Reliable Calibrated Probability Estimation in Classification

Marinka Zitnik*

***Abstract*—Estimating reliable class membership probabilities is of vital importance for many applications in data mining in which classification results are combined with other sources of information to produce decisions. Other sources include domain knowledge, outputs of other classifiers or example-dependent misclassification costs. In this paper we revisit the problem of classification calibration motivated by the issues of the isotonic regression and binning calibration. These methods can behave badly on small or noisy calibration sets, producing inappropriate intervals or boundary generalization. We propose an improvement of the calibration with isotonic regression and binning method by using bootstrapping technique, named boot-isotonic regression and boot-binning, respectively. Confidence intervals obtained by repeatedly calibrating the set sampled with replacement from the original training set are used for merging unreliable or too narrow calibration intervals. This method has been experimentally evaluated with respect to two calibration measures, several classification methods and several problem domains. The results show that the new method outperforms the basic isotonic regression and binning methods in most configurations.**

***Index Terms*—probabilistic calibration for classification, bootstrapping, isotonic regression, binning, quality measures, confidence interval.**

I. Introduction

IT is desirable that a classification method produces membership values that reflect the assessment uncertainty of an observation belonging to a particular class. The aim of calibration is to transform or calibrate unnormalized scores into the probability space to give a reliable and realistic impression of the uncertainty assessment. A comparative study [1] revealed that isotonic regression and Platt calibration [13] are most effective general probabilistic calibration techniques. However, simple binning averaging method [3] often behaves well in transforming the original uncalibrated estimated probabilities.

Loosely speaking we can think of a calibration function as a probabilistic scaler. This works by letting the numerical base classifier predict the examples in the calibration set and then fitting a scaling function that maps the predicted values to probabilities in a way that is optimal with respect to the true classes. Special care must be taken to avoid overfitting the training set and usually predictions used as input for the calibration step are joined in calibration set, separated from the training examples used by classifiers.

Both isotonic regression and binning have many problems when there is not enough data or provided data is noisy. The general problem of the scalers is that the small number of outliers may receive extreme values of the decision function, which gives them the highest influence on the form of the scaling function. Specifically in isotonic regression and binning, the errors in some examples distribute to all examples in the corresponding section, but since there are only a few of them and sections are small, the final calibration returned by isotonic regression or binning method can get greatly influenced by few outliers.

This paper is organized as follows, in section II, some of the most known calibration methods and evaluation measures are reviewed. Next, section III presents our improvement of calibration employing bootstrapping technique for merging unreliable intervals returned by isotonic regression and binning method. An experimental evaluation of proposed methods with respect to various classifiers, problem domains and quality measures is included in section IV. Finally, section V provides conclusions and points out future work.

II. Related Work

Many calibration techniques have been proposed in the literature [9], including softmax scaling, calibration of Gaussian modelling of the decision function, beta scaling, Platt calibration, isotonic regression and binning. The two latter methods are investigated in this paper. An empirical study [1] reviewed the dependency between optimal calibration strategy and learning algorithm. While Platt calibration is suitable for maximum margin methods [11] such as SVMs or boosting, isotonic regression performs consistently well for large data sets, but suffers from overfitting on small data sets.

Estimating probabilities and confidence values is crucial in many real applications. If probabilities are accurately predicted, decisions with good assessment of risks and costs can be made.

Calibrated probability estimates are needed when a cost sensitive decision must be made about each example in a data set, and the costs associated with different examples are different [3]. Additionally, the integration of these techniques with other models such as multi classifiers or with previous knowledge is more robust [7].

In real problems when we have to make several related decisions at a time, the joined best local decisions do not always give the best global result [10]. In this configuration each local model is a data mining model which accompanies its predictions with probabilities. Having single problem unrealistic probabilities will probably not matter but combining several non realistic probabilistic models can make the overall model diverge, as bad local estimations distort the global result.

*M. Zitnik is with the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, Ljubljana, Trzaska 25. Contact: marinka.zitnik@student.uni-lj.si.

Manuscript received January 08, 2012; revised January 11, 2012.

One solution to presented problems above is calibration.

In classification probabilities can represent degrees of confidence, especially in binary classification, thus accompany each prediction with a score of reliability. For this reason, instead of redesigning existing methods to obtain good probabilities, several calibration techniques have been developed. Generally, a calibration technique is a postprocessing technique aimed at improving any existing machine learning method: decision trees, kernel methods, neural networks, instance-based methods, Bayesian methods etc.

We use the following notation. Given a data set T, it contains n examples noted as x_i and c denotes the number of classes. An actual probability of example x_i belonging to class j is represented by $f_{i,j} = P(C = c_j|x_i)$ and is strictly not a probability but an indicator function, p_j denotes the prior probability of class j. The estimated probability of example x_i to be of class j, given a particular classifier, is $p_{i,j} = P_{est}(C = c_j|x_i)$ and calibrated probability is noted as $\widehat{f_{i,j}} = P_{cal}(C = c_j|x_i)$. In terms of calibration univariate means that the set of possible classes consists only of two classes, $c = 2$, which simplifies our notation.

A. Univariate Calibration Methods

Different calibration approaches can be applied depending on the task [10]. The most common calibration approaches are summarized in table I. However in this paper we focus on classification problems and specifically on type of probability calibration.

TABLE I
A TAXONOMY OF CALIBRATION PROBLEMS.

Task	Addressed problem	Calibration target
Classification	Expected class distribution is different from real class distribution	prediction
Classification	Expected probability of correct guesses differs from real proportion	probability
Regression	Expected output differs from real average output	prediction
Regression	Expected density functions are too narrow or broad	probability

This section surveys different approaches to provide calibrated probability estimates, these include (i) simple normalization, (ii) a function mapping from the membership values to calibrated probability, (iii) derivation using Bayes rule, and (iv) optimization of the Beta distributed random variable to obtain calibrated values.

