LETTER Communicated by Olcay Yildiz

Credible Intervals for Precision and Recall Based on a *K*-Fold Cross-Validated Beta Distribution

Yu Wang
wangyu@sxu.edu.cn
Jihong Li*
lijh@sxu.edu.cn
School of Software, Shanxi University, Taiyuan 030006, P.R.C.

In typical machine learning applications such as information retrieval, precision and recall are two commonly used measures for assessing an algorithm's performance. Symmetrical confidence intervals based on *K*-fold cross-validated *t* distributions are widely used for the inference of precision and recall measures. As we confirmed through simulated experiments, however, these confidence intervals often exhibit lower degrees of confidence, which may easily lead to liberal inference results. Thus, it is crucial to construct faithful confidence (credible) intervals for precision and recall with a high degree of confidence and a short interval length. In this study, we propose two posterior credible intervals for precision and recall based on *K*-fold cross-validated beta distributions. The first credible interval for precision (or recall) is constructed based on the beta posterior distribution inferred by all *K* data sets corresponding to *K* confusion matrices from a *K*-fold cross-validation. Second, considering that each data set corresponding to a confusion matrix from a *K*-fold cross-validation can be used to infer a beta posterior distribution of precision (or recall), the second proposed credible interval for precision (or recall) is constructed based on the average of *K* beta posterior distributions. Experimental results on simulated and real data sets demonstrate that the first credible interval proposed in this study almost always resulted in degrees of confidence greater than 95%. With an acceptable degree of confidence, both of our two proposed credible intervals have shorter interval lengths than those based on a corrected *K*-fold cross-validated *t* distribution. Meanwhile, the average ranks of these two credible intervals are superior to that of the confidence interval based on a *K*-fold cross-validated *t* distribution for the degree of confidence and are superior to that of the confidence interval based on a corrected *K*-fold cross-validated *t* distribution for the interval length in all 27 cases of simulated and real data experiments. However, the confidence intervals based on the *K*-fold and corrected *K*-fold cross-validated *t* distributions

*Jihong Li is the corresponding author.

Neural Computation **28, 1694–1722** (2016)
doi:10.1162/NECO_a_00857

are in the two extremes. Thus, when focusing on the reliability of the inference for precision and recall, the proposed methods are preferable, especially for the first credible interval.

1 Introduction

There are multiple candidate models (i.e., algorithms) for a typical machine learning application and we need to choose one or several among many. In classification tasks with two classes of supervised learning, this is done by comparing the misclassification error, which is the sum of false positives and false negatives. However, as Yildiz, Aslan, and Alpaydin (2011) pointed out, misclassification error does not make a distinction between false positives and false negatives. Thus, many other performance measures have been proposed to evaluate candidate models, such as precision and recall. Precision and recall that are based on a binary contingency table are two measures that are commonly used in machine learning applications such as information retrieval (see Tables 1 and 2).

In practice, to be able to eliminate the effect by chance (e.g., variance due to small changes in the training set), one typically does training and validation a number of times, possibly by various resampling methods such as cross-validation and bootstrap (Alpaydin, 1999; Bengio & Grandvalet, 2004; Dietterich, 1998; Efron & Tibshirani, 1993; Hastie, Tibshirani, & Friedman, 2001; Markatou, Tian, Biswas, & Hripcsak, 2005; Nadeau & Bengio, 2003; Wang, Wang, Jia, & Li, 2014; Yildiz, 2013). For example, after deriving K training and validation sets, classification algorithms are trained with the K training sets, and K confusion matrices are subsequently obtained based on the validation sets (Bengio & Grandvalet, 2004; Markatou et al., 2005; Moreno-Torres, Saez, & Herrera, 2012). Then the precision and recall values can be calculated based on the K confusion matrices from K-fold cross-validation, and these are commonly evaluated with two measures: the microaverage and the macroaverage. The so-called microaveraged precision (or recall) is computed based on the average of the corresponding elements of K confusion matrices, while macroaveraged precision (or recall) is the average of K precisions (or recalls) computed by each confusion matrix.

