IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 26, 2020, accepted July 5, 2020, date of publication July 9, 2020, date of current version July 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3008150

An Experimental Investigation of Calibration Techniques for Imbalanced Data

LANLAN HUANG[1], JUNKAI ZHAO[1], BING ZHU[1], (Member, IEEE), HAO CHEN[2], AND SEPPE VANDEN BROUCKE[3]

[1]Business School, Sichuan University, Chengdu 610064, China
[2]China Tobacco Guangxi Industrial Company, Ltd., Nanning 530001, China
[3]Department of Business Informatics and Operations Management, Universiteit Gent, 9000 Ghent, Belgium

Corresponding author: Bing Zhu (zhubing@scu.edu.cn)

This work was supported in part by the Sichuan University under Grant skqy-201742 and Grant 2019ZY-Business04, and in part by the Fundamental Research Funds for the Central Universities under Grant SCUBS-PY-202032.

ABSTRACT Calibration is a technique used to obtain accurate probability estimation for classification problems in real applications. Class imbalance can create considerable challenges in obtaining accurate probabilities for calibration methods. However, previous research has paid little attention to this issue. In this paper, we present an experimental investigation of some prevailing calibration methods in different imbalance scenarios. Several performance metrics are considered to evaluate different aspects of calibration performance. The experimental results show that the performance of different calibration techniques depends on the metrics and the degree of the imbalance ratio. Isotonic Regression has better overall performance on imbalanced datasets than parametric and other complex non-parametric methods. However, it performs unstably in highly imbalanced scenarios. This study provides some insights into calibration methods on imbalanced datasets, and it can be a reference for the future development of calibration methods in class imbalance scenarios.

INDEX TERMS Probability calibration, class imbalance, Isotonic regression.

I. INTRODUCTION

In many real-world classification applications, it is crucial to obtain accurate class probability estimation [1]. Class probability provides more detailed information than no-probability score or class label and supports the decision-making process more effectively. For example, ‘‘the customer has an 80% chance of churn’’ is more informative than just a class label of ‘‘churner’’, which can help the company find profitable customers to launchretention campaigns. Unfortunately, many popular classifiers such as support vector machine, boosted decision tree and even modern deep neural networks cannot produce accurate class probability estimations [2], [3]. To deal with this issue, two ways have been developed. The first way is to develop probabilistic models that are well calibrated. The other way is to use post-processing calibration method, which attempts to transform the output of classifiers to well-calibrated probabilities [4]. The probabilistic models need to redesign and optimize objective functions used in the classifier, which leads to high computational complexity and cost [5]. In contrast, the post-processing calibration method is simpler and can be used in combination with any type of classifier. Therefore, probability calibration has received considerable attention and many calibration approaches have been proposed in the past few years.

Class imbalance is a common issue in many real classification applications such as disease diagnosis [6] and credit scoring [7]. Class imbalance means more instances are labeled as certain classes while fewer instances are labeled as other classes. Addressing class imbalance can be a challenging issue [8]. Standard classification algorithms usually pursue high overall accuracy, thus tending to bias towards the majority classes [9], which yields biased output and results in high uncertainty for probability calibration [10]. Despite the strong demand for probability calibration in imbalance scenarios, there are few investigations on the performance of different calibration methods in the imbalance scenarios.

In this paper, we perform an experimental investigation of calibration techniques on imbalanced datasets. The contributions of our paper are three-fold. First, we launch an extensive comparative study of different calibration techniques across twenty-four binary imbalanced datasets. So far, no such

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

large-scale benchmark comparisons have been performed. Second, we use three different evaluation measures to provide different views on the performance of calibration techniques. Third, we compare the performance of different calibration methods in the highly imbalanced scenarios and the lowly imbalanced scenarios. The experimental results give some insights of calibration on imbalanced datasets and provide useful guidelines for the future development of calibration techniques.

The remainder of this paper is structured as follows. The related work of calibration techniques is provided in section II. In section III, we describe the experimental framework. In section IV, the experimental results are summarized and discussed. In section V, we conclude the whole paper.

II. RELATED WORK

In the past few years, various post-processing calibration methods have been proposed, and they can be classified into two groups: parametric and non-parametric calibration techniques. Platt Scaling (PS) is probably the most prevailing parametric calibration method. It aims to train a sigmoid function to map the original outputs from a classifier to calibrated probabilities [11]. Piecewise logistic regression is an extension of Platt Scaling, which assumes that the log-odds of calibrated probabilities follow a piecewise linear function [12]. There are also other parametric calibration methods. For example, the probability-mapping approach maps raw scores obtained from the classifier to the class probabilities using generalized linear models and generalized additive models [13]. Another method is called shape-restricted polynomial regression, which makes use of monotone polynomials with some semi-definite constraints to satisfy the continuously-constrained requirement of monotonicity [14]. Other parametric calibration methods include asymmetric Laplace method [15] and Beta calibration [16].

The most widely used non-parametric method is Histogram Binning (HB), which divides the outputs of a classifier into several subsets(bins) and uses the proportion of positive class in each bin as the calibrated probability [17]. There are also some extensions of Histogram Binning methods. For example, adaptive calibration of predictions (ACP) also uses the proportion of positive class as the posterior probability in each bin, but it obtains bins from a 95% confidence interval around each individual prediction [6]. Recently, another method called ROC Binning has been proposed, which constructs equal-width bins based on the ROC curves. ROC Binning can be effective when class prevalence between the training and test sets is different [18]. Bayesian binning into quantiles (BBQ) is an ensemble of multiple Histogram Binning models, which uses the combination of different equal frequency Histogram Binning models as the calibration result [19]. Another well-known calibration method is Isotonic Regression which maps the outputs into isotonic probabilities [20]. Some studies attempt to adjust the Isotonic Regression techniques. For example, the calibration method called ensemble of near-isotonic regression (ENIR) uses selective Bayesian averaging to ensemble the nearly-isotonic regression models [21], which makes a trade-off between the isotonicity and goodness-of-fit using a penalty function. Isotonic regression-based techniques can be viewed as non-parametric binning methods mapping the raw scores into a piecewise constant function.

Although various calibration methods have been proposed, there are only a few studies have attempted to find the influence of class imbalance [10], [18]. However, their research only considers Brier score as the performance measure, which only reflects one aspect of results.