1) Simple Normalization: Normalized membership values do not cover the assessment uncertainty of an observation belonging to a class, therefore these methods have no probabilistic background and are only used for pre-calibration [12] – after normalization scores meet the mathematical requirements for a probability. Boundary values such as in regularization of unnormalized scores are transformed to boundary membership probabilities with $p_{i,j} = 0.5$. More sophisticated normalizations are E-calibration and its special variant softmax calibration [9], [10]. Nevertheless these are just (non) linear transformations and should not be used for determination of membership probabilities.

2) Calibration Using Mapping: Univariate calibration with mapping is the search for a function which maps a membership value p_i to a calibrated conditional probability for positive class $\widehat{f}_i$. The mapping function is learned with one of regression techniques, such as logistic, piecewise logistic and isotonic regression.

Binning [5] is the simplest method for calibration that uses mapping function. It works by first ordering the examples by increasing values of predicted probabilities. Then, the ordered examples are divided into a pre-defined number of segments, each of which contains a constant number of examples. These segments are called the bins. For each segment s , its lower and upper bounds are determined and mean predicted probability of the examples that fall into s is estimated, the latter is the corrected probability estimate returned by binning method [3].

Binning is a parametric calibration method as the number of bins must be specified in advance. Its main weakness is the fixed number of examples in each formed segment [10]. It might happen that the algorithm averages probabilities of examples which should be in different segments and thus the error increases for both segments; in the first segment due to examples with higher probabilities and the second segment lacks examples that have been attributed to the first segment.

Platt presents a parametric approach for fitting a sigmoid that maps estimated probabilities into calibrated ones; this method is known as Platt scaling or Platt logistic regression [13]. Accepted method for the calibration of membership values is to model the log odds of the conditional probabilities

$$g(i) = \log \frac{\widehat{f}_i}{1 - \widehat{f}_i}$$

as a (linear) function g of the membership values for the positive class, which leads to the derivation of the calibrated conditional probability

$$\widehat{f}_i = \frac{1}{1 + e^{-g(i)}}$$

Platt selected g to be a linear function $g(i) = A \cdot p_i + B$ with scalar parameters A and B. Using this linear function the calibration function gets typical sigmoidal shape. The estimators $\hat{A}$ and $\hat{B}$ can be found with the optimization procedure model-trust algorithm [13].

Platt's logistic regression has been extended by applying piecewise instead of full logistic regression [20]. Different from Platt's model is that log odds are not regarded as a linear function of membership values but instead as a piecewise linear function with four knots separating membership values into three areas; (i) obvious decision for negative class, (ii) hard to classify and (iii) obvious decision for positive class. The log odds are fitted separately as a linear function in each of these three areas.

Another method that uses mapping function for calibration is isotonic regression. Isotonic regression [12] relies on the monotonicity assumption, meaning the true conditional class

probability f_i is an isotonic (monotonically increasing) function of the values of the learners decision function. This is a non parametric method which leads to a stepwise constant mapping function. Isotonic regression applies the basic model

$$\widehat{f}_i = g(i) + \epsilon_i,$$

where g is an isotonic function and ϵ_i an individual error term. A non-decreasing mapping function $\hat{g}$ can be found given a training set with learned membership values p_i and binary class labels f_i so that

$$\hat{g} = arg\min_h \sum_{i=1}^{n} (f_i - h(p_i))^2$$

holds. Pair-adjacent violators (PAV) algorithm is used to fit the training set according to this mean square error criterion. It has been shown [8] that isotonic regression based calibration using PAV algorithm is equivalent to the ROC convex hull method and instead of using isotonic regression convex hull computation can be used.

Platt scaling is effective when there are sigmoid-shaped distortions in the predicted probabilities [1]. Isotonic regression is more powerful and can correct any monotonic distortions. Unfortunately, sometimes this is not a benefit. A learning curve analysis shows that isotonic regression is prone to overfitting when data is scarce and can perform worse than Platt scaling [1].

3) Calibration via Bayes Rule: It consists of two steps to supply membership probabilities [6]. At first the positive unnormalized class scores are split into two groups according to their true class. Secondly, membership probabilities are determined by application of Bayes theorem to class conditional probabilities and class priors. The latter can be easily estimated from the training set, but the choice of the distribution type for the class conditional probabilities is cruical. Standard assumption is Gaussian or asymmetric Laplacian distribution [9].

4) Calibration Using Assignment Values: Calibration method proposed in [6] splits the unnormalized scores for a chosen class, here instead the membership values are partitioned according to their assignment [21]. Membership values are modelled separately in each partition as Beta distributed random variables.

Methods listed in this subsection are designed for binary classification and for most of them it is not trivial to extend them to multi class problems.

Multivariate extensions are introduced in [9] including the standard procedure of reduction to binary classification tasks with subsequent pairwise coupling and calibrator based on the Dirichlet distribution. With Dirichlet calibration method the output of the binary reduction algorithms are first transformed into Beta distributed random variables and then combined to realizations of a Dirichlet distributed random vector.

B. Calibration Quality Measures

Calibration is defined as the degree of the approximation of the predicted probabilities to the actual probabilities. More precisely, a perfectly calibrated classifier is the one, for which the following holds: for a sample of examples with predicted probability p the expected proportion of positives is close to p. Evaluation measures are extensively surveyed in [14].

Commonly used measures include: (i) mean squared error (MSE), (ii) log-loss [18], (iii) calibration by overlapping bins [16], (iv) calibration loss [19], (v) H-hat and (vi) chi-squared test through Hosmer-Lemeshow C-hat [10].

