Machine Learning (2020) 109:2247–2281
https://doi.org/10.1007/s10994-020-05914-3

Learning with mitigating random consistency from the accuracy measure

Jieting Wang[1] · Yuhua Qian[1] · Feijiang Li[1]

Received: 13 April 2020 / Revised: 27 July 2020 / Accepted: 19 September 2020 /
Published online: 27 October 2020

Abstract

Human beings may make random guesses in decision-making. Occasionally, their guesses may generate consistency with the real situation. This kind of consistency is termed random consistency. In the area of machine leaning, the randomness is unavoidable and ubiquitous in learning algorithms. However, the accuracy (A), which is a fundamental performance measure for machine learning, does not recognize the random consistency. This causes that the classifiers learnt by A contain the random consistency. The random consistency may cause an unreliable evaluation and harm the generalization performance. To solve this problem, the pure accuracy (PA) is defined to eliminate the random consistency from the A. In this paper, we mainly study the necessity, learning consistency and leaning method of the PA. We show that the PA is insensitive to the class distribution of classifier and is more fair to the majority and the minority than A. Subsequently, some novel generalization bounds on the PA and A are given. Furthermore, we show that the PA is Bayes-risk consistent in finite and infinite hypothesis space. We design a plug-in rule that maximizes the PA, and the experiments on twenty benchmark data sets demonstrate that the proposed method performs statistically better than the kernel logistic regression in terms of PA and comparable performance in terms of A. Compared with the other plug-in rules, the proposed method obtains much better performance.

Keywords Random consistency · Accuracy · Pure accuracy · Bayes-risk consistent

Editors: Kee-Eung Kim, Vineeth N. Balasubramanian.

✉ Yuhua Qian
jinchengqyh@126.com

Jieting Wang
jietingwang@email.sxu.edu.cn

Feijiang Li
feijiangli@email.sxu.edu.cn

[1] Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, Shanxi Province, China

1 Introduction

In the process of decision-making, human beings may make random guesses without logical reasoning when they lack sufficient evidence or detailed knowledge. For instance, intern doctors are likely to diagnose patients with colds during flu season, and students are likely to choose a lucky option when faced with a difficult multiple-choices question. Sometimes, these random guesses may generate consistency with the real situation. We term this consistency the random consistency.

In the area of machine learning, randomness is unavoidable and ubiquitous in constructing classifiers, such as collecting and labeling data, selecting the structures or parameters of models and even in setting random operations (Ghahramani 2015). The prediction results of the learning models may also contain the random consistency. The random consistency produces dishonest feedback, misleads the decision direction and harms the improvement of the generalization ability, especially when the tendency of random guesses coincides with the class distribution of the real situation.

Eliminating the random consistency from evaluation measures has been well-studied in the field of educational psychology, where researchers advocate that the expected score for the accurate answer with no insight would be zero rather than one. This elimination has proven helpful in achieving a higher reliability and validity assessment and increasing the performance of examinees (Sabers and Feldt 1968; Diamond and Evans 1973; Wu et al. 2017; Budescu and Bar-Hillel 1993; Espinosa and Gardeazabal 2010). In the field of clustering evaluation, eliminating the random consistency has been an increasingly employed method to improve the quality of clustering evaluation (Hubert and Arabie 1985; Albatineh et al. 2006; Vinh et al. 2009, 2010; Albatineh and Niewiadomska-Bugaj 2011; Qian et al. 2016; Li et al. 2018, 2019).

In the area of classification, the accuracy (A) is a vital performance measure in model evaluation and learning theory. The original learning theories focus on searching the generalization bounds for the error probability (one minus accuracy) (Valiant 1984; Bartlett and Mendelson 2003). The traditional algorithms, including logistic regression, support vector machine and Adaboost are designed to optimize convex surrogate loss functions of the error probability (Zhang 2003; Bartlett et al. 2006). In ensemble learning, accuracy has been used as the preferential measure to evaluate the performance of integration (Zhou et al. 2002; Martinezmunoz and Suarez 2006). Although it is a fundamental performance measure, the accuracy does not recognize the random consistency, which may limit the performance of the algorithms based on it. In this paper, we aim to define a performance measure that eliminates the random consistency from the accuracy and to study the learning performance of the measure theoretically and experimentally.

1.1 Related work

The measure that eliminates the random consistency from the accuracy is referred to as the pure accuracy (PA). The PA measure is a kind of non-decomposable measures. The non-decomposable measures cannot be decomposed into each individual instance (Waegeman et al. 2014; Kotlowski and Dembczynski 2017; Sanyal et al. 2018). Similar measures include the F-measure, AUC, and balanced error rate (Zhao et al. 2013). For the non-decomposable measures, many learning theories and algorithms have been developed.

From the aspect of learning theory, Waegeman et al. (2014) investigated the generalization bound in terms of the F-measure when optimizing the Hamming loss and subset zero-one loss in a multi-label learning setting, and concluded that optimizing such losses as a surrogate of the F-measure leads to a high worst-case regret. Bayes-risk consistency guarantees that by increasing the amount of data, a rule can eventually learn the optimal decision with high probability. Agarwal et al. (2005b) show the Bayes-risk consistency of the AUC based on a new proposed combinatorial parameter. The key step of their proof is the symmetrization by a ghost sample that is the same as that for the classification error rate (Devroye et al. 1996). In this paper, to clarify the surrogate relation of PA and A, we show the upper bound of PA value for A-optimal rule and the upper bound of A value for PA-optimal rule. In addition, we give a Bayes-risk consistency analysis for the pure accuracy based on the Rademacher complexity in a finite hypothesis space and based on the VC dimension in an infinite space.

In optimizing the non-decomposable measures, Musicant et al. (2003) extended the support vector machine to optimize the F-measure by setting appropriate parameters in the standard SVM. Joachims (2005) proposed a large margin machine for maximizing a convex lower bound of non-decomposable measures. Hazan et al. (2010) and Song et al. (2016) trained deep neural networks by inferring the gradients of the non-decomposable measures. Narasimhan and Agarwal (2013) proposed a SVM model for optimizing the AUC via a tight convex upper bound. Waegeman et al. (2014) proposed an exact algorithm for optimizing the F-measure in the context of multi-label learning. Gao et al. (2016) proposed a one-pass AUC optimization algorithm that needed to read the training data only once. These methods directly optimize the non-decomposable measures. In addition to these direct methods, the plug-in rule is an effective method that learns a posterior probability function by the logistic regression method or some other mature methods, and searches a threshold that optimizes the objective measure. For optimizing non-decomposable measures, Narasimhan et al. (2015) simply used the bisection method to determine a threshold. The bisection method require the monotonicity of the function being solved. Then, there is still much room for improving the effectiveness of the plug-in method. Here, we give an interval search method to determine the threshold of the plug-in rule for optimizing the PA.

1.2 Contributions

We aim to verify the learning ability and Bayes-risk consistency of the PA in this paper. First, with regard to the cost-sensitive loss function, we give a non-closed formulation of the optimal rule w.r.t the PA. Based on this formulation, we illustrate that the PA is insensitive to the class distribution of classifiers and gets a low bias in minority accuracy and majority accuracy compared with A. Second, we give a novel lower and an upper bound for the optimal rules w.r.t the A and PA, respectively. These bounds help us clarify the surrogate relation between the PA and A. Furthermore, the generalization upper bounds of the PA in the worst case are given to analyze the consistency. The proof of these bounds employ the same symmetrization technique that was applied to prove the generalization upper bound of the accuracy (Devroye et al. 1996) and AUC (Agarwal et al. 2005a). However, the difference is that the PA has fractional formulation. Thus, the consistent analysis of the PA needs to handle the fractional formulation. Last, we design a plug-in rule in terms of maximizing the PA and experimentally validate its performance.

Briefly, the major contributions of this paper are summarized as follows:

- Some bounds for the optimal rules w.r.t the PA and A are given. These bounds theoretically show that the PA-optimal rule is capable of approaching a satisfactory A value for all distributions.
- Second, we develop an inequality to handle the probability of large deviations of variables in fractional form. The generalization bounds for the PA are shown in finite and infinite hypothesis space. These bounds verify the Bayes-risk consistency of learning by PA.
- We propose a plug-in rule based on the interval search method for optimizing the PA. Through it, we experimentally verify the fairness and performance of PA in learning.

The organization of this paper is presented as follows: We give the definition of the PA in Sect. 2. In Sect. 3, two examples are given to show the necessity of evaluating classifiers by the PA. In Sect. 4, a surrogate analysis between the PA and the A is conducted. In Sect. 5, the generalization upper bounds of the PA are developed. We propose a plug-in rule for optimizing the PA and experimentally validate its performance in Sect. 6. We form a conclusion and propose future work in Sect. 7.

In this paper, definitions and theorems which are tagged with a literature reference are taken from the literature, while the original ones come without such a tag. All the proofs are presented in the "Appendix".

2 Preliminaries

We consider the task of binary classification. Let $\mathcal{X} \subset \mathcal{R}^d$ and $\mathcal{Y} = \{+1, -1\}$ be the feature space and the label space, respectively. The underlying distribution of $\mathcal{X} \times \mathcal{Y}$ is usually unknown, and we only have a collection of empirical data $\mathcal{S}_N = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$ that are drawn independently from this distribution. The goal of classification is to learn a classifier $h(\boldsymbol{x})$ mapping from $\mathcal{X}$ to $\mathcal{Y}$ via $\mathcal{S}_N$. Let $\mathcal{H}$ be the hypothesis space, from which the classifier $h(\boldsymbol{x})$ is learnt. To evaluate the performance of classifiers, the confusion matrix is usually employed. Let TP, FP, FN, TN denote the true positive $\mathbb{P}(h(X) = +1, Y = +1)$, false positive $\mathbb{P}(h(X) = +1, Y = -1)$, false negative $\mathbb{P}(h(X) = -1, Y = +1)$ and true negative $\mathbb{P}(h(X) = -1, Y = -1)$, respectively. Let p and $q(h)$ denote the probability of $\mathbb{P}(Y = +1)$ and $\mathbb{P}(h(X) = +1)$, respectively. The confusion matrix is shown in Table 1.

Based on the confusion matrix, the accuracy (A) and the error probability (L) are defined as:

$$A(h) = \mathbb{P}(h(X) = Y) = TP + TN, \tag{1}$$

$$L(h) = \mathbb{P}(h(X) \neq Y) = FP + FN. \tag{2}$$

Table 1 Confusion matrix

h(X)	Y		
	$Y = +1$	$Y = -1$	Total (h)
$h(X) = +1$	TP	FP	$q(h)$
$h(X) = -1$	FN	TN	$1 - q(h)$
Total (Y)	p	$1 - p$	1

2.1 The definition of PA

To define the pure accuracy (PA), we begin with giving the definition of random accuracy (RA), which aims to measure the random consistency in accuracy. For the classifier $h(\boldsymbol{x})$ to be evaluated, let $\mathcal{H}^{q(h)}$ be the set of all possible binary partitions with the same class distribution as it:

$$\mathcal{H}^{q(h)} = \left\{h' : \mathbb{P}\left(h'(X) = +1\right) = q(h), h'(X) \in \{+1, -1\}\right\}. \tag{3}$$

Considering that the output preference of the classifier (tendency of predicting which instances as positive) is unknown in advance, we suppose the partitions in $\mathcal{H}^{q(h)}$ are uniformly distributed. Because the partitions in $\mathcal{H}^{q(h)}$ have the same output randomness as the classifier to be evaluated, we define RA as the expectation accuracy over the partitions in $\mathcal{H}^{q(h)}$.

Lemma 1 *When the partitions in $\mathcal{H}^{q(h)}$ are distributed uniformly, the expectation accuracy of partitions in $\mathcal{H}^{q(h)}$ is:*

$$\mathbb{E}_{h' \in \mathcal{H}^{q(h)}} A(h') = pq(h) + (1-p)(1-q(h)). \tag{4}$$

Definition 1 The RA is defined as:

$$RA(h) = pq(h) + (1-p)(1-q(h)). \tag{5}$$

Definition 2 The PA is defined as:

$$PA(h) = \frac{A(h) - RA(h)}{1 - RA(h)}. \tag{6}$$

Definition 3 The pure loss (PL) is defined as:

$$PL(h) = 1 - PA(h) = \frac{1 - A(h)}{1 - RA(h)}. \tag{7}$$

The denominator of PA guarantees the maximum value to be 1. Note that the formulation of the PA coincides with the definition of Cohen's κ statistic (Cohen 1960; Scott 1955; Goodman and Kruskal 1963). The difference between them is how to define the random consistency. In the definition of Cohen's κ statistic, random consistency is called as chance agreement. The chance agreement is the agreement degree that the two raters give their ratings independently. The chance agreement between the classifier $h(X)$ and the label label Y is:

$$\mathbb{P}(h(X) = Y) = \sum_{l=\{-1,+1\}} \mathbb{P}(h(X) = Y = l) = pq(h) + (1-p)(1-q(h)). \tag{8}$$

The way we define the RA gives a general framework to measure the random consistency in measures and is helpful to propose new performance measures.

Cohen's κ statistic has been successfully used in the area of psychology (Cameron et al. 2003) and medicine (Blair and Stanley 2008). The advantage of correction for the expected agreement by chance has made Cohen's κ statistic commonly be used as a reliable performance measure in the area of machine learning (Ferri et al. 2009;). In ensemble learning, Kappa-error diagrams have been used to gain insights about the effectiveness of classifier ensembles (Kuncheva 2013) and to prune classifiers (Margineantu and Dietterich 1997). In addition, Cohen's κ statistic has been used for feature selection (Vieira et al. 2010).

3 On the advantages of pure accuracy measure

A learning algorithm sensitive to the class distribution may get a decision boundary that deviates from the optimal one. Thus, the learning objective should be insensitive to the output class distribution. The extensively applied accuracy does not satisfy this property. We employ Example 1 to show that the PA is satisfactory in this respect.

Example 1 (Class distribution insensitivity) In this example, we aim to compare the evaluation result of the A and PA on the prediction results with different class distribution. Under the settings of $N = 100$ and $p = 0.3$, we randomly generate a binary vector as the true label vector. A partition with a fixed class distribution q can be generated by Algorithm 1. The class distribution q is varied from 0 to 1 with a step of 0.05. Under each q, we run Algorithm 1 1000 times to generate 1000 partitions and use A and PA to evaluate the partitions, respectively. The distributions of the A value and PA value are shown in Fig. 1. From Fig. 1, it is easy to observe that the value of A decreases with the increase of q, while the value of PA is always near zero. This finding reflects that the A is sensitive to the class distribution of classifiers, while the PA is not.

Algorithm 1 Generator of Partition with a Fixed Class Distribution

```
Require: Data set $\mathcal{S}_N = \{(\boldsymbol{x}_i, y_i), i = 1, 2..., N\}$, class ratio $q \in [0, 1]$.
for each $i \in N$ do
    Generating $q_0 \sim$ uniform$(0, 1)$.
    if $q_0 < q$ then $h(\boldsymbol{x}_i) = +1$;
    else   $h(\boldsymbol{x}_i) = -1$
    end if
end for
Ensure: The predicted label $h(\boldsymbol{x}_i)$, $i = 1, 2..., N$.
```

Further, we give the classifier that maximizes A and PA, respectively.

