0.01% and for the first two data set MSE increases also less than 2%. For the third one for which we have the narrowest credible intervals, AUC-ROC actually improves slightly but MSE worsens slightly more, almost 7%.

6 Conclusion

In many applications classification is just the starting point. The actual goal is to make a decision e.g. on whether to provide credit to an applicant, whether to address a potential buyer with marketing activities, what investment instruments to include in a portfolio, or which treatments to prescribe to a patient. While the need for well-calibrated probabilities for making justified decisions has been clearly identified [20, 6] and several practical methods for calibrating arbitrary classifiers are available, the literature has been ignoring the reliability of the estimates. The existing methods calibrate the outputs well on training data, but

do not have guarantees on performance on test data and often have very large uncertainty.

In this work we showed how the reliability of the calibration can be analysed using simple statistical analysis and demonstrated it in the context of isotonic regression [21]. We further extended isotonic regression to provide estimates with probabilistic guarantees on reliability. As isotonic regression is one of the most widely applied calibration methods, our results have clear practical value. The formulation has potential also for follow-up work on extending other calibration methods to better account for reliability, even though rigorous statistical analysis will not be trivial for e.g. the ensemble methods [13, 12].

The main conclusion of our work is clear. By explicitly measuring the reliability of calibration and directly optimizing for it, we can improve reliability without notably affecting accuracy. In most cases the deterioration in accuracy metrics was less than 1% while the width of the credible intervals was reduced by 80-90%. Our method guarantees that downstream applications that use individual predictions for further processing can be sufficiently certain that the estimate will be close to the true probability. This was achieved with a deterministic algorithm that terminates after a finite number of steps, demonstrated in practice to perform near optimally even though we lack formal guarantees of optimality.

Acknowledgements

This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI).

References

1. Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E.: An Empirical Distribution Function for Sampling with Incomplete Information. The Annals of Mathematical Statistics **26**(4), 641–647 (1955)
2. de Bruijne, M.: Machine learning approaches in medical image analysis: From detection to diagnosis. Medical Image Analysis **33**(5) (2016)
3. Diemert, E., Betlei, A., Renaudin, C., Massih-Reza, A.: A Large Scale Benchmark for Uplift Modeling. Proceedings of the AdKDD and TargetAd Workshop, KDD (2018)
4. Fawcett, T., Niculescu-Mizil, A.: PAV and the ROC convex hull. Machine Learning **68**(1), 97–106 (2007)
5. Gelman, A., Carlin, J., Stern, H.S., Rubin, D.B.: Bayesian data analysis. Chapman & Hall (2004)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney (2017)
7. Kumar, A., Liang, P., Ma, T.: Verified uncertainty calibration. In: Advances in Neural Information Processing. No. NeurIPS (2019)

8. Louzada, F., Ara, A., Fernandes, G.B.: Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science **21**(2), 117–134 (2016)
9. McMahan, H.B., Golovin, D., Chikkerur, S., Liu, D., Wattenberg, M., Hrafnkelsson, A.M., Boulos, T., Kubica, J., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E.: Ad click prediction: a View from the Trenches. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 1222–1230 (2013)
10. Murphy, A.H., Winkler, R.L.: Reliability of Subjective Probability Forecasts of Precipitation and Temperature. Source Journal of the Royal Statistical Society. Series C (Applied Statistics) **26**(1), 41–47 (1977)
11. Murphy, K.: Machine learning: a probabilistic perspective. The MIT Press (2011)
12. Naeini, M.P., Cooper, G.F.: Binary Classifier Calibration using an Ensemble of Near Isotonic Regression Models. In: 2016 IEEE 16th International Conference on Data Mining. pp. 360–369 (2016)
13. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the 29th AAAI Conference on Artificial Intelligence pp. 2901–2907 (2015)
14. Niculescu-Mizil, A., Caruana, R.: Obtaining Calibrated Probabilities from Boosting. In: Proceedings of Uncertainty in Artificial Intelligence. pp. 413–420 (2005)
15. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning ICML 05. pp. 625–632. No. 1999 (2005)
16. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
17. Provost, F., Fawcett, T.: Robust Classification For Imprecise Environments. In: Machine Learning. vol. 42, pp. 203–231 (2001)
18. Rubinstein, M.: Markowitz's "Portfolio Selection": A Fifty-Year Retrospective. The Journal of Finance **57**(3), 1041–1045 (2002)
19. Shmueli-Scheuer, M., Roitman, H., Carmel, D., Mass, Y., Konopnicki, D.: Extracting user profiles from large scale data. Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud pp. 1–6 (2010)
20. Van Calster, B., McLernon, D.J., Van Smeden, M., Wynants, L., Steyerberg, E.W., Bossuyt, P., Collins, G.S., MacAskill, P., Moons, K.G., Vickers, A.J.: Calibration: The Achilles heel of predictive analytics. BMC Medicine **17**(1), 1–7 (2019)
21. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. Proceedings of International Conference on Machine Learning pp. 1–8 (2001)
22. Zadrozny, B., Elkan, C.: Transforming Classifier Scores into Accurate Multiclass Probability Estimates. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining pp. 694–699 (2002)