III. EXPERIMENT FRAMEWORK

In this section, we present an overview on the experimental framework. First, we show the main features of the datasets and related data preprocessing procedure. Then, we describe the four calibration methods and three evaluation measures used in the experiments. Finally, we introduce the main settings as well as the statistical tests used in the experiments.

A. DATASETS

Our experiments were conducted on twenty-four binary classification datasets, which are commonly used for class imbalance research. Eleven of datasets are from KEEL [22] and thirteen are from the OpenML [23] dataset repository.

Detailed information about the datasets is summarized in Table 1, which includes names (Name), number of attributes (#Attr.), number of instances (#Size) and imbalance ratio (#IR.). Imbalance ratio is defined as:

$$IR = \frac{N_-}{N_+}, \tag{1}$$

where N_+ is the number of minority (or positive) instances and N_- is the number of majority (or negative) instances on the dataset. The datasets vary in the number of instances and attributes. The datasets are arranged in an increasing order of the imbalance ratio in Table 1. As the imbalanced ratio shows, twelve of the datasets have IR values higher than 9, which means they are highly imbalanced. Twelve of the the datasets have IR values lower than 9, which means they are relatively lowly imbalanced. In our experiments, we removed the instances which contain missing values in the data preprocessing step.

B. CALIBRATION METHODS

1) PLATT SCALING

Platt Scaling is a parametric method. It was originally built to calibrate the support vector machine model and is now also applied to other classifiers. Platt Scaling uses a sigmoid function to map the outputs of a binary classifier to calibrated probabilities. The sigmoidal function of this method is defined as follows:

$$P(y_i = 1|s_i) = \frac{1}{1 + exp(As_i + B)}$$

where s_i denotes the uncalibrated classification output of the i-th instance from the classifier, y_i is the true class label of the

TABLE 1. Summary of datasets used in the experiments.

Name	#Attr.	#Size	#IR.
magic	11	19020	1.84
titanic	4	2201	2.10
ilpd	11	583	2.49
diabetes	9	768	2.87
blood	5	748	3.20
vehicle0	18	846	3.25
new-thyroid1	5	215	5.14
JapaneseVowels	15	9961	5.17
segment0	19	2308	6.02
CastMetal1	38	327	6.79
page-blocks0	10	5472	8.79
optdigits	65	5620	8.83
yeast-0-2-5-6_vs_3-7-8-9	8	1004	9.14
climate	21	540	10.74
PizzaCutter1	38	661	11.71
shuttle-c0-vs-c4	9	1829	13.87
ozone	73	2534	14.84
wilt	6	4839	17.5
winequality-red-4	11	1599	29.17
PieChart2	37	745	46
poker-8-9_vs_6	10	1485	58.4
Satellite	37	5100	67
poker-8-9_vs_5	10	2075	82
abalone19	8	4174	129.44

i-th instance with value from {0,1}. A and B are the parameters of the sigmoidal function, which can be determined by minimizing the negative log likelihood function as follows:

$$-\sum_i y_i log(p_i) + (1 - y_i) log(1 - p_i)$$

where p_i represents the estimated probability.

The sigmoidal function can overfit data when there are only a few positive instances. To avoid overfitting, y_i is usually substituted by t_i as follows:

$$t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2}, & y_i = 1 \\ \frac{1}{N_- + 2}, & y_i = 0 \end{cases}$$

where N_+ and N_- represent the number of positive and negative instances [24], respectively.

2) HISTOGRAM BINNING

The most commonly used non-parametric approach is Histogram Binning, which is also called quantile binning. Histogram Binning first sorts uncalibrated instances according to their estimated classification scores and then partitions the score into B equal frequency bins. The calibrated probabilities of an instance in each bin can be estimated using the fraction of observed positive instances in that bin. Therefore, the instances belonging to the same bin have an identical probability [17].

3) ISOTONIC REGRESSION

Isotonic Regression is a non-parametric regression approach with the assumption that calibrated probabilities are isotonic (monotonically increasing). Isotonic Regression assumes that if a classifier ranks the instances exactly, then the mapping from outputs into probabilities should follow an isotonic function. This method relaxes the restriction of a strict form of function, and a preliminary setting of the number of bins can be dropped [20].

Usually, the pair adjacent violators (PAV) algorithm is used to obtain a step-wise constant isotonic function, which has been proved to calibrate the outputs well [25]. PAV first sorts the prediction score of each instance i, $s_1 \leq s_2 \leq \ldots \leq s_N$. Then, it estimates the probability p_i of the i-th instance by y_i, where y_i is the corresponding class label. If the probability estimates are isotonic, which means $p_i \leq p_{i+1}, i \in \{1, \ldots, N-1\}$, there will be no further calibration. If $p_i > p_{i+1}$ occurs, the following transformation is used:

$$p_i^* = p_{i+1}^* = \frac{p_i + p_{i+1}}{2} \tag{5}$$

where p_i^* is the transformed probability for the i-th instance. This process continues until a set of isotonic estimated probabilities are obtained. We can view this method as a kind of binning method, where the bin size and the number of bins depend on how well the classifier ranks examples.

4) BAYESIAN BINNING INTO QUANTILES

Bayesian binning into quantiles is a prevalent extensive form of Histogram Binning. This method considers multiple binning models to produce the calibrated probabilities. Each model differs in the number of bins, and all these models are combined by using a Bayesian score function learned from a Bayesian network [5], [19]. Let s_i and y_i define respectively an uncalibrated classifier prediction scores and the true class of the i-th instance. Also, let D define the set of all training instances. BBQ first sorts the raw scores as $S = \{s_1, s_2, \ldots, s_N\}$, where N is the total number of training data. Then, BBQ uses the partition rule P_a to partition the scores into B equal frequency bins, which can be described as a set $\{t_1, t_2, \ldots, t_B\}$. A binning model M can be defined as:

$$M \equiv \{B, P_a, \Phi\}, \tag{6}$$

where $\Phi = \{\phi_1, \phi_2, \ldots, \phi_B\}$, ϕ_b is the parameter of the binomial distribution used to describe the distribution of the positive class in the b-th bin, which is denoted by $P(y = 1|t_b)$. Therefore, Φ determines all distributions of every bin using the rule P_a. Each binning model M can be scored as follows:

$$Score(M) = P(M) \cdot P(D|M) \tag{7}$$

BBQ has the following assumptions: (1) All instances are i.i.d. (2) The class distribution in each bin follows a binomial distribution and they are independent. (3)ϕ_i follows a Beta distribution with the parameters $\alpha_b = \frac{N^*}{B} p_b$ and $\beta_b = \frac{N^*}{B}(1 - p_b)$, where N^* is a prior parameter describing the strength of our belief in the distribution, p_b is the midpoint of the interval defining the b-th bin in the binning model M.