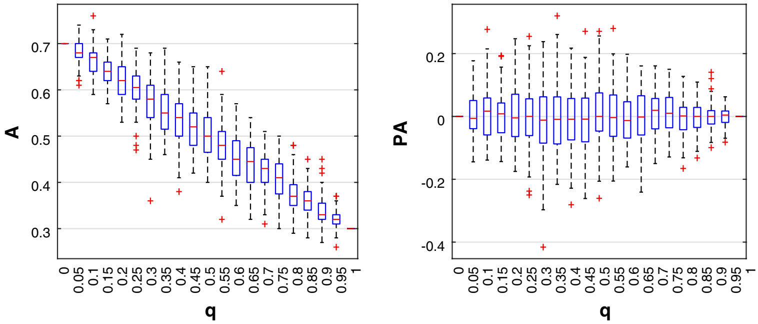


Fig. 1 Distribution of A and PA under different q. Under each q, the box plot depicts the A values (left panel) and PA values (right panel) of Algorithm 1

Lemma 2 (Devroye et al. 1996) *Let $\eta(\boldsymbol{x}) = \mathbb{P}(Y = +1|X = \boldsymbol{x})$ be the conditional class probability given $X = \boldsymbol{x}$. The classifier that maximizes the A or minimizes the L is:*

$$h_A^*(\boldsymbol{x}) = \arg\max_h A(h) = \begin{cases} +1, & \eta(\boldsymbol{x}) > \frac{1}{2}, \\ -1, & otherwise. \end{cases} \tag{9}$$

Correspondingly, the minimal error probability is

$$L^* = L(h_A^*) = \mathbb{E}_X \min\{\eta(X), 1 - \eta(X)\}. \tag{10}$$

Theorem 1 *The classifier that maximizes the PA is*

$$h_{PA}^*(\boldsymbol{x}) = \arg\max_h PA(h) \tag{11}$$

$$= \begin{cases} +1, & \eta(\boldsymbol{x}) > (\frac{1}{2} - p)PA^* + p, \\ -1, & otherwise. \end{cases} \tag{12}$$

where $PA^* = PA(h_{PA}^*)$ and $p = \mathbb{P}(Y = +1)$.

For the cost-sensitive loss $L_\rho = \rho FP + (1 - \rho)FN$, it is known that when ρ is smaller, more attention will be paid to the minority class to get a smaller L_ρ. According to the proof of Theorem 1, PA is equivalent to L_ρ with $\rho = (1/2 - p)PA^* + p$. Due to $PA^* \leq 1$, a smaller p value will generate a smaller $(1/2 - p)PA^* + p$ value. In this case, h_{PA}^* will pay more attention to the minority class. Thus, h_{PA}^* may be insensitive to class distribution.

In learning classifiers, the minority class is often overwhelmed by the majority class to guarantee a higher overall accuracy (He and Garcia 2009). Then the classifiers learnt by optimizing the accuracy or error probability are usually biased to the majority class. This phenomenon is particularly desirable to avoid because the minority class is precious and inadequately represented. We employ Example 2 to show that the pure accuracy can mitigate the classification bias.

Example 2 (Fairness) To measure the bias of the classifier $h(X)$, we use the absolute difference of the two class accuracy:

$$Bias(h) = |\mathbb{P}(h(X) = -1|Y = -1) - \mathbb{P}(h(X) = +1|Y = +1)| \tag{13}$$

Assume that two class data are generated from Gaussian distribution: $\mathcal{N}(\boldsymbol{\mu_1}, \Sigma)$ and $\mathcal{N}(\boldsymbol{\mu_2}, \Sigma)$. The label of the minority class is corrupted by the instance-independent noise at the level s_1: $\mathbb{P}(\widetilde{Y} = -1|Y = +1) = s_1$.

For this learning task, the bias of h_A^* is:

$$Bias(h_A^*) = \left|\Phi\left(\frac{d_0 + \Delta/2}{\sqrt{\Delta}}\right) - 1 + \Phi\left(\frac{d_0 - \Delta/2}{\sqrt{\Delta}}\right)\right|, \tag{14}$$

where $\Phi(\bullet)$ is the cumulative distribution function of the standard normal distribution, $\Delta = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})' \Sigma^- (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ and $d_0 = \ln \frac{1-p}{p} \frac{1}{1-2s_1}$. Due to the formulation of h_{PA}^* is non-closed, the bias of it is simulate through a large number of instances. First, a sample that obey the distribution of this task are generated with a size of 10^4. Then, the threshold that optimizes the PA is searched from the range [0, 1] with a step 10^{-4}, and the bias of h_{PA}^* is calculated through the sample.

Let $\boldsymbol{\mu_1} = -1$, $\Sigma = 1$ and $\boldsymbol{\mu_2}$ vary from 0 to 2, p vary from 0.05 to 0.35 and the one-side noise level s_1 vary from 0 to 0.5. The bias curve of h_A^* (the dashed line) and that of h_{PA}^* (the solid line) are shown in Fig. 2. As Fig. 2 shown, the dashed line is consistently lower than the solid line in each case, which demonstrates that learning by PA is more fair than learning by A under different imbalance degree, overlap degree and noise level.

4 Surrogate analysis of the optimal rules

The task of classification is to predict the labels of future observations. The optimal classifier is usually obtained by minimizing a loss function. From the same hypothesis space, different loss functions usually obtain different optimal classifiers. In this section, we focus on giving some novel bounds for $h_{PA}^*(\boldsymbol{x})$ and $h_A^*(\boldsymbol{x})$ to clarify the substitution relationship between them in learning classifiers. Theorem 2 (derived by Lemma 3) and Theorem 3 (derived by Lemma 4) are major results of this section.

Lemma 3 *For all distributions, the plug-in rule with ρ as the decision threshold*

$$h_\rho(\boldsymbol{x}) = \begin{cases} +1, & \eta(\boldsymbol{x}) > \rho, \quad \textit{where} \quad \rho \in (0, \frac{1}{2}], \\ -1, & \textit{otherwise}, \end{cases} \tag{15}$$

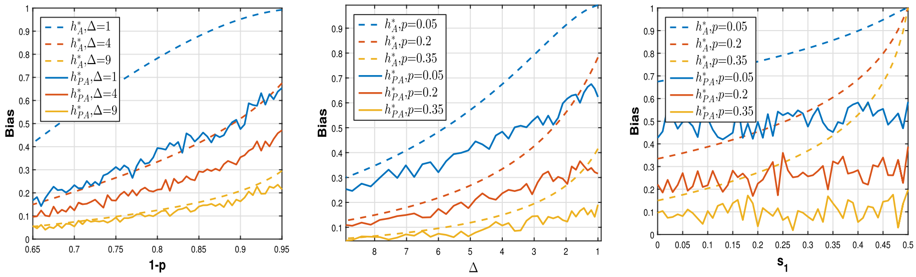
satisfies:

$$L(h_\rho) \le \frac{1-\rho}{\rho} L^*, \tag{16}$$

when $\rho = 1/2$, the equality holds.

Lemma 3 gives an upper bound on the error probability of the plug-in rule. According to Lemma 3, we have:

Theorem 2 *For all distributions, suppose that $p = \mathbb{P}(Y = +1) \le \frac{1}{2}$, the error probability of h_{PA}^* satisfies:*



(a) (b) (c)

Fig. 2 Bias of h_A^* and h_{PA}^* as a function of the ratio of majority class (left panel), the Mahalanobis distance of the two distributions (middle panel) and the one-side noise level (right panel). The dashed line is the bias curve of h_A^*, and the solid line is that of h_{PA}^*

$$L^* \leq L(h^*_{PA}) \leq \left(\frac{1}{(\frac{1}{2} - p)PA^* + p} - 1 \right) L^*. \tag{17}$$

From Theorem 2, we can conclude that the error probability of the optimal classifier learnt by PA satisfies $L(h^*_{PA}) \to L(h^*_A)$ as $PA^* \to 1$ for all distributions.

Lemma 4 *For all distributions, suppose that* $p = \mathbb{P}(Y = +1) \leq \frac{1}{2}$, *the pure loss of* h^*_A *satisfies:*

$$PL(h^*_A) \leq \frac{L^*}{p\left(\frac{3}{2} - p\right) - L^*\left(\frac{1}{2} - p\right)}. \tag{18}$$

Lemma 4 gives the upper bound of the pure loss of h^*_A with respect to L^*. To obtain the convergence relation between $PL(h^*_A)$ with $PL(h^*_{PA})$, we further amplifying L^* in Theorem 3.

Theorem 3 *For all distributions, suppose* $p \leq \frac{1}{2}$, *the pure loss of* h^*_A *satisfies:*

$$PL(h^*_{PA}) \leq PL(h^*_A) \tag{19}$$

$$\leq \frac{2(1 - p)}{p(3 - 2p) - L^*(1 - 2p)} PL(h^*_{PA}). \tag{20}$$

From Theorem 3, we can conclude that the pure loss of the optimal classifier learnt by A satisfies $PL(h^*_A) \to PL(h^*_{PA})$ as $L^* \to 0$ only when $p = \frac{1}{2}$. Based on Theorems 2 and 3, we can infer that learning by PA can obtain a satisfactory A for all distributions, while learning by A can obtain a satisfactory PA only when the class distribution is balanced. We also employ Example 3 to reflect this phenomenon.

Example 3 (Surrogate analysis) In this example, we aim to analyse the surrogate relation of A and PA. Under the settings of $N = 100$ and $p = \{0.1, 0.2, 0.5\}$, we enumerate all possible values of *FP* and *FN* and calculate the A values and PA values. The A value and PA value of each pair of (*FP*, *FN*) under different *p* are shown in Fig. 3. From Fig. 3, we can observe that under the settings $p = \{0.1, 0.2\}$, when the PA value tends to 1, most of the A values tends to 1, while when the A value tends to 1, most of the PA values are low. When $p = 0.5$, the relation between A and PA is linear.

5 Bayes-risk consistency analysis of learning by the pure accuracy measure

The underlying distribution of $\mathcal{X} \times \mathcal{Y}$ is usually unknown, and we only have a collection of the empirical data $\mathcal{S}_N = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$ that is drawn independently from the distribution. In machine learning, the classifier is generally obtained by the principle of empirical risk minimization (ERM). The feasibility of the ERM is guaranteed by the property of Bayes-risk consistency. The corresponding loss function of PA is PL. Therefore, in this section, we validate the learnability of PA by analyzing the Bayes-risk consistency of PL.

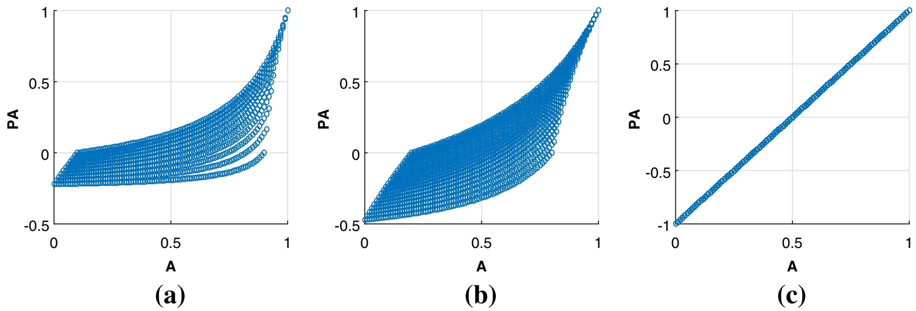


Fig. 3 Surrogate analysis of A and PA when $p = 0.1$ (left panel), $p = 0.2$ (middle panel) and $p = 0.5$ (right panel)

For the risk function R, let $\widehat{R}_N(h)$ be the empirical risk calculated on $\mathcal{S}_N$: $\widehat{R}_N(h) = \mathbb{E}_{X \times Y \in \mathcal{S}_N} R(h(X), Y)$. ERM obtains the optimal rule $h^*_{\widehat{R}_N}$ from a hypothesis space $\mathcal{H}$ by minimizing $\widehat{R}_N(h)$:

$$h^*_{\widehat{R}_N} = \arg\min_{h \in \mathcal{H}} \widehat{R}_N(h). \tag{21}$$

To guarantee the feasibility of the ERM, the property of Bayes-risk consistency is defined as:

Definition 4 (Devroye et al. 1996) The rule $h^*_{\widehat{R}_N}$ is Bayes-risk consistent, if for any small enough ε, it satisfies

$$\lim_{N \to \infty} \mathbb{P}(|R(h^*_{\widehat{R}_N}) - \inf_h R(h)| > \varepsilon) = 0. \tag{22}$$

The Bayes-risk consistency requires that the empirical optimal hypothesis $h^*_{\widehat{R}_N}$ has a large probability of converging to the universal optimal hypothesis as the number of empirical data tends to infinite.

To analysis the Bayes-risk consistency, the gap between $R(h^*_{\widehat{R}_N})$ and $\inf_h R(h)$ is usually upper bounded by (Devroye et al. 1996):

$$R(h^*_{\widehat{R}_N}) - \inf_h R(h) \leq 2 \sup_{h \in \mathcal{H}} |\widehat{R}_N(h) - R(h)|, \tag{23}$$

which is known as the estimation error. The estimation error measures the performance gap between the empirical data and the underlying distribution. The convergence of the estimation error ensures that the rule learnt finite samples can be generalized to infinite samples. The bound of the estimation error, the so-called generalization bound, is the key factor in studying the property of the Bayes-risk consistency.

The Rademacher complexity (Bartlett and Mendelson 2003) and the VC Dimension (Vapnik and Chervonenkis 1971) are two complexity measures of the hypothesis space; they have a crucial role in bounding the estimation error in the sense of accuracy. Here, we use the generalization bounds based on them to analyse the Bayes-risk consistency of learning by PA. To save space, we omit the definitions of the Rademacher complexity, the VC dimension and the corresponding generalization bounds.

5.1 The Bayes-risk consistency of the pure loss measure in a finite hypothesis space

The fractional form of the pure loss leads to that the empirical value of it is not an unbiased estimation of the expected value. Therefore, the techniques in deriving the generalization bounds of the error probability (Theorem 8 in Bartlett and Mendelson (2003) and Theorem 2 in Vapnik and Chervonenkis (1971) cannot be directly applied. Here, we establish a bridge between the estimation error of the pure loss and that of the error probability; and then obtain the Bayes-risk consistency of the pure loss in finite hypothesis space and infinite hypothesis space based on Theorem 8 in Bartlett and Mendelson (2003) and Theorem 2 in Vapnik and Chervonenkis (1971), respectively.