Calibration can clarify the relationship between ROC analysis and the AUC based measures, [8] and [19]. A perfectly calibrated classifier always gives a convex ROC curve. Although a classifier can produce very good rankings and high AUC, the estimated probabilities might differ from the actual probabilities. Calibration loss for class j is defined as the mean squared deviation from empirical probabilities that are derived from the slope of ROC segments

$$CalLoss(j) = \sum_{b=1}^{r_j} \sum_{i \in s_{j,b}} (\widehat{f_{i,j}} - \sum_{i \in s_{j,b}} \frac{f_{i,j}}{|s_{j,b}|}),$$

where r_j is the number of segments in the ROC curve for class j, i.e. the number of different estimated probabilities for class j : $|\widehat{f_{i,j}}|$. Each ROC segment is denoted by $s_{j,b}$ with $b \in 1, 2, \ldots r_j$ and defined as

$$s_{j,b} = \{i \in 1, 2, \ldots n | \forall k \in 1, \ldots n \ : \ \widehat{f_{i,j}} \geq \widehat{f_{k,j}}, i \notin s_{j,d}, \forall d < b\}.$$

In this paper we use mean squared error or Brier score, log-loss and calibration by overlapping binning. Although Brier score was originally not a calibration measure it was decomposed in terms of calibration and refinement loss [17], [19]. An idea of decomposition is that data is organized into bins and the observed probability in each bins is compared to the predicted or global probability. Log-loss is Kullback-Leibler divergence between real and inferred model and is defined as

$$LogLoss = \sum_{j=1}^{c} \sum_{i=1}^{n} \frac{f_{i,j} \log \widehat{f_{i,j}}}{n}.$$

Similarly to decomposition of mean squared error calibration by overlapping bins [16] consists of splitting the test set into several bins. The problem that arises here is the determination of the proper number of used bins, partial solution is to make the bins overlap. Formally, calibration by overlapping bins measure for class j is

$$CalBin(j) = \frac{1}{n-s} \sum_{b=1}^{n-s} \sum_{i^*=b}^{b+s-1} |\widehat{f_{i^*,j}} - \frac{\sum_{i^*=b}^{b+s-1} f_{i^*,j}}{s}|,$$

where s is the bin length and i^* are the new indices of the examples ordered by $\widehat{f_{i,j}}$.

III. Calibration using bootstrapping for merging intervals, boot-binning and boot-isotonic regression

As mentioned in the previous sections, binning method and isotonic regression are often used calibration methods. The latter, although optimal in the sense of minimum square deviations gives poor results on small and noisy calibration

sets which is often shown as too many or too few bins and poor generalization on the edges of bins.

The core of our approach is to use bootstrapping sampling technique to compensate for distortions when the calibration set is insufficient for straightforward calibration.

Based on this reasoning we have improved isotonic regression and binning calibration method by using bootstrapping statistical technique and we refer to our proposed methods as boot-isotonic regression and boot-binning, respectively.

We have split the method into four stages.

In the first stage a new data set is constructed which is obtained by random sampling with replacement from the original data set. This bootstrapped sample which is of the same size as the original data set typically consists of only 63% different original examples and 37% examples are not occurring in the sample. Using this information, a new set is constructed; sampled examples are used as a training set and the remaining 37% examples of original train set as a calibration set.

In the second stage a prediction model is devised on the newly formed training set with a classification method of user's choice. Based on the built classifier's predictions for examples in the calibration set, a procedure for standard binning method or isotonic regression is called, returning calibrated probabilities.

Third stage consists of repeating first and second stage r times. In addition calibrated probabilities are recorded in each repetition.

In the fourth stage the final calibration is performed which combines unreliable and too narrow bins, possibly returned by calibration methods. Using non parametric Wilcoxon signed-rank test a non parametric confidence interval and a pseudo-median estimate is obtained for each example x_i of the data set. This is possible due to recorded calibrated probabilities obtained from third stage.

Based on confidence intervals and pseudomedians we construct final bins by continually merging two sets of intervals obtained from two repetitions of second stage. Individual bin is discarded if it does not contain a minimum number of confidence intervals T_{count} – a confidence interval of example x_i is said to be contained in a bin if a bin intersects a specified fraction T_{in} of it. Probabilities are merged accordingly.

Boot-binning and boot-isotonic regression are parametric methods. Specifically, it must be specified: (i) r, the number of iterations in third stage; (ii) T_{count}, confidence interval threshold; and (iii) T_{in}, intersection between confidence interval and an individual bin.

IV. Results

A. Experimental Configuration

Our main hypothesis is that proposed methods, boot-isotonic regression and boot-binning, significantly improve calibration results obtained with isotonic regression and binning, respectively.

The proposed methods have been evaluated in different settings, on a total of 10 data sets from the well known UCI machine learning repository [15], these are listed in table II. The chosen data sets cover a wide range of complexity, dimensionality and various types of attributes. In all data sets nominal attributes have been dichotomized. Multi class data sets have been converted to binary tasks by joining several smaller classes to one or selecting two largest classes. Type of conversion of multi class data sets is noted in table II.

Five different methods for classification have been tested (i) decision trees (DT), (ii) support vector machines (SVM), (iii) random forests (RF), (iv) boosting (Boost), and (v) naive Bayes (NB).

The calibration methods used in the experiments are (i) binning (Binn), (ii) isotonic regression (IR), (iii) boot-isotonic regression (B-IR) and (iv) boot-binning (B-Binn). Firstly, we compare classification methods without calibration (Base). It is understandable that calibration would not have had any sense if the results had not been improved. Calibration is evaluated with the CalBin and MSE quality measures.

Methods are evaluated in a setting using training, calibration and test set. This is because calibration methods can use or not an additional data set for calibration. Each data set is first split into two different subsets: the training set and the test set with the sizes of 80% and 20% of the examples, respectively. This splitting is done uniformly at random before the testing starts and all methods use the same division of particular data set.

Boot-binning and boot-isotonic regression internally implement splitting the train set to obtain calibration set using bootstrapping as described in section III. That is, in tests with boot-binning and boot-isotonic regression we provide the train set as input to these techniques.

In tests with binning and isotonic regression calibration set is obtained by randomly choosing 30 % from the train set. Newly formed training set is used to train a classifier, probabilities of the model are calibrated on the calibration set and corrected results are tested on the test set.

TABLE II
DATA SETS USED IN THE EXPERIMENTS. SIZE, NUMBER OF NOMINAL AND NUMERIC ATTRIBUTES. IN CASE A DATASET REPRESENTS A MULTI CLASSIFICATION PROBLEM, A CLASS USED IN TESTS IS LISTED.