Then, $P(D|M)$ can be derived as:

$$P(D|M) = \prod_{b=1}^{B} \frac{\Gamma(\frac{N^*}{B})}{\Gamma(N_b + \frac{N^*}{B})} \frac{\Gamma(m_b + \alpha_b)}{\Gamma(\alpha_b)} \frac{\Gamma(n_b + \beta_b)}{\Gamma(\beta_b)} \quad (8)$$

where Γ is the gamma function, m_b and n_b are the numbers of positive and negative instances in the b-th bin, and N_b is the number of total instances in the b-th bin. The term $P(M)$ specifies the prior probability of the binning model M. Usually, a uniform prior is used. After scoring the model M, a calibrated probability can be obtained by using the model average as follows:

$$P(y_i = 1|s_i) = \sum_{k=1}^{T} \frac{Score(M_k)}{\sum_{j=1}^{T} Score(M_j)} P(y_i = 1|s_i, M_k), \quad (9)$$

where T is the number of binning models we use and $P(y_i = 1|s_i, M_k)$ is the predictive probability derived from model M_k for the uncalibrated classifier output.

C. EVALUATION MEASURES

In our experiments, we used three metrics as the evaluation measures, namely, Brier score (BS), expected calibration error (ECE), maximum calibration error (MCE). Bier score and ECE provide numerical measures of overall calibration performance. MCE evaluates the stability of calibration methods. Detailed information about the three metrics is given in this section.

1) BRIER SCORE

Brier score, which is also called mean squared error [26], is a popular metric used to measure the performance of the probability estimator. The main idea of Brier score is that the most accurate calibrated probabilities have the lowest squared deviation from the class label. For the binary classification problem, Brier score can be calculated by the following formula:

$$Brier\ score = \frac{1}{N} \sum_{i=1}^{N} (y_i - p_i)^2 \quad (10)$$

where N is the number of instances, p_i and y_i represent the calibrated probability and class label for the i-th instance, respectively. Brier score can be decomposed into two separate terms as follows [27]:

$$Brier\ Score = \frac{1}{N} \sum_{i=1}^{B} N_i(e_i - o_i)^2 + \frac{1}{N} \sum_{i=1}^{B} N_i(o_i(1 - o_i)), \quad (11)$$

where N_i is the number of instances in the i-th bin, o_i is the fraction of positive instances and e_i is the mean calibrated probability in the i-th bin. The first term of this formula is called calibration loss, which indicates how close the calibrated probabilities are to the actual probabilities. The second term is called refinement loss, which measures how close the calibrated probabilities are to 0 or 1. As the decomposition shows, Brier score not only considers the calibration accuracy, but also takes the certainties of the estimate into account. Therefore, it may prefer the results pushing probability estimation towards 0 and 1, which will negatively affect the results.

2) EXPECTED CALIBRATION ERROR

Expected calibration error measures the overall performance of calibration [3]. To calculate ECE, the calibrated probabilities should be sorted and divided into several bins. Then, the values of ECE can be calculated as follows:

$$ECE = \sum_{i=1}^{B} \pi_i \cdot |o_i - e_i| \quad (12)$$

where π_i is the fraction of instances that fall into the i-th bin, o_i is the fraction of positive instances in the i-th bin and e_i is the mean calibrated probability in that bin. If the probabilities are well calibrated, ECE will be small.

3) MAXIMUM CALIBRATION ERROR

Maximum calibration error is used to measure the stability of calibration. MCE can be calculated as follows:

$$MCE = \max_{i=1}^{B} |o_i - e_i| \quad (13)$$

where o_i is the fraction of positive instances in the i-th bin, e_i is the mean calibrated probability in that bin. If a calibration method is more stable and robust, its MCE value will be smaller than other methods. On imbalanced datasets, the occurrence of one large deviation between true probabilities and calibrated probabilities can cause serious consequences. Therefore, the measurement of stability of calibration is necessary.

D. STATISTICAL TEST

To compare the performance of different calibration methods, we fisrt apply the Iman-Davenport test to determin whether there are significant differences across different calibration methods for one metric, which is a modified form of the Friedman test [28]. The Iman-Davenport test ranks calibration methods on each dataset, and computes the average rank of each calibration method R_j as follows:

$$R_j = \frac{1}{m} \sum_{i=1}^{m} r_{ij} \quad (14)$$

where m is the total number of datasets, r_{ij} is the rank of the j-th calibration method on the i-th dataset. The Iman-Davenport test statistic follows the F distribution with the degrees of freedom $k - 1$ and $(k - 1)(m - 1)$ as follows:

$$F_F = \frac{(m - 1)\chi_F^2}{m(k - 1) - \chi_F^2} \quad (15)$$

where k is the number of methods to be compared and χ_F^2 is defined as follows:

$$\chi_F^2 = \frac{12m}{k(k + 1)} [\sum_{j=1}^{k} R_j^2 - \frac{k(k + 1)^2}{4}] \quad (16)$$

TABLE 2. Experimental performance of calibration methods based on ECE.

Datasets	Method	GBDT	Rank	LR	Rank	RF	Rank	SVM	Rank
All	PS	0.0502	3.08	0.0449	2.92	0.0322	2.92	0.0384	3.00
	HB	0.0285	2.29	0.0350	2.33	0.0287	2.46	0.0274	2.08
	ISO	0.0229	**1.46**	0.0217	**1.63**	0.0213	**1.33**	0.0227	**1.50**
	BBQ	0.0671	3.17	0.0514	3.13	0.0565	3.29	0.0581	3.42
IR<9	PS	0.0442	3.17	0.0415	2.92	0.0408	3.33	0.0490	3.25
	HB	0.0365	2.00	0.0522	2.25	0.0378	2.42	0.0388	1.75
	ISO	0.0310	**1.50**	0.0286	**1.58**	0.0280	**1.42**	0.0302	**1.58**
	BBQ	0.0455	3.33	0.0691	3.25	0.0473	2.83	0.0980	3.42
IR>9	PS	0.0625	3.08	0.0591	3.00	0.0276	2.25	0.0240	2.83
	HB	0.0194	2.50	0.0206	2.25	0.0211	2.75	0.0166	2.50
	ISO	0.0149	**1.33**	0.0143	**1.42**	0.0135	**1.33**	0.0144	**1.42**
	BBQ	0.0896	3.08	0.0421	3.33	0.0696	3.67	0.0263	3.25