First, we give the formulation of the empirical error probability $\widehat{L}_N(h)$ and the empirical random accuracy $\widehat{RA}_N(h)$ to analysis the Bayes-risk consistency:

$$\widehat{L}_N(h) = \sum_{i=1}^{N} \mathbf{I}\{h(\boldsymbol{x}_i) \neq y_i\}, \tag{24}$$

$$\widehat{RA}_N(h) = \frac{1}{N|\mathcal{H}^{q(h)}|} \sum_{j=1}^{|\mathcal{H}^{q(h)}|} \sum_{i=1}^{N} \mathbf{I}\{h_j(\boldsymbol{x}_i) = y_i\}, \tag{25}$$

where $h_j \in \mathcal{H}^{q(h)}$, $|\bullet|$ is the cardinality of a set and $\mathbf{I}\{\bullet\}$ is the indicator function. Then, the empirical pure loss $\widehat{PL}_N(h)$ is

$$\widehat{PL}_N(h) = \frac{\widehat{L}_N(h)}{1 - \widehat{RA}_N(h)}. \tag{26}$$

In practice, according to Lemma 1, the empirical random accuracy is computed by:

$$\widehat{RA}_N(h) = 1 - \widehat{p}_N - (1 - 2\widehat{p}_N)\widehat{q(h)}_N, \tag{27}$$

where

$$\widehat{p}_N = \sum_{i=1}^{N} \mathbf{I}\{y_i = +1\}, \tag{28}$$

$$\widehat{q(h)}_N = \sum_{i=1}^{N} \mathbf{I}\{h(\boldsymbol{x}_i) = +1\}. \tag{29}$$

Lemma 5 *For two random variables* $Z_1, Z_2 \in [0,1]$, *any* $\varepsilon \in (0,1]$, *let* $\alpha = \mathbb{E}Z_1\mathbb{E}Z_2/(2\mathbb{E}Z_1 + \mathbb{E}Z_2)$, *we have*

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{Z_1}{Z_2} - \frac{\mathbb{E}Z_1}{\mathbb{E}Z_2}\right| > \varepsilon\right) \\ &\leq \mathbb{P}\big(|Z_1 - \mathbb{E}Z_1| > \alpha\varepsilon\big) + 3\mathbb{P}\big(|Z_2 - \mathbb{E}Z_2| > \alpha\varepsilon\big). \end{aligned} \tag{30}$$

Lemma 5 links the probability of the large estimation error of the fractional variable to that of the numerator and denominator. Based on Lemma 5, we obtain Theorems 4 and 5.

Theorem 4 *Suppose the cardinality of* $\mathcal{H}$ *is finite:* $|\mathcal{H}| < \infty$, *then for every* $h \in \mathcal{H}$, *any* $\varepsilon \in (0, 1]$, *we have*

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{h\in\mathcal{H}} \left| \widehat{PL}_N(h) - PL(h) \right| > \varepsilon \right\} \\ &\quad \le 8|\mathcal{H}| \exp\left\{ -2N\left(\alpha\varepsilon - \frac{Rc(\mathcal{H})}{2} \right)^2 \right\}, \end{aligned} \tag{31}$$

where $\alpha = \min_{h\in\mathcal{H}} \frac{L(h)}{2+PL(h)}$ *and* $Rc(\mathcal{H})$ *is the Rademacher complexity of* $\mathcal{H}$.

Theorem 4 provides the probability of the large estimation error in terms of the number of the empirical data in finite hypothesis space. From Theorem 4, we can conclude that learning by the PA is Bayes-risk consistency in a finite hypothesis space.

5.2 The Bayes-risk consistency of the pure loss measure in an infinite hypothesis space

In this section, we consider the Bayes-risk consistency in an infinite hypothesis space. For an infinite hypothesis space, the union bound cannot be utilized. We utilize the symmetrization technical to bound the estimation error of the pure loss. Next, we divide the hypothesis space into $N + 1$ subspaces according to the class probability of hypothesis functions, to ensure that each hypothesis subspace has the same degree of random accuracy. Then, we employ the VC bound of the error probability to bound the estimation error of the pure loss.

Lemma 6 *Let* $\mathcal{S}'_N = \{(\boldsymbol{x}'_1, y'_1), ..., (\boldsymbol{x}'_N, y'_N)\}$ *be an independent and identically distributed collection as* $\mathcal{S}_N$ *and* $\widehat{PL}'_N(h)$ *is the corresponding empirical pure loss. Suppose* $N \ge 5(6 + 4\alpha\varepsilon)\alpha^{-2}\varepsilon^{-2}$, *where* $\alpha = \min_{h\in\mathcal{H}} \frac{L(h)}{2PL(h)+1}$, $\varepsilon \in (0, 1]$, *then we have*

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{h\in\mathcal{H}} \left| \widehat{PL}_N(h) - PL(h) \right| > \varepsilon \right\} \\ &\quad \le 2\mathbb{P}\left\{ \sup_{h\in\mathcal{H}} \left| \widehat{PL}_N(h) - \widehat{PL}'_N(h) \right| > \frac{\varepsilon}{2} \right\}. \end{aligned} \tag{32}$$

Theorem 5 *As the same condition as Lemma 6 and suppose the VC dimension of* $\mathcal{H}$ *is finite:* $d_{vc}(\mathcal{H}) < \infty$, *we have:*

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{h\in\mathcal{H}} \left| \widehat{PL}_N(h) - PL(h) \right| > \varepsilon \right\} \\ &\quad \le 4(N + 1) \exp\left\{ -\left(\frac{\varepsilon^2(1 - |2\hat{p}_N - 1|)^2}{16} - \frac{d_{vc}(\mathcal{H}) \ln(2eN/d_{vc}(\mathcal{H}))}{N} \right) N \right\}. \end{aligned} \tag{33}$$

Theorem 5 provides the probability of the large estimation error in terms of the number of the empirical data in infinite hypothesis space. From Theorem 5, we can conclude that learning by the PA is Bayes-risk consistent in an infinite hypothesis space.

6 Performance validation of learning by the pure accuracy measure

By the Bayes-risk consistency, we have shown that the PA can be utilized to learn classifiers through minimizing PL. However, due to the fractional form, optimizing PL is a challenging task. To handle this challenge, we introduce the plug-in rule and propose an interval search method.

The plug-in rule refers to a rule with a formulation of $h_{\delta^*}(\boldsymbol{x}) = sign(\hat{\eta}(\boldsymbol{x}) - \delta^*)$, where $\hat{\eta}(\boldsymbol{x})$ is an estimator of the posterior probability $\eta(\boldsymbol{x}) = \mathbb{P}(Y = +1|X = \boldsymbol{x})$ and δ^* is a threshold (Koyejo et al. 2014). The plug-in method mainly contains the following steps: first, randomly split the training data $\mathcal{S}_N$ into $\mathcal{S}_1$ and $\mathcal{S}_2$; second, learn $\hat{\eta}(\boldsymbol{x})$ by minimizing a loss function on $\mathcal{S}_1$; third, determine δ^* by maximizing the learning objective on $\mathcal{S}_2$.

In Narasimhan et al. (2014), it has been proved that assigning an empirical threshold to a suitable posterior probability estimate can optimize the performance measures expressed as a function of the *TP* and *TN* and p. That is, the plug-in method can optimize a complex performance measure through searching a decision threshold that optimizes the measure for the posterior probability estimate. The major focus of this section is developing an method to search the threshold that optimizes PL rather than to learn the posterior probability $\hat{\eta}(\boldsymbol{x})$.

In this section, first, we introduce the method to learn the posterior probability. Then, we discuss some methods of determining the threshold that optimizes PL and propose a interval search method. Finally, we experimentally validate the performance of the interval search method and the classifier learnt by the PA.

6.1 Learning $\hat{\eta}(\boldsymbol{x})$

Many methods can be employed to learn $\hat{\eta}(\boldsymbol{x})$. Here, we introduce the kernel logistic regression model, which is proven to be a suitable posterior probability estimate (Ingo 2005; Narasimhan et al. 2014; Menon et al. 2013). The kernel logistic regression model is:

$$\max_{\alpha_j} \sum_{i=1}^{|\mathcal{S}_1|} \sum_{j=1}^{|\mathcal{S}_1|} \alpha_j y_j K(\boldsymbol{x}_j, \boldsymbol{x}_i) y_i - \sum_{i=1}^{|\mathcal{S}_1|} \log\left(1 + \exp\left(\sum_{j=1}^{|\mathcal{S}_1|} \alpha_j y_j K(\boldsymbol{x}_j, \boldsymbol{x}_i)\right)\right) \tag{34}$$

where α_i are the variables to be solved and $K(\bullet, \bullet)$ is kernel function. With the optimal α_i^* is obtained by the gradient descent method, we have

$$\hat{\eta}(\boldsymbol{x}) = \frac{1}{1 + \exp(-\sum_{j=1}^{|\mathcal{S}_1|} \alpha_j^* y_j K(\boldsymbol{x}_j, \boldsymbol{x}))}. \tag{35}$$

6.2 The interval search method

As for determining δ^*, different threshold settings correspond to optimizing different learning objective functions.

To optimize the accuracy, the threshold δ^* of the plug-in rule is 0.5, and this is the so-called kernel logistic regression (KLR) method. To optimize the balanced accuracy (BA), the threshold δ^* of the plug-in rule is p (Menon et al. 2013).

For the measures in a fractional form, search strategies are effective and simple. An intuitive approach to determine the optimal threshold is the point-wise search method, namely, evaluating the fractional measure at each possible threshold and outputting the best performing threshold. There is no doubt that exhausting all possible thresholds is impossible. The gird search is a method to handle this, which divide the range of the threshold into multiple equal intervals and set the end points as the candidate thresholds. Besides, the posterior probabilities on $\mathcal{S}_2$ can also be set to the candidate thresholds. We term this search strategy the $\mathcal{S}_2$-search. The gird search method and the $\mathcal{S}_2$-search method search the threshold in a limited range. In addition to the point-wise search methods, the bisection method transforms the fractional measure to a one-dimensional function and obtains the optimal threshold by solving the zero root of the one-dimensional function in binary (Narasimhan et al. 2015). The bisection requires that the objective function be monotone on the interval, while the fractional performance measures are usually non-monotonic with respect to the threshold.

In this subsection, we develop a method for searching the optimal threshold via the interval search method, and use this method to minimize PL. The interval search method is an effective way to search the local minimum of a unimodal function (Chong and Żak 2011). For a unimodal one-dimensional function $f(r)$ defined in $[\alpha, \beta]$, to obtain the minimum r^*, the interval search method is based on the idea that it produces a series of intervals $[\alpha_k, \beta_k]$, where $[\alpha_{k+1}, \beta_{k+1}] \subset [\alpha_k, \beta_k]$ and $\lim_{k\to\infty} \beta_k = \lim_{k\to\infty} \alpha_k = r^*$. Specifically, the interval search method inserts two points in each iteration and produces $[\alpha_k, \lambda_k, \mu_k, \beta_k]$. If $f(\lambda_k) < f(\mu_k)$, then $\alpha_{k+1} = \alpha_k$ and $\beta_{k+1} = \mu_k$; otherwise, $\alpha_{k+1} = \lambda_k$ and $\beta_{k+1} = \beta_k$. When the interval length is reduced by the ratio of $1 - (\sqrt{5} - 1)/2$, the interval search method is so-called gold section method.

For any plug-in rule $h_\delta(\boldsymbol{x}) = sign(\widehat{\eta}(\boldsymbol{x}) - \delta)$, we briefly discuss about whether the PL is a unimodal function of the threshold δ. According to the proof of Theorem 1, the PL is consistent to the cost-sensitive loss with the optimal threshold as the cost weight:

$$L_{\delta^*}(\delta) = \delta^* FP(\delta) + (1 - \delta^*) FN(\delta), \tag{36}$$

where δ^* is the minimum of $L_{\delta^*}(\delta)$:

$$\delta^* = \operatorname*{argmin}_{\delta} L_{\delta^*}(\delta), \tag{37}$$

and $FP(\delta) = \mathbb{P}(\eta(X) > \delta, Y = -1)$, $FN(\delta) = \mathbb{P}(\eta(X) \le \delta, Y = +1)$. Because

$$FP(\delta) = \mathbb{P}(\eta(X) > \delta, Y = -1) = \mathbb{P}(\eta(X) > \delta) - \mathbb{P}(\eta(X) > \delta, Y = +1), \tag{38}$$

we have:

$$\begin{aligned} L_{\delta^*}(\delta) &= \delta^* FP(\delta) + (1 - \delta^*) FN(\delta) \\ &= \delta^* \mathbb{P}(\eta(X) > \delta) - \delta^* \mathbb{P}(Y = +1) + \mathbb{P}(\eta(X) \le \delta, Y = +1). \end{aligned} \tag{39}$$

For $\delta_1 < \delta_2$, we have:

$$\begin{aligned} & L_{\delta^*}(\delta_1) - L_{\delta^*}(\delta_2) \\ & \quad = \delta^* \mathbb{P}(\eta(X) \in (\delta_1, \delta_2]) - \mathbb{P}(\eta(X) \in (\delta_1, \delta_2], Y = +1). \end{aligned} \tag{40}$$

Thus, if

$$\frac{\mathbb{P}(\eta(X) \in (\delta_1, \delta_2], Y = +1)}{\mathbb{P}(\eta(X) \in (\delta_1, \delta_2])} < \delta^* \tag{41}$$

we have $L_{\delta^*}(\delta_1) > L_{\delta^*}(\delta_2)$; otherwise, $L_{\delta^*}(\delta_1) < L_{\delta^*}(\delta_2)$.

The unimodality of PL requires that for $\delta_1 < \delta_2 < \delta^*$, $L_{\delta^*}(\delta_1) > L_{\delta^*}(\delta_2)$ and for $\delta^* < \delta_1 < \delta_2$, $L_{\delta^*}(\delta_1) < L_{\delta^*}(\delta_2)$. Thus, when $\delta^* < \delta_1 < \delta_2$, the unimodality of $L_{\delta^*}(\delta)$ requires that the posteriori probability should satisfy condition (41), which signifies that there exist a small number of positive objects in the objects with small posterior probabilities. When $\delta_1 < \delta_2 < \delta^*$, the unimodal of $L_{\delta^*}(\delta)$ requires that the posteriori probability should satisfy the contrary case of condition (41), which signifies that there exist a large number of positive objects in the objects with large posterior probabilities.

According to the above discussion, if the posteriori probability is sufficiently good, PL is a unimodal function of δ. The interval search method is applied to obtain δ^*. From Theorem 1, we have

$$\delta^* = \left(\frac{1}{2} - p\right)PA^* + p = \frac{1}{2} - \left(\frac{1}{2} - p\right)PL^*. \tag{42}$$

Then, we express the plug-in rule as:

$$h_r(\boldsymbol{x}) = sign\left[\widehat{\eta}(\boldsymbol{x}) - \left(\frac{1}{2} - \left(\frac{1}{2} - p\right)r\right)\right], \tag{43}$$

and apply the interval search method to finding the optimal r that minimizes $\widehat{PL}_{|\mathcal{S}_2|}(h_r(\boldsymbol{x}))$.