No.	Data set	Size	Nom.	Num.	Class
1	Monks1	556	6	0	binary
2	Mushroom	8124	22	0	binary
3	Adult	32561	8	6	binary
4	House voting	435	16	0	binary
5	Tic-Tac	958	8	0	binary
6	Marketing	8993	13	0	"1" vs. rest
7	Yeast	1484	0	8	"NUC" vs. "CYT"
8	Breast Cancer	286	9	0	binary
9	Ionosphere	351	0	34	binary
10	Breast cancer Wisc.	699	0	9	binary

Each test consists of first selecting a data set on which to perform test, classification method and calibration technique. We then perform 30 repetitions for each such combination, that is for 10 data sets, 5 classification methods and 5 calibration approaches a total of 7500 ($7500 = 30 \cdot 10 \cdot 5 \cdot 5$) tests. For each test the calibration results on the test data set are evaluated

with respect to CalBin and MSE quality measure. Final CalBin and MSE measures for a data set–classifier–calibration choice are obtained by averaging CalBin and MSE measures over 30 repetitions.

Table III lists classification algorithms' configurations used in experimental evaluation.

Table IV lists calibration techniques' configurations used in experimental evaluation.

TABLE III
CLASSIFICATION ALGORITHMS' CONFIGURATIONS USED IN EXPERIMENTAL EVALUATION.

Method	Configuration
DT	Gini Index attribute selection; min. instances in leaves = 2; pruning with m-estimate, $m = 2$
SVM	C-SVM; RBF kernel; tolerance, $p = 0.5$; precision $\epsilon = 0.001$; data normalization
RF	100 trees in forest; growth control, min. $g = 5$ examples
Boost	Classification tree weak classifier; number of created classifiers $n_b = 5$
NB	Laplace succession for prior distribution; m-estimate with $m = 2$ for conditional probability

TABLE IV
CALIBRATION TECHNIQUES' CONFIGURATIONS USED IN EXPERIMENTAL EVALUATION.

Method	Configuration
IR	non parametric
Binn	$n_{bin} = 0.07\|Set\|, 0.1\|Set\|, 0.15\|Set\|, 0.18\|Set\|, 0.2\|Set\|$ (5 settings, each used in 6 repetitions)
B-IR	T_{in} = 0.7, 0.5, 0.4, 0.3, 0.1, $r = 20$, T_{count} = $0.1\|Set\|$, $0.2\|Set\|$, $0.25\|Set\|$, $0.3\|Set\|$, $0.4\|Set\|$; (5 settings, each used in 6 repetitions)
B-Binn	T_{in} = 0.7, 0.5, 0.4, 0.3, 0.1, $r = 20$, T_{count} = $0.1\|Set\|$, $0.2\|Set\|$, $0.25\|Set\|$, $0.3\|Set\|$, $0.4\|Set\|$; number of bins, $n_{bin} = 0.07\|Set\|, 0.1\|Set\|, 0.15\|Set\|, 0.18\|Set\|, 0.2\|Set\|$ (5 settings, each used in 6 repetitions)

Evaluation is performed within R environment on a 2.0 GHz Intel Core i7 processor with 8 GB RAM. Table V presents average execution time of calibration over data sets and classifiers based on configurations in table IV. Boot-binning and boot-isotonic regression are computationally more intense and have higher time complexity than binning and isotonic regression. This is expected behaviour because calibration technique using bootstrapping works by repeatedly performing basic calibration (binning or isotonic regression). The number of repetitions depends on the parameter r.

The execution time to obtain final calibration using bootstrapping is approximately r-times larger than execution time for basis calibration. Some overhead is imputed due to calculation of confidence intervals and merging process of bins. As we set the parameter r to a value of $r = 20$ in our tests, there is a factor of 20 increase in time for calibration with bootstrapping.

B. Experimental Evaluation

Table VI shows calibration results with respect to the Brier (MSE) measure for all combinations of tested classifiers, calibration methods and data sets. Results are obtained using configuration described in previous section.

TABLE V
EXECUTION TIME FOR CALIBRATION APPROACHES AVERAGED OVER DATA SETS AND CLASSIFICATION METHOD.

Calibration method	Time [s]
Binning	0.6
Isotonic regression	1.3
Boot-binning	15.4
Boot-isotonic regression	32.1

In table VII we show calibration results with respect to the CalBin measure. Again, results are obtained using configuration described in previous section.

It is important to remark that general comparisons between methods are done in equal conditions. First of all we are comparing to classification methods without calibration estimating the relevance of calibration. The most interesting comparisons are those between binning and boot-binning method, because proposed boot-binning is implemented as an improvement of the original binning method. Analogously, we are interested in comparison between isotonic regression and boot-isotonic regression.

Differences exist when calibration results are evaluated with CalBin measure. This arises from the different nature of the measures. CalBin namely evaluates only calibration and MSE also evaluates other components.

Table VI offers some observations which confirm our previous knowledge about classification methods, their properties and calibration techniques. Specifically, it is known that isotonic regression is more powerful calibration method ([1], [5]) than simple binning averaging. This explains why both binning and boot-binning methods are in general outperformed by isotonic regression and boot-isotonic regression.

Further, there are substantial differences between predictions made by different learning algorithms and true posterior probabilities.

Empirical comparisons in [1] have shown that maximum margin methods such as boosting and SVMs yield characteristic distortions in their predictions, namely they push probability mass away from zero and one and have sigmoid shaped distortions in probability.

Other methods such as naive Bayes have the opposite bias and tend to push predictions closer to extreme values. Analysis in [1] suggests that random forests (among some others such as neural nets and bagged decision trees) are the best learning methods for predicting well calibrated probabilities prior to calibration. After calibration the best methods are SVMs and random forests. Concluding, results in table VI confirm results presented in [1].

Figure 1 shows an example how calibration transforms predictions of decision trees.