Traditionally, when applying a learning algorithm in machine learning, the focus is typically directed at the the single-point micro- and macroaveraged precision and recall values of the algorithm's performance from a K-fold cross-validation. However, as Wang, Li, Li, Wang, and Yang (2015) pointed out, point estimations are rather trivial and do not consider variations of the estimation. In response to this, symmetrical confidence intervals based on K-fold cross-validated t distributions have been proposed. As we confirmed through simulated experiments, however, these confidence intervals often exhibit lower degrees of confidence and short interval lengths (see section 4). This may easily lead to liberal inference results. When confidence intervals are used to compare the performance of two algorithms, for example, the results can be misleading insofar as they can

Table 1: Contingency Table for a Two-Class Classification Problem.

		Predicted Positive	Class Negative	Sum
True class	Positive	TP	FN	P′
	Negative	FP	TN	N′
	Sum	P	N	

Note: TP (resp. TN) is the number of true positives (resp. negatives) and FP (resp. FN) the number of false positives (resp. negatives).

Table 2: Performance Measures.

Name	Formula
Error	$(FP + FN)/(TP + FP + TN + FN)$
Precision	$TP/(TP + FP)$
Recall	$TP/(TP + FN)$
F_1 score	$2TP/(2TP + FP + FN)$
Sensitivity	$TP/(TP + FN)$
Specificity	$TN/(FP + TN)$
True positive rate	$TP/(TP + FN)$
False positive rate	$FP/(FP + TN)$
Matthews correlation coefficient	$\frac{TP*TN-FP*FN}{\sqrt{P*N*P'*N'}}$

imply that two algorithms are significantly different when in fact they are not.

Furthermore, a theoretical analysis of the posterior distributions of precision and recall in Goutte and Gaussier (2005) shows that they follow a beta distribution. As such, these distributions are always nonsymmetrical, owing to the occurrence of two different parameters in the beta distribution, as shown in Figure 1. Of course, when these two parameters are the same, a beta distribution is symmetrical, but this might not always occur because there will always be an unequal number of true positives (TPs) and false positives (FPs) (or false negatives, FNs) in practical applications. (See Goutte & Gaussier, 2005, and Wang et al., 2015.) Consider case finding for rare diseases as a practical example. In case finding, a good case-finding model may always have FP $\gg$ TP (due to class imbalance) and FN $\gg$ TP. Meanwhile, symmetrical confidence intervals may significantly affect the estimation accuracy of the confidence interval in some cases. This is because the values of precision and recall range between 0 and 1, whereas the symmetrical confidence interval can exceed the range of $(0, 1)$ (Wang et al., 2015). Thus, the use of a symmetrical distribution, such as the commonly used t distribution, may be inappropriate for approximating the distribution of precision and recall, and this can result in large bias and a critically false conclusion.

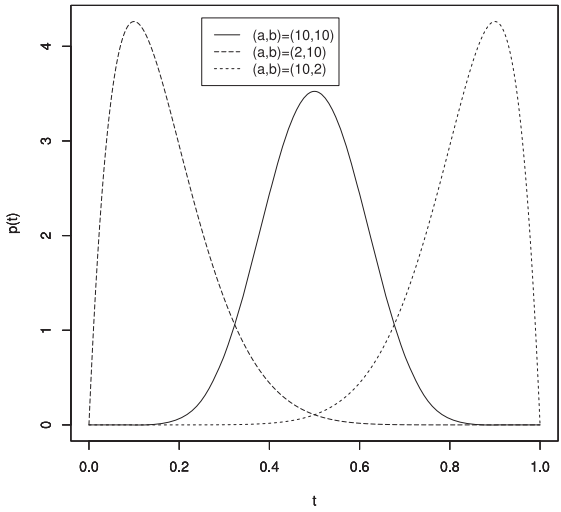

Figure 1: Density curves of beta distribution $B(a, b)$ with different parameter combinations of a and b.

To effectively measure the performance of an algorithm, it is crucial to construct faithful confidence (credible) intervals for precision and recall—that is, intervals with a high degree of confidence and a short interval length. In Bayesian statistics, credible intervals are analogous to confidence intervals in frequentist statistics. The degree of confidence of a credible interval is the probability of the inclusion of the true value in the credible interval. Interval length indicates the accuracy of the credible interval. Thus, in this study, two posterior credible intervals for precision and recall are constructed based on a K-fold cross-validated beta distribution.

The remainder of this study is organized as follows. Section 2 defines the standard precision and recall measures of an algorithm's performance and then gives their (single-point) estimations based on a K-fold cross-validation. Two credible intervals based on K-fold cross-validated beta distributions proposed in this letter and confidence intervals based on K-fold and corrected K-fold cross-validated t distributions are described in section 3. Section 4 discusses the simulated and real data experiments that show how the confidence (credible) intervals behaves compare. Section 5 concludes the study.