A fixed reduction of the interval is employed. In each iteration, the interval length is reduced by the $\tau \in (0, 0.5)$ ratio. The interval search method is thus called as τ-interval search method and the ratio τ is a parameter to be tuned. The interval search method for minimizing the PL is shown as Algorithm 2. The time complexity of the τ-interval search method contains two parts, which are learning $\widehat{\eta}(\boldsymbol{x})$ and searching δ^*. The time complexity of learning $\widehat{\eta}(\boldsymbol{x})$ is the same as the gradient descent method, and the time complexity of search δ^* is $\mathcal{O}(N \log_\tau \epsilon)$, where N is the number of training data, τ is the reduction ratio of the interval and ϵ is the threshold of the stop condition. Learning $\widehat{\eta}(\boldsymbol{x})$ is the main time consuming part. When handling large number of samples, it is suggested to utilize effective gradient descent method.

Algorithm 2 The τ-Interval Search Method for Minimizing the PL

```
Require: The training data $\mathcal{S}_N$
  Randomly split the training data $\mathcal{S}_N$ into $\mathcal{S}_1$ and $\mathcal{S}_2$ with a ratio of 8 : 2 and use $\mathcal{S}_1$ to estimate $\widehat{\eta}(\boldsymbol{x})$
  Set $\alpha = 0$, $\beta = 1$, $t = 0$,$\epsilon = 0.0001$
  Let $\lambda = \alpha + \tau(\beta - \alpha)$ and $\mu = \beta - \tau(\beta - \alpha)$,
  Obtain $h_\lambda(\boldsymbol{x}) = sign[\widehat{\eta}(\boldsymbol{x}) - (\frac{1}{2} - (\frac{1}{2} - p)\lambda)]$, $h_\mu(\boldsymbol{x}) = sign[\widehat{\eta}(\boldsymbol{x}) - (\frac{1}{2} - (\frac{1}{2} - p)\mu)]$ and calculate $\widehat{PL}_{|\mathcal{S}_2|}(h_\lambda, Y)$ and $\widehat{PL}_{|\mathcal{S}_2|}(h_\mu, Y)$ on $\mathcal{S}_2$
  while $\beta - \alpha > \epsilon$, do
    IF $\widehat{PL}_{|\mathcal{S}_2|}(h_\lambda(\boldsymbol{x})) \leq \widehat{PL}_{|\mathcal{S}_2|}(h_\mu(\boldsymbol{x}))$, THEN update $\beta = \mu$ ELSE update $\alpha = \lambda$;
    $\lambda = \alpha + \tau(\beta - \alpha)$, $\mu = \beta - \tau(\beta - \alpha)$ and calculate $\widehat{PL}_{|\mathcal{S}_2|}(h_\lambda(\boldsymbol{x}))$ and $\widehat{PL}_{|\mathcal{S}_2|}(h_\mu(\boldsymbol{x}))$ on $\mathcal{S}_2$;
    $t = t + 1$; $\delta^t = \frac{1}{2} - (\frac{1}{2} - p)\lambda$.
  end while
Ensure: The optimal threshold $\delta^*$.
```

6.3 Experiments

We validate the performance of the τ-interval search on a variety of benchmark data sets. By the benchmark data sets, we show that learning by PA is more fair in majority accuracy and minority accuracy than A and compare the τ-interval search method with some other plug-in rules to show its effectiveness.

The benchmark data sets are downloaded from the KEEL Data Set Repository (Alcalafdez et al. 2008) and the UCI Machine Learning Repository (Dua and Graff 2017). These data sets are briefly described in Table 2, including data ID, name, size, number of attributes and the imbalance ratio(IR). The posterior probability is generated by the kernel logistic regression and the kernel function is the RBF kernel $K(\boldsymbol{x},\boldsymbol{x}') = \exp(-\gamma||\boldsymbol{x} - \boldsymbol{x}'||^2)$.

Each data set is randomly divided into a training set, a validation set and a test set at a ratio of 3:1:1. The methods are compared in the same division. We randomly divide the data set 30 times to obtain an average performance. The parameter γ is chosen from $\{2^{-4}, 2^{-2}, 2^0, 2^2, 2^4, 2^6\}$ and the τ is chosen from $\{0.1, 0.2, 0.3, 0.4\}$ via the validation set. Each attribute is linearly scaled to the range [0, 1] using the maximum and minimum values in the training data. For each data, we also add 3% and 5% random uniform label noise to increase the complexity of the data.

First, to show that learning by PA is more fair than A, we compare the bias [refer to Eq. (13)] of KLR and the τ-interval method. Figure 4 shows the comparison result. Each bar of Fig. 4 is the difference of the mean bias over 30 times between KLR and the τ-interval method on each benchmark data set. As shown in Fig. 4, we observe that 16/20, 17/20, 16/20 bars are greater than zero under 0% noise, 3% noise, 5% noise, respectively. That is, the bias of KLR is large than that of the τ-interval method, which reflects the classifiers learnt by PA is more fair than the classifiers learnt by A.

Table 2 Description of data sets

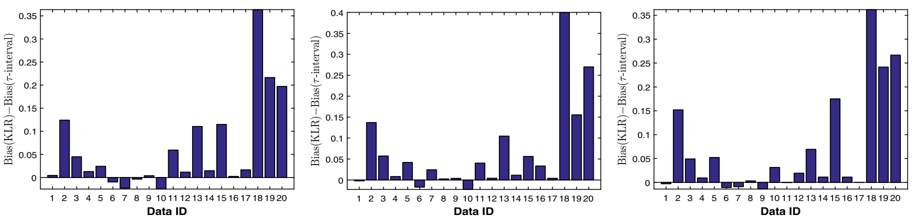
Data ID	Data name	Attribute	Instance	IR	Download
1	First-order theorem proving1	51	6118	1.02	UCI
2	First-order theorem proving2	51	6118	1.28	UCI
3	First-order theorem proving3	51	6118	1.16	UCI
4	First-order theorem proving4	51	6118	1.14	UCI
5	First-order theorem proving5	51	6118	1.27	UCI
6	Crx	15	653	1.21	KEEL
7	Heart	13	270	1.25	KEEL
8	Australian	14	690	1.25	KEEL
9	Wdbc	30	569	1.68	KEEL
10	Bands	19	365	1.70	KEEL
11	Ionosphere	33	351	1.79	KEEL
12	Wisconsin	9	683	1.86	KEEL
13	Pima	8	768	1.87	KEEL
14	Titanic	3	2201	2.10	KEEL
15	German	20	1000	2.33	KEEL
16	Segment	19	2308	6.02	KEEL
17	Dermatology	34	358	16.90	KEEL
18	Wilt	5	4839	17.54	UCI
19	Flare	11	1066	23.79	KEEL
20	Winequality-red	11	1599	29.17	KEEL

Second, to validate the performance the proposed method, the A and PA are employed as evaluation measures. The benchmark methods are KLR, p-cut (with the proportion of the minority class in $\mathcal{S}_2$ as the threshold), grid-search, $\mathcal{S}_2$-search and bisection method. The KLR aims to optimize the A, and p-cut aims to optimize the balanced accuracy. The grid-search and $\mathcal{S}_2$-search aim to optimize the PA. The bisection is used to optimize the F_1-measure and PA, which are noted as Bisection-F_1 and Bisection-PA, respectively. Tables 3, 5 and 7 show the mean and the standard deviation of A over 30 time comparisons with 0%, 3% and 5% label noise, respectively. Tables 4, 6 and 8 show the mean and the standard deviation of PA over 30 time comparisons with 0%, 3% and 5% label noise, respectively. In each row of the tables, the method with the maximal evaluation value is underlined and printed in bold type, and the method with a dot indicates that the τ-interval search is significantly better with regard to the pairwise Student's t test with a level of 0.1. As shown in Tables 3, 4, 5, 6, 7 and 8, the evaluation score obtained by the τ-interval search is highlighted in bold and is underlined in most of the comparisons. In many comparisons, the τ-interval search is statistically better than other methods.

To further analysis the statical performance of each method, for each method, we calculate the gap between the times of the significant wins and the times of significant loses. An algorithm a significantly wins b if its mean and standard deviation are satisfied:

$$\mu_a - 1.96\frac{\sigma_a}{\sqrt{t}} > \mu_b + 1.96\frac{\sigma_b}{\sqrt{t}}, \tag{44}$$

where t is the number of comparison times; otherwise, a significantly loses b (Please refer to reference Li et al. (2016) for more details). Figure 5 shows the results of the statistical comparison. Each bar in Fig. 5 represents the gap between the times of the significant wins and the times of significant loses. As shown in Fig. 5, we observe that the bar of the τ-interval search method is the highest w.r.t PA under different noise level. With respect to A, the bar of the τ-interval search is the highest when the noise level is % and 5%; and when the label is not polluted by noise, the bar of the τ-interval search is the second highest. In general, we can conclude that the τ-interval search method can optimize the PA value better and also can obtain a satisfactory A value.



(a) 0% noise **(b)** 3% noise **(c)** 5% noise

Fig. 4 Bias Gap of KLR and the τ-interval method under different noise level. Each bar is the mean gap over 30 times on each data set

Table 3 Comparison of accuracy on data without noise

ID	KLR	p-cut	Bisection-F_1	Grid-search	$\mathcal{S}_2$-search	Bisection-PA	τ-interval
1	0.669 ± 0.012•	0.670 ± 0.012	0.647 ± 0.011•	0.672 ± 0.012	0.672 ± 0.012	0.670 ± 0.013•	**0.673** ± 0.012
2	0.671 ± 0.013	0.665 ± 0.014•	**0.672** ± 0.014	0.670 ± 0.014	0.671 ± 0.014	0.668 ± 0.013	0.670 ± 0.014
3	0.683 ± 0.013•	0.681 ± 0.012•	0.670 ± 0.013•	0.684 ± 0.014	0.684 ± 0.014•	0.683 ± 0.013•	**0.688** ± 0.013
4	0.663 ± 0.011•	0.662 ± 0.013•	0.654 ± 0.011•	0.661 ± 0.012•	0.661 ± 0.012•	0.662 ± 0.011•	**0.667** ± 0.011
5	0.683 ± 0.013	0.681 ± 0.014•	0.670 ± 0.012•	0.682 ± 0.016•	0.681 ± 0.016•	0.683 ± 0.014	**0.685** ± 0.012
6	0.859 ± 0.026•	0.864 ± 0.023	0.808 ± 0.030•	0.860 ± 0.021•	0.859 ± 0.022•	0.861 ± 0.025	**0.868** ± 0.021
7	0.789 ± 0.049•	0.786 ± 0.050•	0.792 ± 0.053•	0.770 ± 0.048•	0.776 ± 0.057•	0.788 ± 0.051•	**0.817** ± 0.038
8	0.850 ± 0.030	0.850 ± 0.032	0.821 ± 0.038•	0.848 ± 0.031•	0.847 ± 0.032•	0.850 ± 0.029	**0.857** ± 0.025
9	0.954 ± 0.022•	0.955 ± 0.022•	0.952 ± 0.022•	0.955 ± 0.023•	0.952 ± 0.023•	0.955 ± 0.022•	**0.965** ± 0.017
10	0.637 ± 0.059•	0.602 ± 0.069•	0.626 ± 0.060•	0.625 ± 0.063•	0.617 ± 0.074•	0.609 ± 0.067•	**0.654** ± 0.050
11	0.846 ± 0.042•	0.845 ± 0.042•	0.846 ± 0.042•	0.846 ± 0.046•	0.845 ± 0.043•	0.846 ± 0.041•	**0.867** ± 0.032
12	0.966 ± 0.013•	0.967 ± 0.013	0.946 ± 0.019•	0.970 ± 0.014	0.969 ± 0.014	0.966 ± 0.014	**0.971** ± 0.010
13	0.763 ± 0.028	0.750 ± 0.031•	0.758 ± 0.030•	0.755 ± 0.035•	0.756 ± 0.033	0.762 ± 0.029	**0.766** ± 0.029
14	0.779 ± 0.016•	0.767 ± 0.023•	0.777 ± 0.016•	0.779 ± 0.019	0.779 ± 0.019	0.780 ± 0.016	**0.783** ± 0.017
15	0.742 ± 0.022•	0.710 ± 0.027•	0.744 ± 0.024	0.725 ± 0.027•	0.726 ± 0.028•	0.732 ± 0.026•	**0.746** ± 0.023
16	0.993 ± 0.004•	0.992 ± 0.005•	0.992 ± 0.005•	0.994 ± 0.004•	0.995 ± 0.004•	0.993 ± 0.005•	**0.996** ± 0.003
17	0.999 ± 0.004	0.998 ± 0.005•	0.996 ± 0.006•	0.997 ± 0.006•	0.991 ± 0.012•	0.999 ± 0.004	**1.000** ± 0.003
18	**0.946** ± 0.000	0.716 ± 0.066•	0.878 ± 0.063	0.743 ± 0.220•	0.757 ± 0.207•	0.855 ± 0.035	0.843 ± 0.065
19	**0.958** ± 0.007	0.800 ± 0.054•	0.947 ± 0.013	0.934 ± 0.022•	0.935 ± 0.022•	0.927 ± 0.021•	0.942 ± 0.023
20	**0.968** ± 0.002	0.748 ± 0.063•	0.945 ± 0.030	0.932 ± 0.028•	0.935 ± 0.024•	0.924 ± 0.043•	0.940 ± 0.037
Rank	3.15	5.50	4.75	4.40	4.65	4.05	1.50