In experiments (see data set 10 in table VI) where basic calibration methods did not improve probability estimates, that is calibration with basic methods of binning and isotonic regression increased the MSE measure, our proposed methods partially correct this anomaly. Boot-isotonic regression not

TABLE VI
RESULTS GROUPED BY DATA SET WITH MSE EVALUATION MEASURE. INDIVIDUAL ENTRY IS AVERAGE MSE MEASURE FOR A DATA SET–CLASSIFIER–CALIBRATION CHOICE, OBTAINED BY AVERAGING MSE MEASURE OVER 30 REPETITIONS. DATA SETS ARE NUMBERED AS REFERRED IN TABLE II.

Data.	Class.	Base	IR	Binn	B-IR	B-Binn
1	DT	0.3901	0.2942	0.3109	0.2764	0.2817
	SVM	0.3265	0.2142	0.2453	0.2100	0.2265
	RF	0.2442	0.2104	0.2313	0.2019	0.2210
	Boost	0.3425	0.2731	0.2771	0.2556	0.2633
	NB	0.4099	0.2497	0.1290	0.2249	0.1250
2	DT	0.2201	0.1302	0.1402	0.1201	0.1197
	SVM	0.1422	0.0561	0.0726	0.0512	0.0502
	RF	0.1231	0.0729	0.0827	0.0702	0.0725
	Boost	0.1621	0.0921	0.0824	0.0978	0.0912
	NB	0.2330	0.0442	0.0419	0.0323	0.0371
3	DT	0.1096	0.1054	0.1190	0.1043	0.1095
	SVM	0.1203	0.1187	0.1164	0.1181	0.1130
	RF	0.1010	0.0989	0.0831	0.0853	0.0821
	Boost	0.1320	0.1264	0.1245	0.1151	0.1258
	NB	0.1407	0.1115	0.1146	0.1112	0.1113
4	DT	0.4520	0.3011	0.2400	0.2391	0.2039
	SVM	0.3812	0.1301	0.1622	0.1232	0.1376
	RF	0.2814	0.1262	0.1721	0.1231	0.1561
	Boost	0.2214	0.1923	0.2011	0.1703	0.1938
	NB	0.4198	0.2400	0.1647	0.2203	0.1660
5	DT	0.3139	0.1847	0.1982	0.1911	0.1889
	SVM	0.2741	0.1396	0.1436	0.1348	0.1374
	RF	0.2645	0.1516	0.1747	0.1485	0.1662
	Boost	0.2401	0.1390	0.1931	0.1452	0.1220
	NB	0.4410	0.2271	0.1873	0.2142	0.1662
6	DT	0.2641	0.1938	0.1831	0.1737	0.1798
	SVM	0.2213	0.1542	0.1648	0.1524	0.1520
	RF	0.2301	0.1991	0.2201	0.1963	0.2193
	Boost	0.2170	0.1672	0.1801	0.1689	0.1765
	NB	0.3422	0.2913	0.2999	0.2819	0.2716
7	DT	0.1876	0.1568	0.1348	0.1474	0.1322
	SVM	0.1752	0.1253	0.1306	0.1211	0.1275
	RF	0.1630	0.1492	0.1573	0.1476	0.1603
	Boost	0.1502	0.1412	0.1500	0.1375	0.1355
	NB	0.1916	0.1362	0.1342	0.1354	0.1353
8	DT	0.2032	0.1773	0.1814	0.1625	0.1765
	SVM	0.1988	0.1567	0.1623	0.1521	0.1566
	RF	0.1906	0.1736	0.1625	0.1702	0.1526
	Boost	0.2749	0.1790	0.2305	0.1702	0.2249
	NB	0.2449	0.1694	0.1997	0.1523	0.1996
9	DT	0.2561	0.2019	0.2103	0.2063	0.2051
	SVM	0.1920	0.2013	0.2109	0.2054	0.2119
	RF	0.1981	0.1887	0.2013	0.1872	0.1910
	Boost	0.2301	0.1998	0.2100	0.1849	0.1991
	NB	0.3371	0.2374	0.3057	0.2103	0.2802
10	DT	0.0376	0.0353	0.0496	0.0333	0.0450
	SVM	0.0320	0.0311	0.0399	0.0299	0.0401
	RF	0.0310	0.0301	0.0311	0.0300	0.0305
	Boost	0.0561	0.0671	0.0961	0.0501	0.0831
	NB	0.0283	0.0331	0.0313	0.0258	0.0290

TABLE VII
RESULTS GROUPED BY DATA SET WITH CALBIN EVALUATION MEASURE. INDIVIDUAL ENTRY IS AVERAGE CALBIN MEASURE FOR A DATA SET–CLASSIFIER–CALIBRATION CHOICE, OBTAINED BY AVERAGING CALBIN MEASURE OVER 30 REPETITIONS. DATA SETS ARE NUMBERED AS REFERRED IN TABLE II.