The null hypothesis of the Iman-Davenport test states that all calibration methods perform equally well. If the null hypothesis is rejected, which means some calibration methods have significantly different performance from others. Then, we adopt the post-hoc Holm's test to detect whether there is a significant difference between the best method and the others. The Holm's test takes the method with the lowest average rank as the control method and calculate the z-statistic as follows:

$$z_i = \frac{R_j - R^*}{\sqrt{k(k+1)/6m}} \tag{17}$$

where R^* is the average rank of the control method and m is the number of datasets we used. Then, it finds the corresponding p-values according to the normal distribution and sorts them in an increasing order $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$. If p_i is below $\alpha/(k-i)$ for a given statistical significance level, we can reject the corresponding hypothesis that the i-th method performs equally well as the control method.

E. EXPERIMENTAL SETTINGS

To evaluate the performance of calibration methods, each dataset in our experiment was randomly split into 3 subsets as suggested by [17]: training, calibration and testing set, which contained 35%, 35% and 30% of total instances, respectively. The training set was used to learn the classification model. The calibration set was used to train the calibration model, and testing set was used to evaluate the performance of each calibration method. We applied four popular supervised classification algorithms including logistic regression (LR), support vector machine (SVM), gradient boosting decision tree (GBDT) and random forest (RF) on the training sets to obtain classification models. We used four calibration methods described in subsection III-B in our experiments: Platt Scaling, Histogram Binning, Isotonic Regression and Bayesian binning into quantiles, because they are still the state-of-the-art calibration methods and widely used in practice. All of them were well developed on the validation set and used to calibrate the outputs using the test sets. After deriving calibrated probabilities, we used three metrics to evaluate different calibration methods, including ECE, MCE and Brier score on the testing set. To calculate ECE and MCE, the number of bins was set as B=10. We obtained the average scores of each metrics through repeating the procedure mentioned above 50 times and obtained robust outcomes. All the experiments were implemented in R, and we used the default parameters to train the classification models.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The full results of our experiment are presented from Table 6 to Table 8 in the Appendix. The values in each table represent the ranks of different calibration methods in terms of one metric on the datasets. To better illustrate the results, the average scores and ranks of each calibration method using different classifiers are provided from Table 2 to Table 4. Each table indicates the performance of four calibration methods in terms of one metric. In each table, the best ranks among the four calibration methods for a given classifier are marked in bold. To check the behaviors of different calibration methods for different IR ranges, the results of highly imbalanced datasets (IR > 9) and lowly imbalanced datasets (IR < 9) are provided together with the results of all datasets. We also perform global and pairwise comparisons using the statistical test mentioned in subsection III-D at the 1% significance level. If there is no significant difference in terms of the Iman-Davenport test among the calibration methods for a classifier based on one metric, the method with the best rank is marked with a star. The ranks that are significantly worse than the best one in terms of the Holm's post hoc test is underlined.

Table 2 gives the experimental results of different calibration methods in terms of ECE. As the results of all datasets rows in Table 2 show, Isotonic Regression outperforms other methods on all the four classifiers. Histogram Binning takes second place and it is significantly inferior to Isotonic Regression when GBDT and random forest is used. Platt Scaling and BBQ have larger ranks, and they are always significantly inferior to the best ranked method. The results reveal that Isotonic Regression has better performance in the imbalance scenarios in terms of the overall calibration performance and can do well with class imbalance calibration.

TABLE 3. Experimental performance of calibration methods based on Brier Score.

Datasets	Method	GBDT	Rank	LR	Rank	RF	Rank	SVM	Rank
All	PS	0.0708	**1.96**	0.0729	2.71	0.0590	1.88	0.0636	2.17
	HB	0.0639	2.96	0.0693	2.63	0.0617	3.13	0.0627	3.00
	ISO	0.0605	2.25	0.0616	**1.83**	0.0571	**1.83**	0.0581	**1.96**
	BBQ	0.0911	2.83	0.0721	2.83	0.0736	3.17	0.0808	2.88
IR<9	PS	0.0886	**1.50**	0.0944	2.25	0.0843	**1.75**	0.0921	2.33
	HB	0.0935	3.42	0.1056	3.17	0.0907	3.75	0.0913	3.00
	ISO	0.0891	1.83	0.0911	**1.42**	0.0847	2.00	0.0866	**1.42**
	BBQ	0.0933	3.25	0.1121	3.17	0.0904	2.50	0.1373	3.25
IR>9	PS	0.0570	**2.08**⋆	0.0573	2.67	0.0354	**1.75**	0.0328	**2.08**⋆
	HB	0.0339	2.50	0.0348	**2.17**⋆	0.0333	2.63	0.0338	2.92
	ISO	0.0318	2.58	0.0322	2.42	0.0298	2.08	0.0295	2.25
	BBQ	0.0893	2.83	0.0378	2.75	0.0614	3.54	0.0324	2.75

TABLE 4. Experimental performance of calibration methods based on MCE.

Datasets	Method	GBDT	Rank	LR	Rank	RF	Rank	SVM	Rank
All	PS	0.4053	3.63	0.3833	3.58	0.4467	3.79	0.4804	3.83
	HB	0.1058	**1.29**	0.1379	**1.79**	0.1113	**1.38**	0.1054	**1.21**
	ISO	0.2634	2.70	0.2420	2.17	0.2603	2.50	0.2829	2.38
	BBQ	0.2097	2.38	0.2094	2.46	0.2205	2.33	0.2465	2.58
IR<9	PS	0.3486	3.67	0.3183	3.33	0.3994	3.83	0.4799	3.67
	HB	0.1285	**1.08**	0.2050	**2.08**	0.1350	**1.33**	0.1557	**1.33**
	ISO	0.2612	2.75	0.2157	1.92	0.2405	2.58	0.2715	2.33
	BBQ	0.2305	2.50	0.2503	2.67	0.2557	2.25	0.2909	2.67
IR>9	PS	0.4809	3.83	0.4761	3.67	0.5017	3.83	0.4753	4.00
	HB	0.0851	**1.50**	0.0792	**1.33**	0.0839	**1.25**	0.0583	**1.17**
	ISO	0.2690	2.75	0.2667	2.75	0.2600	2.75	0.2760	2.67
	BBQ	0.1837	1.92	0.1777	2.25	0.1743	2.17	0.2311	2.17

Meanwhile, Platt Scaling and BBQ are unreliable for calibrating classifiers on imbalanced datasets. One explanation is that Isotonic Regression has a simpler training process and there is no need to estimate parameters. However, Platt Scaling has 2 parameters to be trained, and BBQ has many parameters to be estimated. Therefore, Platt Scaling and BBQ perform worse than Isotonic Regression in imbalance scenarios.