Table 4 Comparison of pure accuracy on data sets without noise

ID	KLR	p-cut	Bisection-F_1	Grid-search	$\mathcal{S}_2$-search	Bisection-PA	τ-interval
1	0.339 ± 0.025•	0.341 ± 0.025	0.293 ± 0.022•	0.345 ± 0.023	0.345 ± 0.024	0.340 ± 0.025•	**0.347** ± 0.024
2	0.318 ± 0.028•	0.325 ± 0.028	0.305 ± 0.028•	0.328 ± 0.029	**0.329** ± 0.028	0.326 ± 0.026	0.327 ± 0.029
3	0.361 ± 0.029•	0.363 ± 0.024•	0.320 ± 0.027•	0.366 ± 0.026	0.365 ± 0.027	0.364 ± 0.027•	**0.373** ± 0.026
4	0.324 ± 0.022•	0.325 ± 0.025•	0.295 ± 0.023•	0.325 ± 0.024•	0.325 ± 0.024•	0.324 ± 0.023•	**0.332** ± 0.023
5	0.356 ± 0.026•	0.365 ± 0.026	0.312 ± 0.024•	0.362 ± 0.029	0.361 ± 0.029•	0.365 ± 0.025	**0.369** ± 0.022
6	0.717 ± 0.053•	0.727 ± 0.045	0.605 ± 0.064•	0.719 ± 0.042•	0.718 ± 0.043•	0.722 ± 0.050	**0.736** ± 0.041
7	0.574 ± 0.099•	0.570 ± 0.101•	0.571 ± 0.111•	0.539 ± 0.097•	0.548 ± 0.116•	0.574 ± 0.104•	**0.629** ± 0.079
8	0.697 ± 0.060	0.698 ± 0.064	0.628 ± 0.081•	0.693 ± 0.063•	0.690 ± 0.065•	0.698 ± 0.059	**0.711** ± 0.051
9	0.902 ± 0.047•	0.903 ± 0.046•	0.896 ± 0.048•	0.902 ± 0.049•	0.896 ± 0.049•	0.903 ± 0.047•	**0.924** ± 0.037
10	0.196 ± 0.115•	0.181 ± 0.132•	0.188 ± 0.118•	0.186 ± 0.123•	0.179 ± 0.118•	0.185 ± 0.132•	**0.225** ± 0.108
11	0.647 ± 0.100•	0.644 ± 0.099•	0.644 ± 0.102•	0.649 ± 0.101•	0.650 ± 0.096•	0.646 ± 0.098•	**0.698** ± 0.073
12	0.924 ± 0.030•	0.927 ± 0.029	0.877 ± 0.044•	0.935 ± 0.031	0.931 ± 0.031	0.925 ± 0.031•	**0.935** ± 0.022
13	0.449 ± 0.071•	0.466 ± 0.069	0.415 ± 0.082•	0.459 ± 0.072	0.460 ± 0.072	0.470 ± 0.072	**0.474** ± 0.066
14	0.444 ± 0.040•	0.438 ± 0.042•	0.437 ± 0.042•	0.452 ± 0.044	0.451 ± 0.043	0.450 ± 0.042	**0.458** ± 0.042
15	0.345 ± 0.061•	0.372 ± 0.049	0.344 ± 0.070•	0.365 ± 0.054•	0.367 ± 0.054•	0.376 ± 0.059	**0.384** ± 0.067
16	0.973 ± 0.017•	0.968 ± 0.021•	0.969 ± 0.022•	0.975 ± 0.017•	0.978 ± 0.015•	0.973 ± 0.018•	**0.983** ± 0.014
17	0.986 ± 0.043•	0.983 ± 0.044•	0.960 ± 0.068•	0.972 ± 0.051•	0.885 ± 0.195•	0.986 ± 0.043•	**0.996** ± 0.022
18	0.000 ± 0.000•	0.063 ± 0.041	0.033 ± 0.050•	0.057 ± 0.037	0.057 ± 0.037	0.053 ± 0.047	**0.063** ± 0.037
19	0.147 ± 0.140•	0.163 ± 0.060•	0.208 ± 0.148•	0.241 ± 0.125	0.238 ± 0.133	0.231 ± 0.107	**0.250** ± ± 0.136
20	0.004 ± 0.026•	0.051 ± 0.038•	0.110 ± 0.112	0.109 ± 0.107	**0.112** ± 0.113	0.101 ± 0.097	0.110 ± 0.106
Rank	5.15	4.15	6.20	3.60	3.90	3.85	1.15

Table 5 Comparison of accuracy on data sets with 3% noise

ID	KLR	p-cut	Bisection-F_1	Grid-search	$\mathcal{S}_2$-search	Bisection-PA	τ-interval
1	0.660 ± 0.011•	0.659 ± 0.011•	0.642 ± 0.013•	0.659 ± 0.011•	0.660 ± 0.010•	0.659 ± 0.011•	**0.663** ± 0.012
2	**0.657** ± 0.012	0.650 ± 0.012•	0.654 ± 0.011	0.656 ± 0.014	0.655 ± 0.013	0.654 ± 0.012	0.656 ± 0.013
3	0.674 ± 0.015•	0.675 ± 0.012•	0.662 ± 0.016•	0.678 ± 0.013	0.678 ± 0.013	0.674 ± 0.013•	**0.679** ± 0.013
4	0.659 ± 0.012•	0.657 ± 0.013•	0.645 ± 0.012•	0.659 ± 0.012•	0.659 ± 0.012•	0.659 ± 0.013•	**0.663** ± 0.010
5	0.667 ± 0.015	0.663 ± 0.012•	0.658 ± 0.011•	0.665 ± 0.015	0.665 ± 0.014	0.666 ± 0.013	**0.669** ± 0.013
6	0.826 ± 0.029	0.827 ± 0.027	0.782 ± 0.035•	0.830 ± 0.026	0.827 ± 0.028•	0.827 ± 0.026	**0.837** ± 0.028
7	0.770 ± 0.055•	0.765 ± 0.049•	0.767 ± 0.045•	0.764 ± 0.055•	0.760 ± 0.055•	0.766 ± 0.050•	**0.801** ± 0.043
8	0.829 ± 0.033•	0.830 ± 0.033	0.800 ± 0.035•	0.831 ± 0.028•	0.829 ± 0.028•	0.829 ± 0.032•	**0.839** ± 0.026
9	0.904 ± 0.053•	0.898 ± 0.050•	0.893 ± 0.047•	0.900 ± 0.050•	0.904 ± 0.051•	0.903 ± 0.053•	**0.930** ± 0.035
10	0.594 ± 0.054•	0.572 ± 0.067•	0.589 ± 0.061•	0.596 ± 0.052•	0.600 ± 0.051•	0.577 ± 0.067•	**0.637** ± 0.046
11	0.811 ± 0.050•	0.809 ± 0.055•	0.816 ± 0.044•	0.810 ± 0.043•	0.808 ± 0.037•	0.808 ± 0.054•	**0.836** ± 0.044
12	0.935 ± 0.020	**0.939** ± 0.019	0.911 ± 0.019•	0.937 ± 0.020	0.939 ± 0.018	0.936 ± 0.020	0.939 ± 0.019
13	0.744 ± 0.030	0.721 ± 0.029•	0.741 ± 0.029	0.719 ± 0.042•	0.721 ± 0.043•	0.736 ± 0.031•	**0.746** ± 0.034
14	0.765 ± 0.015	0.744 ± 0.020•	0.763 ± 0.015•	0.762 ± 0.019	0.762 ± 0.019•	0.764 ± 0.015•	**0.768** ± 0.014
15	0.727 ± 0.026•	0.689 ± 0.036•	0.723 ± 0.027•	0.717 ± 0.028•	0.717 ± 0.026•	0.710 ± 0.026•	**0.740** ± 0.026
16	0.944 ± 0.019	0.884 ± 0.036•	0.929 ± 0.020•	0.941 ± 0.023•	0.942 ± 0.023•	0.942 ± 0.021•	**0.947** ± 0.020
17	0.914 ± 0.042•	0.914 ± 0.042•	0.914 ± 0.043•	0.917 ± 0.040	0.920 ± 0.037	0.914 ± 0.043•	**0.930** ± 0.030
18	**0.919** ± 0.000	0.658 ± 0.058•	0.833 ± 0.056	0.682 ± 0.192•	0.686 ± 0.197•	0.794 ± 0.051	0.788 ± 0.078
19	**0.929** ± 0.007	0.708 ± 0.052•	0.893 ± 0.032•	0.875 ± 0.046•	0.877 ± 0.047•	0.855 ± 0.062•	0.908 ± 0.039
20	**0.940** ± 0.001	0.673 ± 0.041•	0.886 ± 0.048	0.847 ± 0.136•	0.847 ± 0.139	0.821 ± 0.079•	0.893 ± 0.028
Rank	2.95	5.70	5.00	4.30	4.00	4.70	1.35

Table 6 Comparison of pure accuracy on data sets with 3% noise

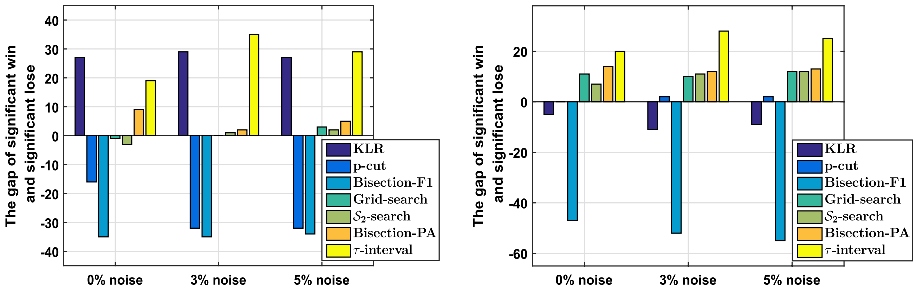
ID	KLR	p-cut	Bisection-F_1	Grid-search	$\mathcal{S}_2$-search	Bisection-PA	τ-interval
1	0.320 ± 0.022•	0.319 ± 0.022•	0.282 ± 0.025•	0.318 ± 0.022•	0.320 ± 0.019•	0.319 ± 0.021•	**0.326** ± 0.024
2	0.288 ± 0.025•	0.296 ± 0.023	0.272 ± 0.023•	**0.301** ± 0.027	0.300 ± 0.027	0.299 ± 0.024	0.299 ± 0.028
3	0.343 ± 0.031•	0.350 ± 0.025	0.304 ± 0.033•	0.353 ± 0.025	0.353 ± 0.025	0.347 ± 0.026•	**0.355** ± 0.027
4	0.316 ± 0.024•	0.316 ± 0.026•	0.277 ± 0.025•	0.319 ± 0.023	0.319 ± 0.023	0.317 ± 0.026	**0.325** ± 0.019
5	0.325 ± 0.032•	0.329 ± 0.024	0.288 ± 0.024•	0.329 ± 0.029	0.328 ± 0.028	0.331 ± 0.025	**0.335** ± 0.026
6	0.652 ± 0.059•	0.654 ± 0.056	0.551 ± 0.074•	0.659 ± 0.051	0.654 ± 0.056•	0.653 ± 0.053	**0.673** ± 0.056
7	0.536 ± 0.112•	0.528 ± 0.100•	0.522 ± 0.095•	0.522 ± 0.112•	0.513 ± 0.114•	0.528 ± 0.102•	**0.597** ± 0.085
8	0.654 ± 0.066•	0.659 ± 0.066	0.584 ± 0.077•	0.659 ± 0.056•	0.656 ± 0.055•	0.657 ± 0.065	**0.676** ± 0.053
9	0.795 ± 0.115•	0.784 ± 0.108•	0.764 ± 0.104•	0.785 ± 0.110•	0.790 ± 0.115•	0.793 ± 0.114•	**0.851** ± 0.074
10	0.108 ± 0.119•	0.122 ± 0.143•	0.114 ± 0.130•	0.124 ± 0.091•	0.130 ± 0.091•	0.126 ± 0.142•	**0.193** ± 0.095
11	0.575 ± 0.108•	0.572 ± 0.117•	0.582 ± 0.098•	0.569 ± 0.095•	0.564 ± 0.083•	0.569 ± 0.115•	**0.632** ± 0.103
12	0.859 ± 0.042	**0.869** ± 0.040	0.800 ± 0.045•	0.864 ± 0.043	0.867 ± 0.039	0.861 ± 0.043	0.868 ± 0.041
13	0.410 ± 0.069•	0.404 ± 0.065•	0.387 ± 0.073•	0.394 ± 0.072•	0.397 ± 0.073•	0.415 ± 0.071•	**0.432** ± 0.074
14	0.418 ± 0.038	0.400 ± 0.040•	0.413 ± 0.037•	0.421 ± 0.037	0.420 ± 0.037	0.420 ± 0.035	**0.428** ± 0.034
15	0.320 ± 0.062•	0.335 ± 0.069•	0.303 ± 0.064•	0.340 ± 0.053•	0.339 ± 0.050•	0.345 ± 0.054•	**0.365** ± 0.057
16	0.782 ± 0.079	0.641 ± 0.079•	0.695 ± 0.100•	0.781 ± 0.079	0.783 ± 0.079	0.783 ± 0.076	**0.800** ± 0.072
17	0.498 ± 0.169	0.500 ± 0.169	0.495 ± 0.175•	0.508 ± 0.168	0.509 ± 0.159	0.502 ± 0.169	**0.558** ± 0.141
18	0.000 ± 0.000•	0.047 ± 0.041	0.025 ± 0.048•	**0.047** ± 0.033	0.046 ± 0.032	0.035 ± 0.043	0.044 ± 0.036
19	0.046 ± 0.087•	0.100 ± 0.051•	0.131 ± 0.096	0.128 ± 0.083	0.128 ± 0.081	**0.135** ± 0.076	0.133 ± 0.108
20	0.005 ± 0.022•	0.063 ± 0.043•	0.103 ± 0.088	0.103 ± 0.080	0.096 ± 0.078•	0.100 ± 0.069	**0.121** ± 0.077
Rank	5.10	4.50	6.10	3.45	3.80	3.70	1.35

Table 7 Comparison of accuracy on data sets with 5% noise

ID	KLR	p-cut	Bisection-F_1	Grid-search	$\mathcal{S}_2$-search	Bisection-PA	τ-interval
1	0.654 ± 0.012•	0.654 ± 0.012•	0.639 ± 0.014•	0.656 ± 0.012•	0.657 ± 0.012•	0.654 ± 0.012•	**0.661** ± 0.011
2	0.650 ± 0.013	0.646 ± 0.014•	0.650 ± 0.013	0.650 ± 0.014	**0.651** ± 0.014	0.648 ± 0.015	0.651 ± 0.014
3	0.668 ± 0.017•	0.666 ± 0.017•	0.659 ± 0.016•	0.670 ± 0.018•	0.669 ± 0.018•	0.668 ± 0.017•	**0.673** ± 0.016
4	0.641 ± 0.013•	0.642 ± 0.013•	0.632 ± 0.013•	0.643 ± 0.011	0.643 ± 0.011	0.643 ± 0.012	**0.646** ± 0.011
5	0.657 ± 0.011•	0.654 ± 0.014•	0.651 ± 0.013•	0.658 ± 0.012	0.658 ± 0.012	0.657 ± 0.013•	**0.661** ± 0.013
6	0.824 ± 0.028•	0.824 ± 0.026•	0.776 ± 0.036•	0.822 ± 0.027•	0.822 ± 0.027•	0.825 ± 0.028•	**0.835** ± 0.025
7	0.743 ± 0.051•	0.741 ± 0.058•	0.734 ± 0.055•	0.728 ± 0.059•	0.728 ± 0.057•	0.740 ± 0.055•	**0.770** ± 0.048
8	0.812 ± 0.029•	0.812 ± 0.027•	0.787 ± 0.033•	0.810 ± 0.032•	0.809 ± 0.035•	0.813 ± 0.030•	**0.822** ± 0.032
9	0.888 ± 0.034•	0.885 ± 0.031•	0.853 ± 0.029•	0.886 ± 0.032•	0.888 ± 0.032•	0.889 ± 0.035•	**0.911** ± 0.024
10	0.571 ± 0.059•	0.539 ± 0.055•	0.554 ± 0.059•	0.569 ± 0.066•	0.555 ± 0.079•	0.545 ± 0.058•	**0.595** ± 0.050
11	0.763 ± 0.062•	0.759 ± 0.060•	0.766 ± 0.063•	0.762 ± 0.067•	0.765 ± 0.062•	0.760 ± 0.063•	**0.802** ± 0.048
12	0.912 ± 0.023•	0.915 ± 0.025	0.880 ± 0.027•	0.914 ± 0.034	0.914 ± 0.031	0.914 ± 0.023	**0.919** ± 0.022
13	**0.731** ± 0.026	0.703 ± 0.031•	0.731 ± 0.030	0.712 ± 0.030•	0.712 ± 0.031•	0.717 ± 0.025•	0.726 ± 0.031
14	0.750 ± 0.017	0.732 ± 0.022•	0.748 ± 0.017•	0.750 ± 0.018	0.750 ± 0.018	0.750 ± 0.017•	**0.754** ± 0.017
15	0.719 ± 0.022	0.686 ± 0.027•	0.717 ± 0.023	0.702 ± 0.030•	0.702 ± 0.030•	0.704 ± 0.024•	**0.721** ± 0.028
16	**0.935** ± 0.012	0.868 ± 0.025•	0.915 ± 0.009•	0.933 ± 0.015	0.934 ± 0.015	0.932 ± 0.016	0.934 ± 0.013
17	0.885 ± 0.037	0.880 ± 0.043	0.885 ± 0.037	**0.888** ± 0.039	0.876 ± 0.049	0.885 ± 0.037	0.885 ± 0.037
18	**0.902** ± 0.000	0.650 ± 0.054•	0.803 ± 0.076	0.631 ± 0.189•	0.624 ± 0.189•	0.763 ± 0.049•	0.794 ± 0.080
19	**0.912** ± 0.008	0.695 ± 0.036•	0.890 ± 0.028	0.870 ± 0.040•	0.869 ± 0.047•	0.845 ± 0.062•	0.884 ± 0.027
20	**0.921** ± 0.002	0.629 ± 0.042•	0.843 ± 0.054	0.817 ± 0.110	0.819 ± 0.109	0.766 ± 0.086•	0.836 ± 0.079
Rank	3.05	5.70	4.90	4.25	4.10	4.35	1.65