Data.	Class.	IR	Binn	B-IR	B-Binn
1	DT	0.2642	0.2119	0.2261	0.2713
	SVM	0.1972	0.2151	0.19062	0.1622
	RF	0.2023	0.2003	0.1729	0.1630
	Boost	0.2651	0.2719	0.2102	0.2533
	NB	0.2019	0.1391	0.1931	0.1242
2	DT	0.1289	0.1382	0.1032	0.1207
	SVM	0.0761	0.0526	0.0413	0.0422
	RF	0.0739	0.0757	0.0702	0.0625
	Boost	0.0572	0.1012	0.0478	0.1023
	NB	0.0433	0.0591	0.0403	0.0451
3	DT	0.1113	0.1098	0.0961	0.1013
	SVM	0.1751	0.1196	0.1163	0.1042
	RF	0.0762	0.0853	0.0853	0.0912
	Boost	0.1232	0.1233	0.1151	0.0923
	NB	0.1204	0.1041	0.1091	0.11371
4	DT	0.2511	0.2242	0.2013	0.1924
	SVM	0.1242	0.1691	0.1366	0.1126
	RF	0.1532	0.1651	0.1231	0.0923
	Boost	0.1123	0.1931	0.1623	0.1382
	NB	0.1500	0.1647	0.1043	0.1242
5	DT	0.1357	0.1482	0.111	0.1371
	SVM	0.1261	0.1213	0.1112	0.1071
	RF	0.1152	0.1415	0.0971	0.1122
	Boost	0.1076	0.1633	0.1010	0.1098
	NB	0.1873	0.1671	0.1451	0.1261
6	DT	0.1529	0.1753	0.1044	0.1108
	SVM	0.1612	0.1347	0.1023	0.1403
	RF	0.1623	0.1901	0.1203	0.1873
	Boost	0.1561	0.1761	0.1302	0.1145
	NB	0.1698	0.1983	0.1438	0.1726
7	DT	0.1030	0.1428	0.0964	0.1224
	SVM	0.1201	0.1409	0.1017	0.1355
	RF	0.1302	0.1503	0.1216	0.1424
	Boost	0.1514	0.1603	0.1287	0.1599
	NB	0.1340	0.1420	0.1284	0.1356
8	DT	0.1483	0.1154	0.1215	0.1325
	SVM	0.1521	0.1539	0.1231	0.1234
	RF	0.1234	0.1324	0.1201	0.1299
	Boost	0.1680	0.1935	0.1405	0.1529
	NB	0.1414	0.1496	0.1353	0.1366
9	DT	0.1918	0.2134	0.1633	0.1682
	SVM	0.1833	0.2039	0.1634	0.1829
	RF	0.1865	0.2113	0.1761	0.1871
	Boost	0.1872	0.1973	0.1799	0.1811
	NB	0.2145	0.2737	0.2001	0.2228
10	DT	0.0321	0.0315	0.0298	0.0301
	SVM	0.0398	0.0403	0.0351	0.0400
	RF	0.0311	0.0362	0.0268	0.0321
	Boost	0.0521	0.0531	0.0469	0.0532
	NB	0.0561	0.0601	0.0551	0.0609

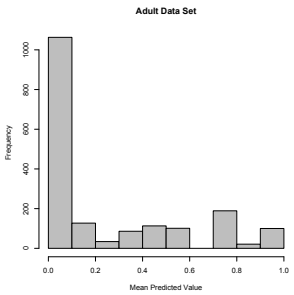
only did not increase the MSE measure but has also improved probability estimates returned by classifier. Based on empirical evaluation, boot-isotonic regression could also be used in settings where the benefits of calibration are not assured.

If we observe table VI, it is important to remark how our methods outperform significantly binning averaging and isotonic regression in terms of MSE. Following are the details of the statistical tests.

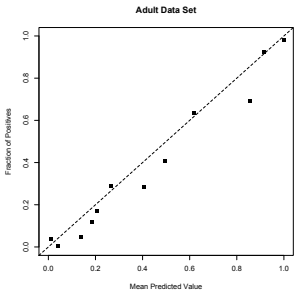
Further we make additional experiments, grouped by classification method and calibration technique, then we use suitable statistics to confirm the differences in the results are significant. We want to detect if the boot-binning (boot-isotonic regression) significantly improves binning averaging (isotonic regression) method.

Experiments are grouped because we want to avoid testing for significance that can be attributed to used classification method and calibration technique. Namely, we already know [1] that some learning methods predict well-calibrated probabilities prior to calibration and that isotonic regression is more powerful than binning averaging in general.

Based on this reasoning we use two-sided paired Wilcoxon signed-rank test for comparison of two classifiers over multiple



(a) Histogram of predicted values for decision trees with MDL splitting criteria.



(b) Reliability diagram obtained with boot-isotonic regression method, r = 20, $T_{in} = 0.7$, $T_{count} = 50$.

Fig. 1. Single decision tree has high variance, figure 3(a) shows histogram of predictions before calibration. After calibration reliability plot closely follows the diagonal line in figure 3(b).

data sets. This test is a non parametric alternative to the paired t-test, which ranks the differences in performances of two classifiers for each data set and compares the ranks for the positive and negative differences, ignoring the signs. Tables VIII and IX show p-values and 95% confidence intervals obtained by testing. For each classifier we are trying to reject the null hypothesis that binning (isotonic regression) and boot-binning (boot-isotonic regression) perform equally well. In that sense we form ten hypotheses, one for each combination of classifier and our proposed calibration method (boot-binning or boot-isotonic regression).

From results in table VIII can be seen, that the p-values are less than the significance level $\alpha = 0.05$ and we can reject the null hypotheses concerning binning approach. Boot-binning significantly outperforms binning calibration method for (i) decision trees, (ii) SVMs, (iii) random forests, (iv) boosting, and (v) naive Bayes.

From results in table IX can be seen, that boot-isotonic regression significantly outperforms isotonic regression for (i) decision trees, (ii) boosting, and (iii) naive Bayes. However we cannot reject the isotonic regression null hypotheses for SVMs and random forest.

Boot-binning and boot-isotonic regression are parametric methods and proper settings of parameters greatly influence the performance. Parameters include r, number of iterations; T_{count}, confidence interval threshold and T_{in}, intersection threshold between confidence interval and calibration bin.

Our empirical evaluation has shown that increasing the

TABLE VIII
P-VALUES FOR THE TWO-SIDED PAIRED WILCOXON SIGNED-RANK TEST AND CONFIDENCE INTERVALS AT SIGNIFICANCE LEVEL $\alpha = 0.05$.

	Classifier	Binning
Boot-binning	DT	0.00195; [0.0043, 0.0205]
	SVM	0.00977; [0.0026, 0.0158]
	RF	0.01898; [-0.0103, -0.0008]
	Boost	0.02734; [0.0021, 0.0349]
	NB	0.01953; [0.0010, 0.0158]

TABLE IX
P-VALUES FOR THE TWO-SIDED PAIRED WILCOXON SIGNED-RANK TEST AND 95% CONFIDENCE INTERVALS AT SIGNIFICANCE LEVEL $\alpha = 0.05$.