2 Precision and Recall Measures of an Algorithm's Performance

In studies on a two-class classification problem of machine learning, the performance of the learning algorithm is always assessed with empirical measures, based on the TP, FP, true negative (TN), and FN values of a

2×2 confusion matrix. In practice, a number of such measures have been developed depending on the type of error under consideration, including the precision value, the recall value, the F_1 score, sensitivity, specificity, the TP rate, the FP rate, the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and the Matthews correlation coefficient as shown in Table 2 (Powers, 2011; Fawcett, 2006; Flach, 2003; Goutte & Gaussier, 2005; Lobo, Jimenez, & Real, 2008; Nadeau & Bengio, 2003; Wang et al., 2015; Yang & Liu, 1999). In this study, we focus on two important performance indicators in machine learning: precision and recall values.

Strictly speaking, the precision and recall values are estimations of the theoretical precision and recall measures for a specific practical application. Thus, we first discuss theoretical precision and recall measures.

2.1 Theoretical Precision and Recall Measures. Without loss of generality, in this study, we consider only the following two-class classification problems: each class is associated with a binary label $l = \{+, -\}$, which accounts for the correctness of the class with respect to the task considered, and the classification algorithm produces a prediction z indicating whether it believes the class to be correct. Then precision may be defined as the probability that a class is positive $(+)$ given that it is returned by the classification algorithm, while the recall is the probability that a positive class is returned (Goutte & Gaussier, 2005; Wang et al., 2015):

$$p = P(l = +|z = +), \tag{2.1}$$

$$r = P(z = +|l = +). \tag{2.2}$$

2.2 Precision and Recall Values Based on a Confusion Matrix. For a specific two-class classification problem, the experimental outcome may be conveniently summarized in a 2×2 confusion matrix:

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}.$$

From these counts, one can obtain the empirical precision and recall values shown in Table 2:

$$p \triangleq \frac{TP}{TP + FP}, \tag{2.3}$$

$$r \triangleq \frac{TP}{TP + FN}. \tag{2.4}$$

It is obvious that the precision and recall values are estimations of the theoretical precision and recall measures.

2.3 Microaveraged Precision and Recall Values Based on a *K*-Fold Cross-Validation. In practice, in order to eliminate the effect by chance (e.g., variance due to small changes in the training set), a resampling method is always used. *K*-fold cross-validation is probably the simplest and most widely used resampling method. It uses all available examples as training and test examples; it mimics *K* training and test sets by using some of the data to fit the model and some to test it.

Formally, the data set S is split into K disjoint and equal-sized blocks, denoted as $T_k, k = 1, 2, \ldots, K$. Let S_k be the training set obtained by removing the elements in T_k from S; $TP(A(S_k), T_k)$, $FP(A(S_k), T_k)$, $FN(A(S_k), T_k)$, and $TN(A(S_k), T_k)$ be the elements of confusion matrix returned by algorithm A trained on the set S_k and tested on T_k (briefly denoted as TP_k, FP_k, FN_k, and TN_k). The averaged confusion matrix based on their respective averages of K TP_ks, FP_ks, FN_ks, and TN_ks has the following form:

$$\begin{pmatrix} \frac{1}{K}\sum_{k=1}^{K} TP_k & \frac{1}{K}\sum_{k=1}^{K} FN_k \\ \frac{1}{K}\sum_{k=1}^{K} FP_k & \frac{1}{K}\sum_{k=1}^{K} TN_k \end{pmatrix}.$$

Then, from equations 2.3 and 2.4, the microaveraged precision and recall values based on a *K*-fold cross-validation can be obtained:

$$p^{Micro} \triangleq \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FP_k}, \tag{2.5}$$

$$r^{Micro} \triangleq \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FN_k}. \tag{2.6}$$

2.4 Macroaveraged Precision and Recall Values Based on a *K*-Fold Cross-Validation. The so-called macroaveraged precision (recall) value based on a *K*-fold cross-validation is the average of K precisions (recalls) computed by a confusion matrix obtained based on each S_k and T_k for $k = 1, 2, \ldots, K$.