Table 8 Comparison of pure accuracy on data sets with 5% noise

ID	KLR	p-cut	Bisection-F_1	Grid-search	$\mathcal{S}_2$-search	Bisection-PA	τ-interval
1	0.308 ± 0.024•	0.309 ± 0.024•	0.275 ± 0.029•	0.313 ± 0.023•	0.314 ± 0.023•	0.308 ± 0.024•	**0.322** ± 0.023
2	0.276 ± 0.028•	0.288 ± 0.027	0.267 ± 0.027•	0.290 ± 0.028	**0.290** ± 0.028	0.286 ± 0.030	0.290 ± 0.028
3	0.332 ± 0.035•	0.332 ± 0.033•	0.300 ± 0.032•	0.337 ± 0.035•	0.336 ± 0.036•	0.335 ± 0.035•	**0.345** ± 0.033
4	0.279 ± 0.026•	0.286 ± 0.026	0.252 ± 0.025•	0.289 ± 0.021	0.289 ± 0.021	0.287 ± 0.024	**0.291** ± 0.022
5	0.305 ± 0.023•	0.311 ± 0.027•	0.274 ± 0.028•	0.314 ± 0.022	0.314 ± 0.021	0.312 ± 0.026•	**0.319** ± 0.027
6	0.648 ± 0.057•	0.647 ± 0.053•	0.539 ± 0.075•	0.643 ± 0.054•	0.644 ± 0.054•	0.649 ± 0.058•	**0.670** ± 0.049
7	0.478 ± 0.105•	0.476 ± 0.117•	0.451 ± 0.118•	0.449 ± 0.119•	0.446 ± 0.117•	0.474 ± 0.112•	**0.534** ± 0.097
8	0.620 ± 0.058•	0.621 ± 0.053•	0.559 ± 0.069•	0.617 ± 0.067•	0.614 ± 0.072•	0.622 ± 0.060•	**0.642** ± 0.064
9	0.761 ± 0.071•	0.758 ± 0.065•	0.672 ± 0.068•	0.759 ± 0.068•	0.763 ± 0.068•	0.764 ± 0.073•	**0.807** ± 0.054
10	0.051 ± 0.127•	0.044 ± 0.113•	0.049 ± 0.117•	0.064 ± 0.094•	0.059 ± 0.096•	0.050 ± 0.117•	**0.111** ± 0.094
11	0.471 ± 0.136•	0.464 ± 0.132•	0.474 ± 0.139•	0.469 ± 0.145•	0.468 ± 0.139•	0.466 ± 0.138•	**0.552** ± 0.112
12	0.807 ± 0.052•	0.816 ± 0.053	0.728 ± 0.064•	0.816 ± 0.071	0.815 ± 0.066	0.812 ± 0.050•	**0.824** ± 0.049
13	0.391 ± 0.062	0.379 ± 0.063	0.376 ± 0.079	0.382 ± 0.060	0.381 ± 0.063	0.387 ± 0.052	**0.394** ± 0.066
14	0.389 ± 0.042	0.373 ± 0.042•	0.385 ± 0.041•	0.393 ± 0.044	0.393 ± 0.044	0.393 ± 0.040	**0.400** ± 0.042
15	0.306 ± 0.057•	0.331 ± 0.055•	0.303 ± 0.066•	0.320 ± 0.066•	0.319 ± 0.071•	0.339 ± 0.051	**0.355** ± 0.060
16	**0.759** ± 0.049	0.598 ± 0.055•	0.655 ± 0.042•	0.754 ± 0.054	0.755 ± 0.055	0.752 ± 0.053	0.758 ± 0.047
17	0.408 ± 0.148	0.402 ± 0.151	0.405 ± 0.146	**0.411** ± 0.147	0.392 ± 0.155	0.410 ± 0.143	0.408 ± 0.148
18	-0.000 ± 0.000•	0.043 ± 0.053	0.043 ± 0.055	0.049 ± 0.032	0.047 ± 0.030	0.040 ± 0.055	**0.053** ± 0.035
19	0.069 ± 0.092•	0.108 ± 0.063•	0.161 ± 0.094	0.159 ± 0.098	0.150 ± 0.105•	0.163 ± 0.104•	**0.195** ± 0.110
20	0.003 ± 0.015•	0.051 ± 0.036	**0.061** ± 0.062	0.043 ± 0.057	0.041 ± 0.058	0.058 ± 0.048	0.046 ± 0.054
Rank	4.70	4.95	5.75	3.45	4.10	3.60	1.45



(a) A **(b)** PA

Fig. 5 Statistical comparison under different noise level. Each bar is the gap between the times of the significant wins and the times of significant loses. The significant lose and win is defined according to Eq. (44)

7 Conclusion

With an increase in the complexity of the data, eliminating random consistency from learning algorithms has great potential to improve the generalization ability. In this paper, first, we have shown that the PA is insensitive to the class distribution of classifiers in evaluation and is more fairer than the A in learning classifiers through two vivid examples. Second, we have given some novel bounds to show that learning by PA can approach to the optimal A and have shown that the empirical risk minimization process of the PA is Bayes-risk consistent. Based on these theoretical guarantees, we have proposed a plug-in rule model that optimizes the PA. The experimental results have shown the fairness and effectiveness of learning by PA. An interesting future work is to establish the other strategies to define the random consistency. An analysis of the random consistency for each instance maybe a promising direction.

Acknowledgements This work is supported by National Key R&D Program of China (No. 2018YFB1004300), the National Natural Science Foundation of China (Nos. 61672332, 61872226, 61976129), the Overseas Returnee Research Program of Shanxi Province (Nos. 2017023, 2018172), the Natural Science Foundation of Shanxi Province (No. 201701D121052), and Program for the San Jin Young Scholars of Shanxi.

Appendix: Proofs

Lemma 1 *When the partitions in $\mathcal{H}^{q(h)}$ are distributed uniformly, the expectation accuracy of partitions in $\mathcal{H}^{q(h)}$ is:*

$$\mathbb{E}_{h' \in \mathcal{H}^{q(h)}} A(h') = pq(h) + (1-p)(1-q(h)). \tag{45}$$

Proof Without loss of generality, we assume that $q(h) < p$. Assuming that the size of data is N, we have

$$\mathbb{P}_{h' \in \mathcal{H}^{q(h)}} \left(TP(h') = \frac{j}{N} \right) = \frac{C_{Np}^{j} C_{N-Np}^{Nq(h)-j}}{C_{N}^{Nq(h)}}, \tag{46}$$

where $j = 0, ..., Nq(h)$ and C_n^m is the number of combinations of n items taken m at a time. From (46), we know that $N \cdot TP(h')$ follows the hypergeometric distribution with the size of the population selected from be N, Np elements of the population belonging to one group and $N - Np$ belonging to the other group, and the number of samples drawn from the population be $Nq(h)$. Thus,

$$\mathbb{E}_{h' \in \mathcal{H}^{q(h)}} TP(h') = pq(h). \tag{47}$$

Then, according to $TN(h') = 1 - p - \big(q(h) - TP(h')\big)$, we have:

$$\begin{aligned} \mathbb{E}_{h' \in \mathcal{H}^{q(h)}} A(h') &= \mathbb{E}_{h' \in \mathcal{H}^{q(h)}} 1 - p - q(h) + 2TP(h') \\ &= 1 - p - q(h) + 2pq(h). \end{aligned} \tag{48}$$

□

Example 2 Assume that two-class data are generated from two Gaussian distributions with uncommon means $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, but a common covariance Σ:

$$\boldsymbol{m}(\boldsymbol{x}|y = +1) = \mathcal{N}(\boldsymbol{\mu_1}, \Sigma), \tag{49}$$

$$\boldsymbol{m}(\boldsymbol{x}|y = -1) = \mathcal{N}(\boldsymbol{\mu_2}, \Sigma) \tag{50}$$

and the probability of the positive class is $p = \mathbb{P}(Y = +1)$. The label of the minority class is corrupted by the instance-independent noise at the level s_1: $\mathbb{P}(\widetilde{Y} = -1|Y = +1) = s_1$. For this learning task, the bias of h_A^* is:

$$\begin{aligned} &Bias(h_A^*) \\ &= \Big|\mathbb{P}\big(d(X) < d_0|Y = -1\big) - \mathbb{P}\big(d(X) > d_0|Y = +1\big)\Big| \end{aligned} \tag{51}$$

$$= \left|(\boldsymbol{\Phi}\left(\frac{d_0 + \Delta/2}{\sqrt{\Delta}}\right) - 1 + \boldsymbol{\Phi}\left(\frac{d_0 - \Delta/2}{\sqrt{\Delta}}\right)\right|, \tag{52}$$

where $\boldsymbol{\Phi}(\bullet)$ is the cumulative distribution function of the standard normal distribution, $\Delta = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})' \Sigma^{-} (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ and $d_0 = \ln \frac{1-p}{p} \frac{1}{1-2s_1}$.

Proof According to Lemma 2, the corrupted conditional class probability $\mathbb{P}(\widetilde{Y} = +1|X = \boldsymbol{x})$ is needed. Based on the Bayes' theorem:

$$\mathbb{P}(\widetilde{Y} = +1|X = \boldsymbol{x}) \tag{53}$$

$$= \frac{\boldsymbol{m}(\boldsymbol{x}|\widetilde{y}=+1)\mathbb{P}(\widetilde{Y}=+1)}{\sum_{l\in\{-1,+1\}}\boldsymbol{m}(\boldsymbol{x}|\widetilde{y}=l)\mathbb{P}(\widetilde{Y}=l)}. \tag{54}$$

Because $\mathbb{P}(\widetilde{Y}=+1)=p(1-s_1)$, $\boldsymbol{m}(\boldsymbol{x}|\widetilde{y}=+1)=\mathcal{N}(\boldsymbol{\mu_1},\Sigma)$ and

$$\boldsymbol{m}(\boldsymbol{x}|\widetilde{y}=-1) \\ = \sum_{l\in\{+1,-1\}}\boldsymbol{m}(\boldsymbol{x}|y=l,\widetilde{y}=-1)\mathbb{P}(Y=l|\widetilde{Y}=-1) \tag{55}$$

$$= \sum_{l\in\{+1,-1\}}\boldsymbol{m}(\boldsymbol{x}|y=l)\frac{\mathbb{P}(\widetilde{Y}=+1|Y=l)\mathbb{P}(Y=l)}{\mathbb{P}(\widetilde{Y}=-1)} \tag{56}$$

$$= \frac{(1-p)}{(1-p)+ps_1}\mathcal{N}(\boldsymbol{\mu_2},\Sigma)+\frac{ps_1}{(1-p)+ps_1}\mathcal{N}(\boldsymbol{\mu_1},\Sigma), \tag{57}$$

where $\boldsymbol{m}(\boldsymbol{x}|y=l,\widetilde{y}=-1)=\boldsymbol{m}(\boldsymbol{x}|y=l)$ is satisfied because the label noise is independent on instance: $\mathbb{P}(\widetilde{Y}=-1|Y=l,X=\boldsymbol{x})=\mathbb{P}(\widetilde{Y}=-1|Y=l)$, we have:

$$\mathbb{P}(\widetilde{Y}=+1|X=\boldsymbol{x})=\frac{1-s_1}{1+\exp(\boldsymbol{w}^T\boldsymbol{x}+b)} \tag{58}$$

with $\boldsymbol{w}^T=(\boldsymbol{\mu_2}-\boldsymbol{\mu_1})^T\Sigma^{-1}$ and $b=\frac{1}{2}\boldsymbol{\mu_1}^T\Sigma^{-1}\boldsymbol{\mu_1}-\ln p-\frac{1}{2}\boldsymbol{\mu_2}^T\Sigma^{-1}\boldsymbol{\mu_2}+\ln(1-p)$.

Further, according to Lemma 2, the optimal classifier in the sense of accuracy is

$$h_A^*(\boldsymbol{x})=\begin{cases}+1, & d(\boldsymbol{x})>d_0,\\ -1, & otherwise.\end{cases} \tag{59}$$

where $d(\boldsymbol{x})=\boldsymbol{x}'\Sigma^-(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})-\frac{1}{2}(\boldsymbol{\mu_1}+\boldsymbol{\mu_2})'\Sigma^-(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$ and $d_0=\ln\frac{1-p}{p}\frac{1}{1-2s_1}$.

According to the additivity of the Gaussian distribution, we obtain the probability mass function of $d(\boldsymbol{x})$:

$$\boldsymbol{m}(d(\boldsymbol{x})|y=+1)=\mathcal{N}(\Delta/2,\Delta), \tag{60}$$

$$\boldsymbol{m}(d(\boldsymbol{x})|y=-1)=\mathcal{N}(-\Delta/2,\Delta) \tag{61}$$

where $\Delta=(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})'\Sigma^-(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$. Then

$$Bias(h_A^*) \\ = \Big|\mathbb{P}\big(d(X)<d_0|Y=-1\big)-\mathbb{P}\big(d(X)>d_0|Y=+1\big)\Big| \tag{62}$$

$$= \left|(\boldsymbol{\Phi}\left(\frac{d_0+\Delta/2}{\sqrt{\Delta}}\right)-1+\boldsymbol{\Phi}\left(\frac{d_0-\Delta/2}{\sqrt{\Delta}}\right)\right|, \tag{63}$$

where $\boldsymbol{\Phi}(\bullet)$ is the cumulative distribution function of the standard normal distribution.