	Classifier	Isotonic regression
Boot-isotonic regression	DT	0.03711; [0.0015, 0.0288]
	SVM	0.06251; [0.0195, 0.0488]
	RF	0.05413; [0.0076, 0.0254]
	Boost	0.04409; [0.0010, 0.0329]
	NB	0.03091; [0.0019, 0.0291]

number of iterations improves the performance of calibration. Nevertheless a compromise must be taken as larger r increases the execution time. Value $r = 1$ performs basic calibration, a noticeable improvement of calibration results in terms of MSE and CalBin starts at r being around 10 and stabilizes at around 30.

The other two parameters depend largely on the characteristics of a data set. Usually, setting T_{in} below 0.5 gives poor results, that is, we require that at least half of confidence interval width is contained in the calibration bin. Best results are obtained for T_{in} being around 0.7, but this parameter can be fine tuned.

Value of confidence interval threshold should depend on the size of the calibration set and also influences the final number of bins returned by calibration method. Increasing T_{count} obstructs the merging of bins and the overall number of returned bins is larger. Setting this parameter to any value greater than $0.5 \cdot |S_{cal}|$ usually gives very poor results.

C. Anomalies Detection and Removal

In previous subsection we reject the null hypothesis of calibration using bootstrapping and bins' merging performs equally well as basic calibration of isotonic regression and binning. However, our main aim in this paper is to reduce problems that emerge using isotonic regression on small and noisy data sets. These include poor boundary generalizations and inappropriate intervals (bins) on small and noisy calibration sets.

In this section we demonstrate the performance of proposed method on an artificial data sets described in table X. We examine and compare calibration bins using reliability diagrams and calibration functions.

We test the behaviour of calibration functions in specific anomalies using data sets A1 and A2. The basis for these sets are the joint probability estimates generated with pseudo-random number generator and sorted in ascending order with corresponding class labels. Class labels are assigned in a way

that the share of positive class increases linearly with respect to the probability estimate. The positive class is marked as 1, the negative as 0. The data set contains 100 examples. Learnt calibration functions are tested on independent test set, which has the properties of data set A0.

In data sets D1 and D2 we test the response of calibration approaches to the gradual addition of noise in the data. Again, probability estimates are obtained by pseudo-random number generator. The positive class distribution is the same as in the A0 data set, classes are balanced and data sets D1 and D2 both contain 100 examples. The noise is added by changing class labels of the number of examples corresponding to the noise rate. Mapping functios learnt on data sets D1 and D2 are tested on independent set D0.

TABLE X
DESCRIPTION OF ARTIFICIAL DATA SETS WITH ANOMALIES (A1, A2) AND NOISE (D1, D2).

No.	Description
A0	class labels 0 are before class labels 1; transition is at probability estimate 0.5; linearly increase positive class share
A1	data set A0 with six class labels 1 at the beginning of the data set
A2	data set A0 with six class labels 1 around probability estimate 0.3
D0	linearly increase positive class share
D1	linearly increase positive class share with 15% of noise
D2	linearly increase positive class share with 25% of noise

Best results are obtained for parameters: $n_{bin} = 13$ for binning; $r = 15$, $T_{in} = 0.7$, $T_{count} = 15$ for boot-isotonic regression; and $r = 10$, $T_{in} = 0.6$, $T_{count} = 15$, $n_{bin} = 13$ for boot-binning. Figure 2 shows calibration plots for data sets with anomalies and figure 3 for data sets with imposed noise.

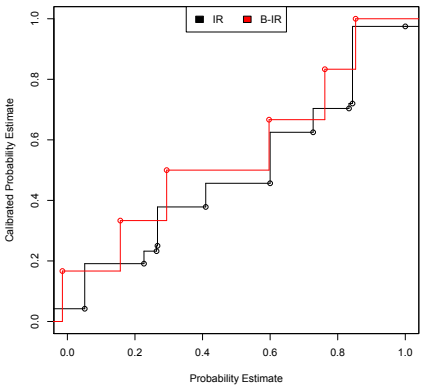
Table XI shows results achieved in terms of MSE measure. Boot-isotonic regression outperforms isotonic regression on both data sets with induced anomalies and noise. Boot-binning performs better than binning data sets with anomalies but gives weaker results than binning on data sets with noise.

TABLE XI
RESULTS ON DATA SETS A0–A2 AND D0–D2 IN TERMS OF MSE.

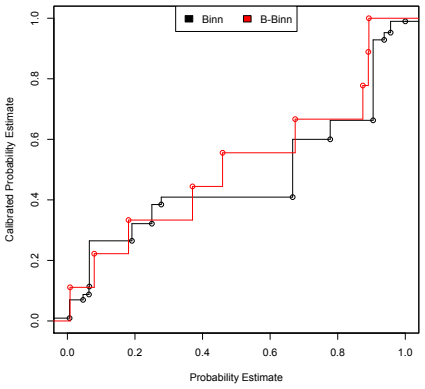
Data set	IR	Binn	B-IR	B-Binn
A0	0.2433	0.2501	0.2389	0.2410
A1	0.2481	0.2533	0.2391	0.2400
A2	0.2412	0.2481	0.2356	0.2378
D0	0.2319	0.2393	0.2148	0.2256
D1	0.2322	0.2331	0.2267	0.2357
D2	0.2491	0.2510	0.2451	0.2501

V. CONCLUSION AND FUTURE WORK

In this paper we discussed the problem of class probability distribution. We have proposed two methods, boot-isotonic regression and boot-binning method, both obtained by improving isotonic regression and binning calibration method, respectively. In this way the problems of poor calibration results on small or noisy calibration sets have been eliminated.



(a) Reliability diagram for isotonic regression (black) and boot-isotonic regression (red).



(b) Reliability diagram for binning (black) and boot-binning (red).