If denoting p_k as the precision value computed based on the kth confusion matrix (TP_k, FP_k, FN_k, and TN_k) and r_k as the corresponding recall, the macroaveraged precision and recall values based on a *K*-fold cross-validation are defined as the averages of precisions and recalls on K groups:

$$p^{Macro} \triangleq \frac{1}{K}\sum_{k=1}^{K} p_k, \tag{2.7}$$

$$r^{Macro} \triangleq \frac{1}{K}\sum_{k=1}^{K} r_k, \tag{2.8}$$

where $p_k = \frac{TP_k}{TP_k+FP_k}$, $r_k = \frac{TP_k}{TP_k+FN_k}$. $TP_k + FN_k = P'$ are all identical for $k = 1, 2, \ldots, K$ in r^{Macro}. Then we have

$$r^{Macro} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k+FN_k} = \frac{\frac{1}{K}\sum_{k=1}^{K} TP_k}{P'} = \frac{\frac{1}{K}\sum_{k=1}^{K} TP_k}{\frac{1}{K}\sum_{k=1}^{K}(TP_k+FN_k)} = r^{Micro}.$$

Remark 1. From above analysis, we can see that the macroaveraged and microaveraged recall values are identical. However, for the macroaveraged and microaveraged precision values, there is no similar conclusion.

3 Credible Intervals for Precision and Recall Measures

In this section, we present four credible and confidence intervals that can be used to infer precision and recall measures. The first two are the posterior credible intervals we propose, and the third and fourth confidence intervals have already been discussed in the literature. The first credible interval for precision (or recall) measure is provided by studying the posterior distribution of the precision (or recall) inferred by all data sets corresponding to K confusion matrices from a K-fold cross-validation. The second credible interval for precision (or recall) is constructed based on the average of K beta posterior distributions, in which each beta posterior distribution is inferred by a data set corresponding to a confusion matrix from K-fold cross-validation. For convenience, we provide several useful lemmas.

Lemma 1. *Observed TP, FP, FN, and TN counts follow a multinomial distribution with parameters $\pi = (\pi_{TP}, \pi_{FP}, \pi_{FN}, \pi_{TN})$, denoted by $D|\pi \sim M(n; \pi)$,*

$$\begin{aligned} &P(D=(TP, FP, FN, TN)) \\ &\quad= \frac{n!}{TP!FP!FN!TN!}(\pi_{TP})^{TP}(\pi_{FP})^{FP}(\pi_{FN})^{FN}(\pi_{TN})^{TN}, \end{aligned}$$

where $\pi_{TP} + \pi_{FP} + \pi_{FN} + \pi_{TN} = 1$, $TP + FP + FN + TN = n$. If rewriting TP, FP, FN, TN be n_1, n_2, n_3, n_4, $\pi_{TP}, \pi_{FP}, \pi_{FN}, \pi_{TN}$ be $\pi_1, \pi_2, \pi_3, \pi_4$, we have the following properties:

Property 1: Each component n_i of D follows a binomial distribution $B(n; \pi_i)$ for $i = 1, 2, 3, 4$.

Property 2. Each component n_i of D conditioned on another component n_j follows a binomial distribution $B(n - n_j; \pi_i/(1 - \pi_j)$ for $i,\ j = 1, 2, 3, 4$ and $i \neq j$.

Property 3. The sum of n_i and n_j also follows a binomial distribution $B(n; \pi_i + \pi_j)$.

Property 4. The distribution of n_i given the number of returned objects $n_i + n_j$ is a binomial with parameters $n_i + n_j$ and $\pi_i/(\pi_i + \pi_j)$ for $i,\ j = 1, 2, 3, 4$ and $i \neq j$.

The proof of lemma 1 and properties 1 to 4 can be found in Goutte and Gaussier (2005). Furthermore, Goutte and Gaussier (2005) revealed that the distributions of precision and recall have the following forms:

Lemma 2.

$$P(p|D) \propto P(D|p)P(p) = p^{TP+\lambda-1}(1-p)^{FP+\lambda-1}$$

that is, $p|D \sim Be(TP + \lambda, FP + \lambda)$*(beta distribution), where* $P(p)$ *is the prior distribution and* $p \sim Be(\lambda, \lambda)$, $D = (TP, FP, FN, TN)$, λ *is the prior parameter. A similar development yields the posterior distribution for the recall:* $r|D \sim Be(TP + \lambda, FN + \lambda)$.

Lemma 3. *Given two independent variables with binomial distributions* $X \sim B(n; \pi)$ *and* $Y \sim B(m; \pi)$ *with identical parameter* π*, the following property holds:* $X + Y \sim B(n + m; \pi)$.