□

Theorem 1 *The classifier that maximizes the PA is*

$$h^*_{PA}(\boldsymbol{x}) = \arg\max_h PA(h) \tag{64}$$

$$= \begin{cases} +1, & \eta(\boldsymbol{x}) > (\frac{1}{2} - p)PA^* + p, \\ -1, & otherwise. \end{cases} \tag{65}$$

where $PA^* = PA(h^*_{PA})$ *and* $p = \mathbb{P}(Y = +1)$.

Proof The formulation of the pure accuracy measure is fractional, which hinders obtaining the optimal classifier. Here, we resort to the cost-sensitive loss to obtain a non-closed-form solution. We begin this proof with two existing definitions and two lemmas:

Definition 5 (Kotlowski and Dembczynski 2017) We refer to a measure as a linear-fractional performance measure if it is non-increasing with FP, FN and formalized as

$$\Psi(FP, FN) = \frac{a_0 + a_1 FP + a_2 FN}{b_0 + b_1 FP + b_2 FN}, \tag{66}$$

where $a_0, a_1, a_2, b_0, b_1, b_2 \in \mathcal{R}$ and $b_0 + b_1 FP + b_2 FN \geq C_1 > 0$.

Definition 6 (Elkan 2001) The cost-sensitive loss is defined as $L_\rho(h) = \rho FP(h) + (1 - \rho)FN(h)$, where $\rho \in (0, 1)$.

Lemma 7 (Kotlowski and Dembczynski 2017) *The regret w.r.t the linear-fractional performance measure* $\Psi(FP, FN)$ *can be bounded by that w.r.t.* $L_\rho(h)$ *when* $\rho = \frac{\Psi^* b_1 - a_1}{\Psi^*(b_1+b_2)-(a_1+a_2)}$

$$\Psi^* - \Psi(h) \leq C_2(L_\rho(h) - L^*_\rho), \tag{67}$$

where $\Psi^* = \max_h \Psi(h)$, $L^*_\rho = \min_h L_\rho(h)$ *and* $C_2 = \frac{1}{C_1}\big(\Psi^*(b_1 + b_2) - (a_1 + a_2)\big)$.

Lemma 8 (Elkan 2001) *The classifier that minimizes* L_ρ *is*

$$h^*_{L_\rho}(\mathbf{x}) = \begin{cases} +1, & \eta(\mathbf{x}) > \rho, \\ -1, & otherwise. \end{cases} \tag{68}$$

where $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 | X = \boldsymbol{x})$.

Because $A = 1 - FN - FP$, $RA = 1 - p - q(h) + 2pq$ and $q(h) = p + FP - FN$, we have

$$PA = \frac{A - RA}{1 - RA} = \frac{p(1-p) - pFP - (1-p)FN}{p(1-p) + (\frac{1}{2} - p)(FP - FN)}. \tag{69}$$

According to Lemma 7, the regret of the PA can be bounded by that of L_ρ with $\rho = (\frac{1}{2} - p)PA^* + p$. Then by Lemma 8, we obtain the formulation. □

Lemma 3 *For all distributions, the plug-in rule with* ρ *as the decision threshold*

$$h_\rho(\boldsymbol{x}) = \begin{cases} +1, & \eta(\boldsymbol{x}) > \rho, \quad where \quad \rho \in (0, \frac{1}{2}], \\ -1, & otherwise, \end{cases} \tag{70}$$

satisfies:

$$L(h_\rho) \leq \frac{1-\rho}{\rho} L^*, \tag{71}$$

when $\rho = 1/2$, the equality holds.

Proof

$$L(h_\rho) = \mathbb{P}(h_\rho(X) = -1, Y = +1) + \mathbb{P}(h_\rho(X) = +1, Y = -1) \tag{72}$$

$$= \mathbb{E}_{X:\eta(X)<\rho}\eta(X) + \mathbb{E}_{X:\eta(X)\geq\rho}(1-\eta(X)) \tag{73}$$

$$\leq \mathbb{E}_{X:\eta(X)<\rho}(1/\rho - 1)\eta(X) + \mathbb{E}_{X:\eta(X)\geq\rho}(1-\eta(X)) \tag{74}$$

$$= \mathbb{E}_X \min\{(1/\rho - 1)\eta(X), 1-\eta(X)\} \tag{75}$$

$$\leq (1/\rho - 1)\mathbb{E}_X \min\{\eta(X), 1-\eta(X)\}. \tag{76}$$

□

Lemma 4 *For all distributions, suppose that $p = \mathbb{P}(Y = +1) \leq \frac{1}{2}$, the pure loss of h_A^* satisfies:*

$$PL(h_A^*) \leq \frac{L^*}{p\left(\frac{3}{2} - p\right) - L^*\left(\frac{1}{2} - p\right)}. \tag{77}$$

Proof Let $q_A^* = \mathbb{P}(h_A^* = +1)$, $FP_A^* = \mathbb{P}(h_A^* = +1, Y = -1)$ and $FN_A^* = \mathbb{P}(h_A^* = -1, Y = +1)$. By definition,

$$PL(h_A^*) = \frac{L^*}{p + (1-2p)q_A^*}. \tag{78}$$

To obtain the upper bound of $PL(h_A^*)$, we derive the lower bound of q_A^*. Because:

$$L^* = \mathbb{E}_{X:\eta(X)\leq 1/2}\eta(X) + \mathbb{E}_{X:\eta(X)>1/2}(1-\eta(X)) \tag{79}$$

$$= \mathbb{E}_X\eta(X) - \mathbb{E}_{X:\eta(X)>1/2}\eta(X) + \mathbb{E}_{X:\eta(X)>1/2}(1-\eta(X)) \tag{80}$$

$$= p - \mathbb{E}_X \max\{2\eta(X) - 1, 0\}, \tag{81}$$

and then, we have

$$q_A^* = \mathbb{E}_X\mathbf{I}\{\eta(X) - 1/2 \geq 0\} \tag{82}$$

$$\geq \mathbb{E}_X \max\{\eta(X) - 1/2, 0\} \tag{83}$$

$$= \frac{1}{2}(p - L^*), \tag{84}$$

where $\mathbf{I}\{\bullet\}$ is the indicator function. Putting the lower bound of q_A^* into the formulation of $PL(h_A^*)$, we obtain the upper bound of $PL(h_A^*)$. □

Theorem 3 For all distributions, suppose $p \le \frac{1}{2}$, the pure loss of h_A^* satisfies:

$$PL(h_{PA}^*) \le PL(h_A^*) \tag{85}$$

$$\le \frac{2(1-p)}{p(3-2p) - L^*(1-2p)} PL(h_{PA}^*). \tag{86}$$

Proof For any $q(h)$, we have

$$1 - RA = p + (1-2p)q(h) \le 1 - p, \tag{87}$$

hence

$$L = (1 - RA)PL \le (1-p)PL. \tag{88}$$

Further amplifying the upper bound in Lemma 4:

$$L^* \le L(h_{PA}^*) \le (1-p)PL(h_{PA}^*), \tag{89}$$

we obtain the result. □

Lemma 5 *For two random variables* $Z_1, Z_2 \in [0,1]$, *any* $\varepsilon \in (0,1]$, let $\alpha = \mathbb{E}Z_1\mathbb{E}Z_2/(2\mathbb{E}Z_1 + \mathbb{E}Z_2)$, *we have*

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{Z_1}{Z_2} - \frac{\mathbb{E}Z_1}{\mathbb{E}Z_2}\right| > \varepsilon\right)\\ &\quad\le \mathbb{P}\big(|Z_1 - \mathbb{E}Z_1| > \alpha\varepsilon\big) + 3\mathbb{P}\big(|Z_2 - \mathbb{E}Z_2| > \alpha\varepsilon\big).\end{aligned} \tag{90}$$

Proof For $\beta \in [0,1]$ and $\gamma > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{Z_1}{Z_2} - \frac{\mathbb{E}Z_1}{\mathbb{E}Z_2}\right| > \varepsilon\right)\\ &= \mathbb{P}\left(\left|\frac{Z_1 - \mathbb{E}Z_1}{(Z_2 - \mathbb{E}Z_2) + \mathbb{E}Z_2} + \frac{(\mathbb{E}Z_2 - Z_2)\mathbb{E}Z_1}{(Z_2 - \mathbb{E}Z_2)\mathbb{E}Z_2 + (\mathbb{E}Z_2)^2}\right| > \varepsilon\right)\end{aligned} \tag{91}$$

$$\le \mathbb{P}\left(\left|\frac{Z_1 - \mathbb{E}Z_1}{(Z_2 - \mathbb{E}Z_2) + \mathbb{E}Z_2}\right| > \beta\varepsilon\right) + \mathbb{P}\left(\left|\frac{(\mathbb{E}Z_2 - Z_2)\mathbb{E}Z_1}{(Z_2 - \mathbb{E}Z_2)\mathbb{E}Z_2 + (\mathbb{E}Z_2)^2}\right| > (1-\beta)\varepsilon\right) \tag{92}$$

$$\begin{aligned}&\le \mathbb{P}\big(|Z_1 - \mathbb{E}Z_1| > \beta|\mathbb{E}Z_2 - \gamma\varepsilon|\varepsilon\big) + 2\mathbb{P}\big(|Z_2 - \mathbb{E}Z_2| > \gamma\varepsilon\big)\\ &\quad + \mathbb{P}\left(\big|Z_2 - \mathbb{E}Z_2\big| > \frac{(1-\beta)\mathbb{E}Z_1}{\mathbb{E}Z_2}\big|\mathbb{E}Z_2 - \gamma\varepsilon\big|\varepsilon\right),\end{aligned} \tag{93}$$

where the first inequality is obtained by

$$\mathbb{P}(|a+b| > \varepsilon) \le \mathbb{P}(|a| > \beta\varepsilon) + \mathbb{P}(|b| > (1-\beta)\varepsilon), \tag{94}$$

and the second inequality is obtained by

$$\mathbb{P}(B_1) = \mathbb{P}(B_1|B_2)\mathbb{P}(B_2) + \mathbb{P}(B_1|B_2^c)\mathbb{P}(B_2^c) \tag{95}$$

$$\leq \mathbb{P}(B_1|B_2) + \mathbb{P}(B_2^c), \tag{96}$$

for any events B_1, B_2, where B_2^c complementary set of B_2. Here, we take the event $|Z_2 - \mathbb{E}Z_2| \leq \gamma\varepsilon$ as B_2 to divide the two terms of the first inequality.

Let

$$\beta = \mathbb{E}Z_1/(\mathbb{E}Z_1 + \mathbb{E}Z_2), \tag{97}$$

$$\gamma = \mathbb{E}Z_1\mathbb{E}Z_2/(\mathbb{E}Z_1 + \varepsilon\mathbb{E}Z_1 + \mathbb{E}Z_2), \tag{98}$$

we have

$$\begin{aligned}&\beta|\mathbb{E}Z_2 - \gamma\varepsilon|\\ &= (1-\beta)\mathbb{E}Z_1|\mathbb{E}Z_2 - \gamma\varepsilon|/\mathbb{E}Z_2\end{aligned} \tag{99}$$

$$= \gamma. \tag{100}$$

Finally, by the assumption $\varepsilon \leq 1$, we have $\gamma \geq \alpha$ and then get the result. □

Theorem 4 *Suppose the cardinality of $\mathcal{H}$ is finite: $|\mathcal{H}| < \infty$, then for every $h \in \mathcal{H}$, any $\varepsilon \in (0, 1]$, we have*

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\} \leq 8|\mathcal{H}|\exp\left\{-2N\left(\alpha\varepsilon - \frac{Rc(\mathcal{H})}{2}\right)^2\right\}, \tag{101}$$

where $\alpha = \min_{h\in\mathcal{H}} \frac{L(h)}{2PL(h)+1}$ and $Rc(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$.

Proof First, we process the superior limit in probability according to the union bound:

$$\begin{aligned}&\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\}\\ &= \mathbb{P}\left\{\exists h \in \mathcal{H} : \left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\}\\ &\leq \sum_{h\in\mathcal{H}}\mathbb{P}\left\{\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\}\\ &\leq |\mathcal{H}|\sup_{h\in\mathcal{H}}\mathbb{P}\left\{\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\}.\end{aligned} \tag{102}$$

Second, we transform the gap in the sense of PL into that of L by Lemma 5. For every $h \in \mathcal{H}$, let $\alpha = \min_{h\in\mathcal{H}} \frac{L(h)}{2PL(h)+1}$, we have:

$$\begin{aligned}&\mathbb{P}\left\{\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\}\\ &\leq \mathbb{P}\left\{\left|\widehat{L}_N - L\right| > \alpha\varepsilon\right\} + 3\mathbb{P}\left\{\left|\widehat{RA}_N - RA\right| > \alpha\varepsilon\right\}.\end{aligned} \tag{103}$$

Third, applying Theorem 8 in Bartlett and Mendelson (2003), for every $h \in \mathcal{H}$, with probability at least $1 - \delta/4$, we obtain that:

$$\left|\widehat{L}_N - L\right| \le \frac{Rc(\mathcal{H})}{2} + \sqrt{\frac{\ln(8/\delta)}{2N}}. \tag{104}$$

Let $\delta = 8\exp\left\{-2N(\alpha\varepsilon - Rc(\mathcal{H})/2)^2\right\}$, and then:

$$\mathbb{P}\left\{\left|\widehat{L}_N - L\right| > \alpha\varepsilon\right\} \le \delta/4. \tag{105}$$

For the second term in (103), by $|\mathcal{H}^{q(h)}| = C_N^{Nq(h)}$ and the triangle inequality, we have:

$$\left|\widehat{RA}_N(h) - RA(h)\right| \le \frac{1}{C_N^{Nq(h)}} \sum_{j=1}^{C_N^{Nq(h)}} \left|\widehat{L}_N(h_j) - L(h_j)\right|. \tag{106}$$

According to Theorem 8 in Bartlett and Mendelson (2003), for every function $h_j \in \mathcal{H}^{q(h)}$, with probability at least $1 - \delta/4$, holds that:

$$\left|\widehat{L}_N(h_j) - L(h_j)\right| \le \frac{Rc(\mathcal{H}^{q(h)})}{2} + \sqrt{\frac{\ln(8/\delta)}{2N}}, \tag{107}$$

because $\mathcal{H}^{q(h)} \subseteq \mathcal{H}$, we have $Rc(\mathcal{H}^{q(h)}) \le Rc(\mathcal{H})$, and then for every function $h_j \in \mathcal{H}^{q(h)}$, with probability at least $1 - \delta/4$, holds that:

$$\left|\widehat{RA}_N(h) - RA(h)\right| \le \frac{Rc(\mathcal{H})}{2} + \sqrt{\frac{\ln(8/\delta)}{2N}}. \tag{108}$$

Putting δ into inequality (108), we obtain for every $h \in \mathcal{H}$:

$$\mathbb{P}\left\{\left|\widehat{RA}_N(h) - RA(h)\right| > \alpha\varepsilon\right\} \le \delta/4. \tag{109}$$

Thus, combining (102), (103), (105)and (109), we obtain the final result. □

Lemma 6 *Let $\mathcal{S}'_N = \{(\boldsymbol{x}'_1, y'_1), \ldots, (\boldsymbol{x}'_N, y'_N)\}$ be an independent and identically distributed collection as $\mathcal{S}_N$ and $\widehat{PL}'_N(h)$ is the corresponding empirical pure loss. Suppose $N \ge 5(6 + 4\alpha\varepsilon)\alpha^{-2}\varepsilon^{-2}$, where $\alpha = \min_{h\in\mathcal{H}} \frac{L(h)}{2PL(h)+1}$, $\varepsilon \in (0, 1]$, then we have*

$$\begin{aligned} &\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\} \\ &\quad \le 2\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - \widehat{PL}'_N(h)\right| > \frac{\varepsilon}{2}\right\}. \end{aligned} \tag{110}$$

Proof There exists at least one function $h_0 \in \mathcal{H}$ satisfies $\left|\widehat{PL}_N(h_0) - PL(h_0)\right| \ge \varepsilon$. For h_0,

$$\begin{aligned} &\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - \widehat{PL}'_N(h)\right| > \frac{\varepsilon}{2}\right\} \\ &\quad \ge \mathbb{E}_{\mathcal{S}_N}\left[\mathbf{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right)\mathbb{P}\left\{\left|\widehat{PL}'_N(h_0) - PL(h_0)\right| < \frac{\varepsilon}{2}\middle|\mathcal{S}_N\right\}\right]. \end{aligned} \tag{111}$$

Here, we omit the detail proof of this inequality because the technique is the same as Lemma 2 in Vapnik and Chervonenkis (1971) on accuracy.