The performance comparison of calibration methods for different IR ranges is also offered in the remaining rows in Table 2. It shows that Isotonic Regression ranks better than the other 3 calibration methods on all classifiers in low IR scenarios, and the superiority is more obvious in high IR scenarios. Histogram Binning performs not significantly worse than Isotonic Regression in the low IR scenarios. However, its performance deteriorates in the high IR scenarios, and it is statistically inferior to Isotonic Regression for three classifiers. Platt Scaling and BBQ are significantly inferior to the Isotonic Regression in both the low IR scenarios and the high IR scenarios.

Table 3 shows the Brier score of each calibration method. From the results of all datasets in the top rows, we can find that Isotonic Regression performs better than the other methods over SVM, logistic regression, random forest, which validates the results based on ECE. Isotonic Regression takes the second place and is not significantly worse than the best method when GBDT is used as the classifier. Plat Scaling has slightly worse performance than Isotonic Regression. Plat Scaling is the top method when GBDT is used and ranks second with random forest and SVM as classifier. Histogram Binning and BBQ have larger Brier scores and they are significantly worse than Isotonic Regression for random forest, SVM and logistic regression.

As shown in the middle part for the low IR scenario rows, Histogram Binning and BBQ also perform significantly worse than the best method on all the classifiers. There is no significant difference between Isotonic Regression and Platt Scaling on all the classifiers. However, in the high IR scenarios, there is no significant difference among all the calibration methods when GBDT, logistic regression and SVM are used. BBQ performs significantly worse than the best method when random forest is used as the base classifier.

Table 4 shows the results of each calibration method with four base classifiers with respect to MCE. As the results of top all datasets rows show, the average ranks of Histogram Binning based on MCE are lower than other methods for all four base classifiers. As MCE measures the stability of the calibration method, it reveals that Histogram Binning performs more stably on imbalanced datasets. Platt Scaling has the highest average MCE scores with all classifiers and it is also significantly worse than Histogram Binning. Therefore,

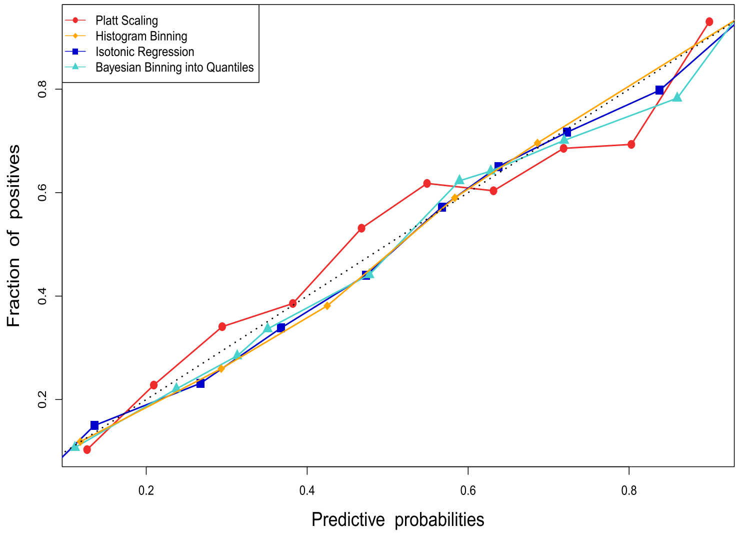

FIGURE 1. Reliability diagram for SVM calibration on *magic* dataset.

TABLE 5. Runtime of different calibration methods on each dataset (in seconds).

	GBDT				LR				RF				SVM			
Datasets	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ
abalone19	0.067	0.076	0.070	0.181	0.026	0.050	0.024	0.674	0.038	0.067	0.039	0.334	0.056	0.054	0.050	0.177
blood	0.012	0.014	0.012	0.047	0.007	0.014	0.006	0.078	0.012	0.017	0.016	0.044	0.014	0.015	0.014	0.047
CastMetal1	0.024	0.024	0.022	0.034	0.015	0.018	0.015	0.023	0.017	0.019	0.017	0.034	0.018	0.018	0.017	0.033
climate	0.021	0.025	0.017	0.036	0.010	0.013	0.009	0.038	0.013	0.017	0.013	0.040	0.016	0.018	0.016	0.037
diabetes	0.014	0.017	0.014	0.066	0.007	0.013	0.007	0.073	0.014	0.019	0.014	0.077	0.018	0.021	0.019	0.071
ilpd	0.015	0.016	0.014	0.049	0.007	0.011	0.007	0.042	0.012	0.016	0.012	0.046	0.014	0.016	0.014	0.042
JapaneseVowels	0.136	0.145	0.149	0.511	0.046	0.095	0.063	1.016	0.131	0.161	0.133	0.958	1.304	1.314	1.328	1.757
magic	0.285	0.246	0.223	0.809	0.070	0.169	0.065	0.865	0.320	0.391	0.304	0.735	5.354	5.319	5.286	6.001
newthyroid1	0.008	0.008	0.008	0.014	0.006	0.008	0.006	0.011	0.006	0.009	0.007	0.014	0.007	0.007	0.007	0.014
optdigits	0.184	0.185	0.173	0.366	0.179	0.194	0.169	0.343	0.181	0.194	0.176	0.487	3.054	3.056	3.056	3.218
ozone	0.109	0.113	0.112	0.242	0.059	0.079	0.056	0.240	0.086	0.098	0.088	0.279	0.190	0.196	0.187	0.320
pageblocks0	0.073	0.077	0.071	0.362	0.026	0.056	0.023	0.511	0.068	0.122	0.063	0.473	0.173	0.178	0.173	0.511
PizzaCutter1	0.029	0.029	0.029	0.060	0.017	0.020	0.017	0.042	0.020	0.023	0.020	0.054	0.023	0.024	0.022	0.069
poker89vs5	0.032	0.035	0.032	0.133	0.019	0.023	0.012	0.183	0.027	0.046	0.027	0.271	0.025	0.030	0.027	0.140
poker89vs6	0.024	0.027	0.024	0.093	0.010	0.018	0.010	0.176	0.020	0.032	0.020	0.255	0.028	0.027	0.025	0.169
Satellite	0.124	0.125	0.110	0.498	0.081	0.098	0.082	0.181	0.081	0.108	0.075	0.327	0.177	0.183	0.188	0.716
segment0	0.044	0.047	0.045	0.135	0.029	0.047	0.027	0.110	0.033	0.053	0.034	0.231	0.147	0.149	0.147	0.226
shuttlec0vsc4	0.027	0.030	0.027	0.049	0.016	0.025	0.017	0.084	0.015	0.024	0.016	0.053	0.019	0.021	0.020	0.222
titanic	0.026	0.029	0.024	0.164	0.009	0.021	0.009	0.167	0.018	0.030	0.017	0.184	0.049	0.055	0.052	0.120
vehicle0	0.021	0.022	0.021	0.060	0.013	0.018	0.013	0.032	0.017	0.021	0.016	0.056	0.028	0.030	0.028	0.080
wilt	0.055	0.059	0.056	0.245	0.026	0.065	0.025	0.810	0.041	0.061	0.041	0.289	0.126	0.125	0.124	0.310
winequalityred4	0.027	0.030	0.027	0.130	0.011	0.027	0.012	0.278	0.032	0.028	0.019	0.093	0.025	0.028	0.026	0.098
winequalitywhite3vs7	0.018	0.020	0.018	0.067	0.009	0.016	0.009	0.123	0.013	0.022	0.013	0.098	0.016	0.017	0.016	0.073
yeast0256vs3789	0.017	0.019	0.017	0.094	0.009	0.016	0.009	0.168	0.016	0.026	0.016	0.100	0.019	0.021	0.019	0.159
Average	0.058	0.059	0.055	0.185	0.029	0.046	0.029	0.261	0.051	0.067	0.050	0.231	0.454	0.455	0.453	0.609