Lemma 4. *Let* $U_1, U_2, \ldots, U_K$ *be random variables with common mean* β *and the following covariance structure*

$$Var(U_k) = \delta, \forall\, k, Cov(U_k, U_{k'}) = \gamma, \forall\, k \neq k'.$$

Let $\rho = \gamma/\delta$ *be the correlation between* U_k *and* $U_{k'}$ *and* $\bar{U} = \frac{1}{K}\sum_{k=1}^{K} U_k$ *the sample mean; then* $\mathrm{Var}(\bar{U}) = \frac{\delta}{K}(1 + (K - 1)\rho)$.

Lemmas 2 to 4 can be found in Goutte and Gaussier (2005) and Dietterich (1998), respectively.

3.1 Credible Intervals Constructed Based on Beta Posterior Distributions Inferred by *K* Data Sets. First, we consider the posterior distribution of precision inferred by the K data sets corresponding to K confusion matrices from a K-fold cross-validation. These data sets are denoted D_1, D_2, and D_K, where $D_k = (TP_k, FP_k, FN_k,$ and $TN_k)$ and $k = 1, \ldots, K$. By assuming that the D_ks are independent for $k = 1, \ldots, K$, lemma 5 can be obtained:

Lemma 5. *Provided that the* D_k*s are independent for* $k = 1, \ldots, K$*, the conditional random variables* $\left(\sum_{k=1}^{K} TP_k\right)\Big|\left(\sum_{k=1}^{K}(TP_k + FP_k)\right)$ *and* $\sum_{k=1}^{K}(TP_k|(TP_k$

$+ FP_k$)) *have the same distribution, and they all follow a binomial distribution* $B\big(\sum_{k=1}^{K}(TP_k + FP_k);\ p = \frac{\pi_{TP}}{\pi_{TP}+\pi_{FP}}\big)$.

Proof. From lemma 1 and property 1, we know that TP_k follows $B(n; \pi_{TP})$ for $k = 1, 2, \ldots, K$. Combined with the assumption of the independence of TP_ks, we have $\sum_{k=1}^{K} TP_k$ follows $B(Kn; \pi_{TP})$ from lemma 3. For the variable $TP_k + FP_k$, it is obvious that its distribution is $B(n; \pi_{TP} + \pi_{FP})$ from properties 1, 2, and 3. Then we have $\sum_{k=1}^{K}(TP_k + FP_k)$ follows $B(Kn; (\pi_{TP} + \pi_{FP}))$.

Thus, property 4 postulates that

$$\left(\sum_{k=1}^{K} TP_k\right) \Bigg| \left(\sum_{k=1}^{K}(TP_k + FP_k)\right) \sim B\left(\sum_{k=1}^{K}(TP_k + FP_k);\ p = \frac{\pi_{TP}}{\pi_{TP} + \pi_{FP}}\right). \tag{3.1}$$

Similarly, from property 4, we know that

$$TP_k|(TP_k + FP_k) \sim B\left(TP_k + FP_k;\ p = \frac{\pi_{TP}}{\pi_{TP} + \pi_{FP}}\right)$$

for $k = 1, \ldots, K$. Then, combining the independence assumption of D_ks, we have

$$\sum_{k=1}^{K}(TP_k|(TP_k + FP_k)) \sim B\left(\sum_{k=1}^{K}(TP_k + FP_k);\ p = \frac{\pi_{TP}}{\pi_{TP} + \pi_{FP}}\right). \tag{3.2}$$

A similar conclusion can be obtained for recall. From equations 3.1 and 3.2 that we can see that $\big(\sum_{k=1}^{K} TP_k\big)\big|\big(\sum_{k=1}^{K}(TP_k + FP_k)\big)$ and $\sum_{k=1}^{K}(TP_k|(TP_k + FP_k))$ follow a binomial distribution $B\big(\sum_{k=1}^{K}(TP_k + FP_k);\ p = \frac{\pi_{TP}}{\pi_{TP}+\pi_{FP}}\big)$.

Lemma 2 tells us that if we assume that p has a prior distribution of $Be(\lambda, \lambda)$, we can infer the posterior distribution of p based on equations 3.1 and 3.2.