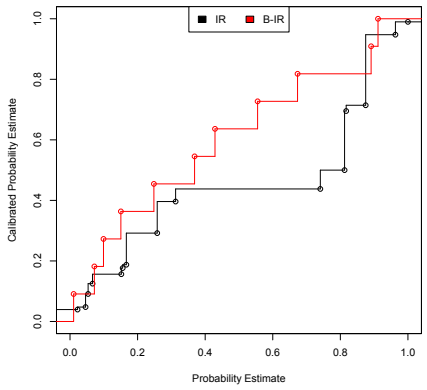
Fig. 2. Reliability diagram for isotonic regression and boot-isotonic regression on data set A1 in figure 3(a) Reliability diagram for binning and boot-binning on data set A1 in figure 3(b).

Using statistical tests we reject the null hypotheses, that binning (isotonic regression) and boot-binning (boot-isotonic regression) perform equally well for majority of classifiers. Calibration results show significant improvement using the bootstrapping approach. We are unable to reject the null hypotheses of equal performance between isotonic regression and its bootstrapping improvement for SVMs and random forests.

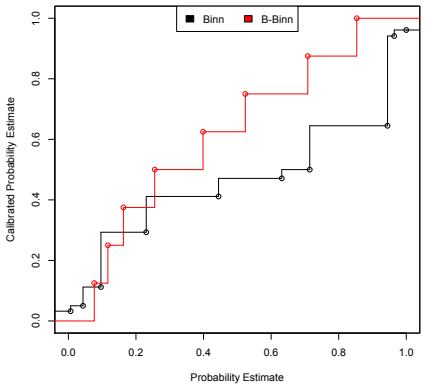
As future work we propose the analysis of the newly devised methods for multi class problems and comparison of our calibration method with other approaches. Our proposed methods are parametric and careful choice of parameters increases the overall complexity of the approach. In addition the idea of bootstrapping could be applied to various other calibration methods.

ACKNOWLEDGMENT

The present work benefited from the input of M. R. Sikonja, associate professor at The Faculty of Computer and Information Science at University of Ljubljana and member of



(a) Reliability diagram for isotonic regression (black) and boot-isotonic regression (red).



(b) Reliability diagram for binning (black) and boot-binning (red).

Fig. 3. Reliability diagram for isotonic regression and boot-isotonic regression on data set D1 in figure 3(a) Reliability diagram for binning and boot-binning on data set D1 in figure 3(b).

Laboratory for Cognitive Modelling, who provide valuable ideas to the writing summarised here.

REFERENCES

[1] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Machine Learning Conference*. ACM Press, 2005, pp. 625–632.

[2] I. Cohen and M. Goldszmidt. Properties and Benefits of Calibrated Classifiers. In 8th *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD*. Springer, 2004, pp. 125–136.

[3] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, 2001, pp. 609–616.

[4] F. Provost and P. Domingos. Tree Induction for Probability-based Ranking. *Machine Learning*, vol. 52, 2002, pp. 199–215.

[5] A. Bella, C. Ferri, J. Hernández-Orallo and M. J. Quintana. Similarity-binning averaging: a generalisation of binning calibration. In *Proceedings of the 10th International Conference on Intelligent Data Engineering and Automated Learning*, 2009, Springer, pp. 341–349.

[6] P. N. Bennett. Using asymmetric distributions to improve text classifier probability estimates: A comparison of new and standard parametric methods. *Technical Report CMU-CS-02-126*, Carnegie Mellon, School of Computer Science, 2002.

[7] S. Ruping. Robust Probabilistic Calibration. In *Proceedings of the 17th European Conference on Machine Learning*, 2006, Springer, pp. 743–750.

[8] T. Fawcett and A. Niculescu-Mizil. PAV and the ROC Convex Hull. *Machine Learning*, 2007, vol. 68/1, pp. 97–106.

[9] M. Gebel. *Multivariate calibration of classifier scores into the probability space*. Ph.D. dissertation, Faculty of Statistics, Dortmund Univ., Duisburg, Germany, 2009.

[10] A. Bella. *An Evaluation of Calibration Methods for Data Mining Models in Simulation Problems*. M. S. Thesis, Dept. Sys. Inf. Comp., University of Valencia, Spain, 2008.

[11] A. Niculescu-Mizil and R. Caruana. Obtaining Calibrated Probabilities from Boosting. In *Proceedings of International Conference on Uncertainty in Artificial Intelligence*, 2005, pp. 413–420.

[12] B Z.adrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (KDD '02). ACM, pp. 694–699.

[13] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A., Bartlett, P., Sch olkopf, B., Schuurmans, D., eds.: *Advances in Large Margin Classifiers*. MIT Press, 1999.

[14] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recogn. Lett.*, 30(1), 2009, pp. 27-38.

[15] A. Frank and A. Asuncion. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2010. Available: http://archive.ics.uci.edu/ml.

[16] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proc. of the 10th Intl. Conference on Knowledge Discovery and Data Mining*, 2004, pp. 69-78.

[17] A. H. Murphy. Scalar and vector partitions of the probability score: Part ii. n-state situation. *Journal of Applied Meteorology*, 11, 1972, 182-1192.

[18] D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. In 3rd Conf. on Maths and Computers in Sport, vol. 14, 1996, pp. 233-241.

[19] P. Flach and E. Matsubara. A simple lexicographic ranker and probability estimator. In *18th European Conference on Machine Learning*, Springer, 2007, pp. 575-582.

[20] Zhang, J. and Yang, Y. Probabilistic score estimation with piecewise logistic regression. In *Proceedings of the 21st International Conference on Machine Learning*, ACM Press, 2004.

[21] U. M. Garczarek and C. Weihs, C. Standardizing the comparison of partitions. *Computational Statistics*, 18 (1), 2003, 143–162.

Marinka Zitnik was born in Ljubljana, Slovenia, in 1989. Zitnik is currently a student in the final year of interdisciplinary university program of computer science and mathematics at The Faculty of Computer and Information Science Ljubljana and Faculty of Mathematics and Physics Ljubljana, University of Ljubljana.