According to Lemma 5, let $\alpha = \min_{h\in\mathcal{H}} \frac{L(h)}{2PL(h)+1}$, we have:

$$\begin{aligned}&\mathbb{P}\left\{\left|\widehat{PL}'_N(h_0) - PL(h_0)\right| > \frac{\varepsilon}{2}\middle|\mathcal{S}_N\right\}\\ &\quad\leq \mathbb{P}\left\{\left|\widehat{A}'_N(h_0) - A(h_0)\right| > \frac{\alpha\varepsilon}{2}\middle|\mathcal{S}_N\right\} + 3\mathbb{P}\left\{\left|\widehat{RA}'_N(h_0) - RA(h_0)\right| > \frac{\alpha\varepsilon}{2}\middle|\mathcal{S}_N\right\}.\end{aligned} \tag{112}$$

For the first term of (112), according to the Bernstein's inequality, we have

$$\begin{aligned}&\mathbb{P}\left\{\left|\widehat{A}'_N(h_0) - A(h_0)\right| > \frac{\alpha\varepsilon}{2}\middle|\mathcal{S}_N\right\}\\ &\leq 2\exp\left\{-\frac{\frac{\alpha^2\varepsilon^2N}{4}}{2\left(A(h_0)(1-A(h_0)) + \frac{\alpha\varepsilon}{6}\right)}\right\}\\ &\leq 2\exp\left\{-\frac{3\alpha^2\varepsilon^2N}{6+4\alpha\varepsilon}\right\}\\ &\leq 2\left\{1 + \frac{3\alpha^2\varepsilon^2N}{6+4\alpha\varepsilon}\right\}^{-1} \leq \frac{1}{8},\end{aligned} \tag{113}$$

where the second inequality is because for any $\rho \in [0,1]$, it is satisfied that $\rho(1-\rho) \leq 1/4$, the third inequality is obtained by $e^{-x} \leq (1+x)^{-1}$ for $x > 0$ and the last inequality is obtained by the assumption $N \geq 5(6+4\alpha\varepsilon)\alpha^{-2}\varepsilon^{-2}$.

For the second term of (112), by the definition of $\widehat{RA}'_N(h_0)$, the only difference in the proof of the two terms in (112) is the number of terms for summation. Under the assumption on N, we have $NC_N^{Nq(h)} \geq 5(6+4\alpha\varepsilon)\alpha^{-2}\varepsilon^{-2}$, and then using the same technique as (113), we have

$$\mathbb{P}\left\{\left|\widehat{RA}'_N(h_0) - RA(h_0)\right| > \frac{\alpha\varepsilon}{2}\middle|\mathcal{S}_N\right\} \leq \frac{1}{8}. \tag{114}$$

Combing the inequalities (112), (113), (114), we have

$$\mathbb{P}\left\{\left|\widehat{PL}'_N(h_0) - \widehat{PL}(h_0)\right| > \frac{\varepsilon}{2}\middle|\mathcal{S}_N\right\} \leq \frac{1}{2}. \tag{115}$$

Thus, according to (111) and (115), we obtain the final result. □

Theorem 5 *As the same condition as Lemma 6 and suppose the VC dimension of $\mathcal{H}$ is finite: $d_{vc}(\mathcal{H}) < \infty$, we have*

$$\begin{aligned}&\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h) - PL(h)\right| > \varepsilon\right\}\\ &\quad\leq 4(N+1)\exp\left\{-\left(\frac{\varepsilon^2(1-|2\hat{p}_N - 1|)^2}{16} - \frac{d_{vc}(\mathcal{H})\ln(2eN/d_{vc}(\mathcal{H}))}{N}\right)N\right\}.\end{aligned} \tag{116}$$

Proof By Lemma 6,

$$\begin{aligned}&\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h)-PL(h)\right|>\varepsilon\right\}\\&\leq 2\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h)-\widehat{PL}'_N(h)\right|>\frac{\varepsilon}{2}\right\}.\end{aligned}\tag{117}$$

We divide the hypothesis space $\mathcal{H}$ into $N+1$ subspaces according to the class distribution of hypothesis function: $\mathcal{H}=\bigcup_{\widehat{q}_N\in\{0,\frac{1}{N},\ldots,1\}}\mathcal{H}^{\widehat{q}_N}$, where $\mathcal{H}^{\widehat{q}_N}=\{h:\frac{1}{N}\sum_{i=1}^{N}\mathbf{I}[h(X_i)=+1]=\widehat{q}_N, h\in\mathcal{H}\}$

. Thus, according to the definition of pure loss and the union bound, we have

$$\begin{aligned}&\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{PL}_N(h)-\widehat{PL}'_N(h)\right|>\frac{\varepsilon}{2}\right\}\\&=\mathbb{P}\left\{\sup_{\widehat{q}_N}\sup_{h\in\mathcal{H}^{\widehat{q}_N}}\left|\widehat{PL}_N(h)-\widehat{PL}'_N(h)\right|>\frac{\varepsilon}{2}\right\}\\&\leq (N+1)\sup_{\widehat{q}_N}\mathbb{P}\left\{\sup_{h\in\mathcal{H}^{\widehat{q}_N}}\left|\widehat{PL}_N(h)-\widehat{PL}'_N(h)\right|>\frac{\varepsilon}{2}\right\}\\&=(N+1)\sup_{\widehat{q}_N}\mathbb{P}\left\{\sup_{h\in\mathcal{H}^{\widehat{q}_N}}\left|\widehat{L}_N(h)-\widehat{L}'_N(h)\right|>\frac{\varepsilon(1-\widehat{RA}_N)}{2}\right\}.\end{aligned}\tag{118}$$

We employ Theorem 3.1 in Vapnik and Chervonenkis (1971) for the error terms. Besides, for any $\widehat{q}_N$, it satisfies that $1-\widehat{RA}_N\geq\frac{1-|2\widehat{p}_N-1|}{2}$ and $d_{vc}(\mathcal{H}^{\widehat{q}_N})\leq d_{vc}(\mathcal{H})$ for $\mathcal{H}^{\widehat{q}_N}\subseteq\mathcal{H}$. Then we have:

$$\begin{aligned}&\sup_{\widehat{q}_N}\mathbb{P}\left\{\sup_{h\in\mathcal{H}^{\widehat{q}_N}}\left|\widehat{L}_N(h)-\widehat{L}'_N(h)\right|>\frac{\varepsilon(1-\widehat{RA}_N)}{2}\right\}\\&\leq 2\sup_{\widehat{q}_N}\exp\left\{-\left(\frac{\varepsilon^2(1-\widehat{RA}_N)^2}{4}-\frac{d_{vc}(\mathcal{H}^{\widehat{q}_N})[\ln(2eN/d_{vc}(\mathcal{H}^{\widehat{q}_N}))]}{N}\right)N\right\}\\&\leq 2\exp\left\{-\left(\frac{\varepsilon^2(1-|2\widehat{p}_N-1|)^2}{16}-\frac{d_{vc}(\mathcal{H})\ln(2eN/d_{vc}(\mathcal{H}))}{N}\right)N\right\}.\end{aligned}\tag{119}$$

Combining (117), (118) and (119), we obtain the final result. □

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., & Roth, D. (2005a). Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, *6*(2), 393–425.

Agarwal, S., Harpeled, S., & Roth, D. (2005b). A uniform convergence bound for the area under the ROC curve. In *Proceedings of the international conference on artificial intelligence and statistics* (pp. 1–8).

Albatineh, A. N., & Niewiadomska-Bugaj, M. (2011). Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification*, *5*(3), 179–200.

Albatineh, A. N., Niewiadomskabugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*(2), 301–313.

Alcalafdez, J., Sanchez, L., Garcia, S., Jesus, M. J. D., Ventura, S., Garrell, J. M., et al. (2008). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307–318.

Bartlett, P. L., & Mendelson, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*(3), 463–482.

Bartlett, P. L., Jordan, M. I., & Mcauliffe, J. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Blair, E., & Stanley, F. J. (2008). Interobserver agreement in the classification of cerebral palsy. *Developmental Medicine & Child Neurology*, *27*(5), 615–622.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277–291.

Cameron, M. L., Briggs, K. K., & Steadman, J. R. (2003). Reproducibility and reliability of the outerbridge classification for grading chondral lesions of the knee arthroscopically. *American Journal of Sports Medicine*, *31*(1), 83–86.

Chong, E. K. P., & Żak, S. H. (2011). *An introduction to optimization* (3rd ed.). New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.

Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research*, *43*(2), 181–191.

Dua, D., & Graff, C. (2017). UCI machine learning repository. Retrieved May 26, 2018, from http://archive.ics.uci.edu/ml.

Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the International Joint Conference on Artificial Intelligence*, *17*, 973–978.

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical psychology*, *54*(5), 415–425.

Ferri, C., Hernandezorallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38.

Gao, W., Wang, L., Jin, R., Zhu, S., & Zhou, Z. (2016). One-pass AUC optimization. *Artificial Intelligence*, *236*, 1–29.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459.

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classifications. *Publications of the American Statistical Association*, *49*(268), 732–764.

Hazan, T., Keshet, J., & Mcallester, D. A. (2010). Direct loss minimization for structured prediction. In *Proceedings of the advances in neural information processing systems* (pp. 1594–1602).

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Ingo, S. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, *51*(1), 128–142.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the International Conference on Machine learning* (pp. 377–384).

Kotlowski, W., & Dembczynski, K. (2017). Surrogate regret bounds for generalized classification performance metrics. *Machine Learning*, *106*(4), 549–572.

Koyejo, O.O., Natarajan, N., Ravikumar, P.K., & Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In *Proceedings of the advances in neural information processing systems* (pp. 2744–2752).

Kuncheva, L. I. (2013). A bound on kappa-error diagrams for analysis of classifier ensembles. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 494–501.

Li, F., Qian, Y., Wang, J., & Liang, J. (2016). Multigranulation information fusion: A dempster-shafer evidence theory-based clustering ensemble method. *Information Sciences*, *378*(1), 58–63.

Li, F., Qian, Y., Wang, J., Dang, C., & Liu, B. (2018). Cluster's quality evaluation and selective clustering ensemble. *ACM Transactions on Knowledge Discovery from Data*, *12*(5), 60.

Li, F., Qian, Y., Wang, J., Dang, C., & Jing, L. (2019). Clustering ensemble based on sample's stability. *Artificial Intelligence*, *273*, 37–55.

Margineantu, D. D., & Dietterich, T. G. (1997). Pruning adaptive boosting. In *Proceedings of the fourteenth international conference on machine learning* (pp. 211–218).

Martinezmunoz, G., & Suarez, A. (2006). Pruning in ordered bagging ensembles. In *International conference on machine learning* (pp. 609–616).

Menon, A. K., Narasimhan, H., Agarwal, S., & Chawla, S. (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the international conference on machine learning* (pp. 603–611).

Musicant, D. R., Kumar, V., & Ozgur, A. (2003). Optimizing F-measure with support vector machines. In *Proceedings of the Florida AI Research Society* (pp. 356–360).

Narasimhan, H., & Agarwal, S. (2013). A new support vector method for optimizing partial AUC based on a tight convex upper bound. In *Proceedings of the conference on knowledge discovery and data mining*.

Narasimhan, H., Vaish, R., & Agarwal, S. (2014). On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in neural information processing systems* (pp. 1493–1501).

Narasimhan, H., Ramaswamy, H. G., Saha, A., & Agarwal, S. (2015). Consistent multiclass algorithms for complex performance measures. In *Proceedings of the international conference on machine learning* (pp. 2398–2407).

Qian, Y., Li, F., Liang, J., Liu, B., & Dang, C. (2016). Space structure and clustering of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(10), 2047–2059.

Sabers, D. L., & Feldt, L. S. (1968). An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. *Journal of Educational Measurement*, *5*(3), 251–258.

Sanyal, A., Kumar, P., Kar, P., Chawla, S., & Sebastiani, F. (2018). Optimizing non-decomposable measures with deep networks. *Machine Learning*, *107*, 1597–1620.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325.

Song, Y., Schwing, A. G., Zemel, R. S., & Urtasun, R. (2016). Training deep neural networks via direct loss minimization. In *Proceedings of the International Conference on Machine learning* (pp. 2169–2177).

Valiant, L. G. (1984). A theory of the learnable. *Communications of ACM*, *27*(11), 1134–1142.

Vapnik, V., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, *16*(2), 264–280.

Vieira, S. M., Kaymak, U., & Sousa, J. (2010). Cohen's kappa coefficient as a performance measure for feature selection. In *Proceedings of the international conference on fuzzy systems* (pp. 1–8).

Vinh, N.X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the international conference on machine learning* (pp. 1073–1080).

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.

Waegeman, W., Dembczyńki, K., Jachnik, A., Cheng, W., & Hüllermeier, E. (2014). On the Bayes-optimality of F-measure maximizers. *Journal of Machine Learning Research*, *15*(1), 3333–3388.

Wu, Q., Laet, T.D., & Janssen, R. (2017). Elimination scoring versus correction for guessing: A simulation study. In *Proceedings of the meeting of the psychometric society*.

Zhang, T. (2003). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, *32*(1), 56–134.

Zhao, M., Edakunni, N. U., Pocock, A. C., & Brown, G. (2013). Beyond Fano's inequality: Bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, *14*(1), 1033–1090.

Zhou, Z., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*, 239–263.