Proposition 1. *Provided that the* D_k*s are independent for* $k = 1, \ldots, K$*, the posterior distribution of precision is a beta distribution with parameters* $\sum_{k=1}^{K} TP_k + \lambda$ *and* $\sum_{k=1}^{K} FP_k + \lambda$*, that is,* $p|(D_1, D_2, \ldots, D_K) \sim Be\big(\sum_{k=1}^{K} TP_k + \lambda, \sum_{k=1}^{K} FP_k + \lambda\big)$*. A similar development yields the posterior distribution for the recall:* $r|(D_1, D_2, \ldots, D_K) \sim Be\big(\sum_{k=1}^{K} TP_k + \lambda, \sum_{k=1}^{K} FN_k + \lambda\big)$.

Proof. Similar to lemma 2, based on equations 3.1 and 3.2, we can write the likelihood of p as

$$L(p) = P((D_1, D_2, \ldots, D_K)|p) \propto p^{\sum_{k=1}^{K} TP_k}(1-p)^{\sum_{k=1}^{K} FP_k}$$

Inference on p can then be performed using Bayes' rule:

$$\begin{aligned} P(p|(D_1, D_2, \ldots, D_K)) &\propto P((D_1, D_2, \ldots, D_K)|p)P(p) \\ &\propto p^{\sum_{k=1}^{K} TP_k+\lambda-1}(1-p)^{\sum_{k=1}^{K} FP_k+\lambda-1}. \end{aligned}$$

That is, $p|(D_1, D_2, \ldots, D_K) \sim Be\big(\sum_{k=1}^{K} TP_k + \lambda, \sum_{k=1}^{K} FP_k + \lambda\big)$. If replacing FP by FN, a similar conclusion can be obtained for recall.

Remark 2. Obtaining proposition 1 requires that D_ks be independent. However, the training sets from any two independent partitions in a K-fold cross-validation contain common samples regardless of how the data set is split. In other words, the training sets are related. Furthermore, Bengio and Grandvalet (2004) pointed out that the correlations of training sets in a K-fold cross-validation should not be negligible. Thus, the TPs, FPs, and FNs are actually not independent. This results in parameters of $Be\big(\sum_{k=1}^{K} TP_k + \lambda, \sum_{k=1}^{K} FP_k + \lambda\big)$ and $Be\big(\sum_{k=1}^{K} TP_k + \lambda, \sum_{k=1}^{K} FN_k + \lambda\big)$ that are greater than the true parameters of them; that is, the true parameters are actually smaller than $\sum_{k=1}^{K} TP_k + \lambda, \sum_{k=1}^{K} FP_k + \lambda, \sum_{k=1}^{K} FN_k + \lambda$. For this, the precision and recall should follow the distributions of $Be\big(\omega \cdot \big(\sum_{k=1}^{K} TP_k\big) + \lambda, \omega \cdot \big(\sum_{k=1}^{K} FP_k\big) + \lambda\big)$ and $Be\big(\omega \cdot \big(\sum_{k=1}^{K} TP_k\big) + \lambda, \omega \cdot \big(\sum_{k=1}^{K} FN_k\big) + \lambda\big)$ with $1/K \le \omega \le 1$. Here, the problem is that ω is unknown and needs to be estimated appropriately. When the correlations of the TPs, FPs, and FNs are large, the ω tends to be small. By contrast, when the correlations of these variables are small, the ω becomes large. Intuitively, using the average ω in the interval $(1/K, 1)$ as the value of ω is a natural selection, denoted as $\omega_0 = (K+1)/2K$. Indeed, the average ω may not be the best choice; however, it provides a solution that is close to the best ω with a closed form and greatly saves computational cost. (See the discussion based on the simulated experiments in the next section.)

Thus, the resulting credible intervals, defined as CI_{p^M} and CI_{r^M}, for precision and recall measures based on the percentiles of the beta distribution are

$$\begin{aligned} CI_{p^M} = \Bigg[& Be\left(\frac{K+1}{2K}\left(\sum_{k=1}^{K} TP_k\right) + \lambda, \frac{K+1}{2K}\left(\sum_{k=1}^{K} FP_k\right) + \lambda\right)_{\alpha/2}, \\ & Be\left(\frac{K+1}{2K}\left(\sum_{k=1}^{K} TP_k\right) + \lambda, \frac{K+1}{2K}\left(\sum_{k=1}^{K} FP_k\right) + \lambda\right)_{1-\alpha/2} \Bigg], \end{aligned} \quad (3.3)$$

$$CI_{r^M} = \left[Be\left(\frac{K+1}{2K}\left(\sum_{k=1}^{K} TP_k\right)+\lambda,\ \frac{K+1}{2K}\left(\sum_{k=1}^{K} FN_k\right)+\lambda\right)_{\alpha/2},\right.$$
$$\left. Be\left(\frac{K+1}{2K}\left(\sum_{k=1}^{K} TP_k\right)+\lambda,\ \frac{K+1}{2K}\left(\sum_{k=1}^{K} FN_k\right)+\lambda\right)_{1-\alpha/2}\right], \quad (3.4)$$

where $Be(\cdot)_\alpha$ denotes the α percentiles of beta distribution.