Platt Scaling cannot produce robust probability estimation in imbalance scenarios. One explanation may be that Histogram Binning uses the fraction of positive instances in each bin as the estimate. Therefore, it performs steadily on each dataset with less variation. The parameters of Platt Scaling need to be trained every time. As a result, the performance of Platt Scaling will change considerably in different imbalance scenarios. Isotonic Regression and BBQ are in the middle. Isotonic Regression performs slightly better than BBQ. It is better than BBQ on 2 classifiers and inferior on 2 classifiers.

For the results of different IR ranges, we can see that Histogram Binning performs better than the other three methods in both low IR scenarios and high IR scenarios. Isotonic Regression has a relatively low average rank in the low IR scenarios. However, it has a higher average rank in the high IR scenarios compared to Histogram Binning and performs significantly worse than Histogram Binning, which means Isotonic Regression can be unstable when the datasets are highly imbalanced. Platt Scaling always has highest average MCE scores and performs significantly worse than Histogram

TABLE 6. Experimental results on individual data set (ECE).

	GBDT				LR				RF				SVM			
Datasets	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ
abalone19	2	4	1	3	4	2	1	3	4	3	1	2	4	3	2	1
blood	4	1	2	3	3	2	1	4	3	4	2	1	3	2	1	4
CastMetal1	4	3	1	2	4	3	1	2	4	3	1	2	4	3	1	2
climate	4	3	1	2	2	3	1	4	4	3	1	2	4	2	1	3
diabetes	1	2	3	4	4	2	1	3	3	2	1	4	3	2	1	4
ilpd	3	4	1	2	4	2	1	3	4	3	1	2	4	3	1	2
JapaneseVowels	2	1	3	4	4	1	2	3	4	2	1	3	3	1	2	4
magic	4	1	2	3	4	1	3	2	3	2	1	4	3	1	2	4
newthyroid1	4	2	1	3	2	3	1	4	4	1	2	3	4	1	2	3
optdigits	4	1	2	3	1	3	2	4	3	1	2	4	3	1	2	4
ozone	3	2	1	4	4	1	2	3	3	1	2	4	3	1	2	4
pageblocks0	3	1	2	4	4	1	2	3	3	1	2	4	3	1	2	4
PizzaCutter1	4	3	1	2	2	3	1	4	2	3	1	4	1	3	2	4
poker89vs5	4	3	1	2	4	2	1	3	3	2	1	4	4	3	1	2
poker89vs6	4	3	1	2	4	2	1	3	2	3	1	4	2	3	1	4
Satellite	1	3	2	4	1	3	2	4	2	3	1	4	2	1	3	4
segment0	4	2	1	3	2	4	1	3	3	4	1	2	3	2	1	4
shuttlec0vsc4	2	3	1	4	2	3	1	4	2	3	1	4	2	4	1	3
titanic	2	3	1	4	3	2	1	4	2	3	1	4	2	3	1	4
vehicle0	3	1	2	4	1	4	2	3	4	1	2	3	4	1	2	3
wilt	3	1	2	4	4	1	3	2	3	1	2	4	3	1	2	4
winequalityred4	3	4	1	2	3	2	1	4	2	3	1	4	2	4	1	3
winequalitywhite3vs7	3	2	1	4	4	2	1	3	1	4	2	3	3	2	1	4
yeast0256vs3789	3	2	1	4	3	1	2	4	2	3	1	4	3	2	1	4

TABLE 7. Experimental results on individual data set (Brier Score).

	GBDT				LR				RF				SVM			
Datasets	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ
abalone19	4	1.5	3	1.5	4	1	2	3	4	1	3	2	4	1.5	3	1.5
blood	2	1	3	4	2	3	1	4	1	4	2	3	3	1	2	4
CastMetal1	4	3	2	1	4	3	2	1	4	3	2	1	4	3	1	2
climate	1	3	4	2	1	2	4	3	4	2	3	1	2	3	4	1
diabetes	1	4	2	3	2	3	1	4	1	3	2	4	1	3	2	4
ilpd	1	4	2	3	2	4	1	3	2	4	1	3	1	4	3	2
JapaneseVowels	1	4	2	3	3	4	2	1	2	4	1	3	2	4	1	3
magic	4	2	1	3	4	3	1	2	1	3	2	4	2	3	1	4
newthyroid1	2	4	3	1	1	3	2	4	1	4	2	3	1	4	3	2
optdigits	1	4	2	3	2	4	1	3	1	4	2	3	3	2	1	4
ozone	1	3	2	4	4	2	1	3	2	4	1	3	2	3	1	4
pageblocks0	2	4	1	3	3	4	1	2	3	4	1	2	3	4	2	1
PizzaCutter1	1	3	4	2	1	2	3	4	1	3	2	4	1	2	3	4
poker89vs5	4	1.5	3	1.5	4	1	3	2	4	1	2	3	4	1.5	3	1.5
poker89vs6	4	2	3	1	4	2	1	3	1	3	2	4	1	3	2	4
Satellite	1	3	2	4	2	3	1	4	1	4	2	3	2	4	1	3
segment0	2	4	1	3	2	4	1	3	1	4	2	3	3	2	1	4
shuttlec0vsc4	1	3	2	4	1	4	3	2	1	3	2	4	1	4	3	2
titanic	2	3	1	4	3	2	1	4	2	3	1	4	2	3	1	4
vehicle0	3	1	2	4	2	4	1	3	2	3	1	4	1	4	2	3
wilt	2	4	1	3	3	4	2	1	2	4	1	3	2	4	1	3
winequalityred4	1	2	4	3	4	1	2	3	2	1	3	4	4	2	3	1
winequalitywhite3vs7	1	3	2	4	4	2	1	3	1	3	2	4	1	3	2	4
yeast0256vs3789	1	4	2	3	1	2	3	4	1	3	2	4	2	4	1	3

Binning on all classifiers, which validates our findings from all dataset rows in Table 4.

Figure 1 presents a reliability diagram [29], which demonstrates the calibration results on *magic* dataset with SVM. From the figure, we can see that Isotonic Regression and Histogram Binning results are close to the diagonal line, which means they perform better than the other 2 calibration methods. The furthest distance between the points of Histogram Binning and the diagonal line is much smaller than that of Platt Scaling, which means Histogram Binning performs more steadily than Platt Scaling. The information from Figure 1 confirms the results from the three measures and similar behaviors can be found on other datasets.

The runtime of different calibration methods on each dataset is shown in Table 5. The last row shows the average across all the datasets. We can see from Table 5, that there is no significant difference among the three methods: Platt Scaling, Histogram Binning and Isotonic Regression.

TABLE 8. Experimental results on individual data set (MCE).

	GBDT				LR				RF				SVM			
Datasets	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ	PS	HB	ISO	BBQ
abalone19	4	2	3	1	4	1	3	2	4	1	3	2	4	2	3	1
blood	4	1	3	2	4	1	2	3	4	2	3	1	4	1	3	2
CastMetal1	4	2	3	1	4	3	2	1	4	2	3	1	4	1	3	2
climate	4	1	3	2	4	1	3	2	4	1	3	2	4	1	3	2
diabetes	2	1	4	3	3	1	4	2	2	1	4	3	2	1	3	4
ilpd	3	1	4	2	4	1	3	2	4	2	3	1	4	1	3	2
JapaneseVowels	2	1	4	3	4	1	2	3	4	1	2	3	4	1	3	2
magic	4	1	2	3	4	1	3	2	3	1	2	4	2	1	4	3
newthyroid1	4	1	2	3	1	3	2	4	4	3	2	1	4	1	2	3
optdigits	4	1	3	2	3	4	2	1	4	1	2	3	4	1	2	3
ozone	4	1	3	2	4	1	3	2	4	1	3	2	4	1	3	2
pageblocks0	2	1	3	4	4	1	3	2	4	1	2	3	4	1	2	3
PizzaCutter1	4	1	3	2	4	1	3	2	4	1	3	2	4	1	3	2
poker89vs5	4	2	3	1	4	1	3	2	4	1	3	2	4	2	3	1
poker89vs6	4	2	3	1	4	1	3	2	4	1	3	2	4	1	3	2
Satellite	4	1	3	2	4	1	3	2	4	1	3	2	4	1	2	3
segment0	4	1	3	2	2	4	1	3	4	1	2	3	4	2	1	3
shuttlec0vsc4	2	3	1	4	3	4	1	2	2	3	1	4	4	1	2	3
titanic	4	2	1	3	4	2	1	3	4	2	1	3	4	2	1	3
vehicle0	4	1	2	3	1	4	2	3	4	1	2	3	4	1	3	2
wilt	4	1	3	2	4	1	2	3	4	1	2	3	4	1	2	3
winequalityred4	4	1	3	2	4	1	3	2	4	1	3	2	4	2	3	1
winequalitywhite3vs7	4	1	3	2	4	1	3	2	4	2	3	1	4	1	3	2
yeast0256vs3789	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3

Bayesian Binning always runs slowest among the four calibration methods.

As the experimental results demonstrated above, we can obtain the findings summarized as follows:

(1) The performance of calibration methods on imbalanced datasets can be different in terms of different metrics. Therefore, we need to choose calibration methods based on the application goal and the corresponding performance metric.

- Overall performance (reliability). Isotonic Regression is better than the other calibration methods in terms of overall reliability. Platt Scaling and BBQ are significantly worse than Isotonic Regression on calibrating the imbalanced data.
- Stability. Histogram Binning is better than other methods, followed by Isotonic Regression and BBQ. Platt Scaling performs worst among them.

(2) Isotonic Regression is more suitable for the imbalance scenarios than other methods. Our experimental results also show that non-parametric methods perform better than parametric methods. Methods with more parameters may not perform better than other methods.

(3) The behavior of calibration methods on imbalanced datasets with different ranges of IR can be different. Therefore, we need to determine the level of imbalance before choosing calibration methods.

- Overall performance (reliability). Isotonic Regression is better than other calibration methods in terms of overall reliability in both the low IR scenarios and the high IR scenarios. Platt Scaling and BBQ are significantly worse than Isotonic Regression on calibrating the imbalanced data in both the low IR scenarios and the high IR scenarios.
- Stability. Histogram Binning is always better than other methods. Isotonic Regression performs more unstably on calibrating highly imbalanced datasets than lowly imbalanced datasets. Thus, we should be more careful when we use Isotonic Regression in highly imbalanced scenarios.

V. CONCLUSION

Calibration on imbalanced data can be a challenging issue and research for this topic is under-developed. In this paper, we conduct a large experimental investigation of calibration in different imbalance scenarios. Different evaluation metrics are considered to provide more insights into the calibration performance from different dimensions. The experimental results show that we can adapt calibration methods based on different needs for imbalanced scenarios. We recommend using Isotonic Regression on imbalanced datasets due to its good probability estimation ability. We also show that there can be some deterioration of the reliability of Histogram Binning and the stability of Isotonic Regression in highly imbalanced scenarios. In future work, we will carry out more research to find out the mechanisms behind the experimental results and propose some new methods that will improve the calibration performance.

APPENDIX

See Table 6 to Table 8.

REFERENCES

[1] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana, ''On the effect of calibration in classifier combination,'' *Appl. Intell.*, vol. 38, no. 4, pp. 566–585, 2013.

[2] A. Niculescu-Mizil and R. Caruana, ''Predicting good probabilities with supervised learning,'' in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 625–632.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, ''On calibration of modern neural networks,'' in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[4] I. Cohen and M. Goldszmidt, ''Properties and benefits of calibrated classifiers,'' in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Berlin, Germany: Springer, 2004, pp. 125–136.

[5] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, ''Binary classifier calibration using a Bayesian non-parametric approach,'' in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2015, pp. 208–216.

[6] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado, ''Calibrating predictive model estimates to support personalized medicine,'' *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 2, pp. 263–274, Mar. 2012.

[7] A. I. Marqués, V. García, and J. S. Sánchez, ''On the suitability of resampling techniques for the class imbalance problem in credit scoring,'' *J. Oper. Res. Soc.*, vol. 64, no. 7, pp. 1060–1070, Jul. 2013.

[8] H. He and E. A. Garcia, ''Learning from imbalanced data,'' *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[9] N. Japkowicz and S. Stephen, ''The class imbalance problem: A systematic study,'' *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[10] B. C. Wallace and I. J. Dahabreh, ''Improving class probability estimates for imbalanced data,'' *Knowl. Inf. Syst.*, vol. 41, no. 1, pp. 33–52, Oct. 2014.

[11] J. C. Platt, ''Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,'' *Adv. Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[12] J. Zhang and Y. Yang, ''Probabilistic score estimation with piecewise logistic regression,'' in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, pp. 115–123.

[13] K. Coussement and W. Buckinx, ''A probability-mapping algorithm for calibrating the posterior probabilities: A direct marketing application,'' *Eur. J. Oper. Res.*, vol. 214, no. 3, pp. 732–738, Nov. 2011.

[14] Y. Wang, L. Li, and C. Dang, ''Calibrating classification probabilities with shape-restricted polynomial regression,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1813–1827, Aug. 2019.

[15] P. N. Bennett, ''Using asymmetric distributions to improve text classifier probability estimates,'' in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2003, pp. 111–118.

[16] M. Kull, T. S. Filho, and P. Flach, ''Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers,'' in *Proc. Artif. Intell. Statist.*, 2017, pp. 623–631.

[17] B. Zadrozny and C. Elkan, ''Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers,'' in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 609–616.

[18] M. Sun and S. Cho, ''Obtaining calibrated probability using ROC binning,'' *Pattern Anal. Appl.*, vol. 21, no. 2, pp. 307–322, May 2018.

[19] M. P. Naeini, G. Cooper, and M. Hauskrecht, ''Obtaining well calibrated probabilities using Bayesian binning,'' in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2901–2907.

[20] B. Zadrozny and C. Elkan, ''Transforming classifier scores into accurate multiclass probability estimates,'' in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2002, pp. 694–699.

[21] M. P. Naeini and G. F. Cooper, ''Binary classifier calibration using an ensemble of piecewise linear regression models,'' *Knowl. Inf. Syst.*, vol. 54, no. 1, pp. 151–170, Jan. 2018.

[22] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, and S. García, ''KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,'' *J. Multiple-Valued Logic Soft Comput.*, vol. 17, nos. 2–3, pp. 255–287, 2011.

[23] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, ''OpenML: Networked science in machine learning,'' *ACM SIGKDD Explorations Newslett.*, vol. 15, no. 2, pp. 49–60, Jun. 2014.

[24] H.-T. Lin, C.-J. Lin, and R. C. Weng, ''A note on Platt's probabilistic outputs for support vector machines,'' *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, Aug. 2007.

[25] M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman, ''An empirical distribution function for sampling with incomplete information,'' *Ann. Math. Statist.*, vol. 26, no. 4, pp. 641–647, Dec. 1955.

[26] G. W. Brier, ''Verification of forecasts expressed in terms of probability,'' *Monthly Weather Rev.*, vol. 78, no. 1, pp. 1–3, Jan. 1950.

[27] G. Blattenberger and F. Lad, ''Separating the brier score into calibration and refinement components: A graphical exposition,'' *Amer. Statistician*, vol. 39, no. 1, pp. 26–32, Feb. 1985.

[28] S. García, A. Fernández, J. Luengo, and F. Herrera, ''Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power,'' *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, May 2010.

[29] M. H. DeGroot and S. E. Fienberg, ''The comparison and evaluation of forecasters,'' *J. Roy. Stat. Soc., Ser. D (Statistician)*, vol. 32, nos. 1–2, pp. 12–22, 1983.

LANLAN HUANG is currently pursuing the Ph.D. degree with the Business School, Sichuan University, Chengdu, China. Her research interests include knowledge management and data mining.

JUNKAI ZHAO is currently a Research Assistant with the Data Science Research Group, Business School, Sichuan University, Chengdu, China. His research interests include machine learning and data mining.

BING ZHU (Member, IEEE) received the M.S. and Ph.D. degrees in management science and engineering from Sichuan University, Chengdu, in 2008 and 2011, respectively. He is currently an Associate Professor with the Business School, Sichuan University. He has authored or coauthored more than 30 published articles. His research interests include machine learning and business intelligence.

HAO CHEN received the Ph.D. degree from Beihang University, Beijing, in 2016. He is currently a Researcher with the Internet Research Laboratory, China Tobacco Guangxi Industrial Company, Ltd. His research interests include big data analytics and enterprise digital transformation.

SEPPE VANDEN BROUCKE received the Ph.D. degree in applied economics from KU Leuven, Belgium, in 2014. He is currently working as an Assistant Professor with the Department of Business Informatics, Universiteit Gent, Belgium, and is also a Lecturer with KU Leuven. His work has been published in well-known international journals and presented at top conferences. His research interests include business data mining and analytics, machine learning, process management, and process mining.