3.2 Credible Intervals Based on the Average of the *K* Beta Posterior Distributions. From lemma 2, we know that for a data set D_k corresponding to a confusion matrix from K-fold cross-validation, we have $p|D_k \sim Be(TP_k+\lambda, FP_k+\lambda)$ for $D_k = (TP_k, FP_k, FN_k,$ and $TN_k)$ and $k = 1, \ldots, K$. However, the posterior distribution of p depends exclusively on a fractional sample set D_k. To use all of the samples to infer the precision and recall, we might consider implementing the average of all $p|D_k$. We might also seek to determine whether $p^A \triangleq \sum_{k=1}^{K}(p|D_k)/K$ similarly follows a beta distribution.

If assuming that the D_ks are independent of each other, the distribution of $p^A = \sum_{k=1}^{K}\phi_k = \sum_{k=1}^{K}(\frac{1}{K}(p|D_k))$ can be expressed as

$$F_{p^A}(t) = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f_{\phi_1}(t_1)\cdots f_{\phi_{k-1}}(t_{k-1}) f_{\phi_{k+1}}(t_{k+1})\cdots f_{\phi_K}(t_K) F_{\phi_k}(t - t_1$$
$$- \cdots - t_{k-1} - t_{k+1} - \cdots - t_K) dt_1 \cdots dt_{k-1} dt_{k+1} \cdots dt_K, \quad (3.5)$$

where

$$f_{\phi_k}(t_k) = K f_{Be(TP_k+\lambda, FP_k+\lambda)}(Kt_k)$$

and

$$F_{\phi_k}(t_k) = \int_{-\infty}^{Kt_k} f_{Be(TP_k+\lambda, FN_k+\lambda)}(t)dt$$

are the probability density and distribution functions of random variable ϕ_k for $k = 1, \ldots, K$, respectively. $f_{Be(a,b)}(t)$ denotes the density function of the beta distribution with parameters of a and b.

From equation 3.5, we can see that despite the independence of the D_ks, the distribution of p^A is nevertheless complex and cannot be used directly to construct a credible interval. A straightforward method, however, is to approximate this distribution with a beta distribution, given that p^A is an

average of the K random variables following a beta distribution. Intuitively, its distribution should be close to a beta distribution:

Proposition 2. *Recalling that* $p^A \triangleq \sum_{k=1}^{K}(p|D_k)/K$ *and* $r^A \triangleq \sum_{k=1}^{K}(r|D_k)/K$, *where* $p|D_k$ *and* $r|D_k$ *follow* $Be(TP_k + \lambda, FP_k + \lambda)$ *and* $Be(TP_k + \lambda, FN_k + \lambda)$, *respectively, for* $D_k = (TP_k, FP_k, FN_k, \text{and } TN_k)$ *and* $k = 1, \ldots, K$, *the distributions of* p^A *and* r^A *can be approximated with* $Be(a_p^A, b_p^A)$ *and* $Be(a_r^A, b_r^A)$, *that is,*

$$p^A \overset{P}{\approx} Be(a_p^A, b_p^A), r^A \overset{P}{\approx} Be(a_r^A, b_r^A), \tag{3.6}$$

where

$$a_p^A = \frac{E_p}{V_p}(E_p - E_p^2 - V_p), b_p^A = \frac{1 - E_p}{V_p}(E_p - E_p^2 - V_p),$$

$$a_r^A = \frac{E_r}{V_r}(E_r - E_r^2 - V_r), b_r^A = \frac{1 - E_r}{V_r}(E_r - E_r^2 - V_r),$$

$$E_p = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k + \lambda}{TP_k + FP_k + 2\lambda}, E_r = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k + \lambda}{TP_k + FN_k + 2\lambda},$$

$$V_p = \left(1 + \frac{K-1}{K}\right)\frac{1}{K^2}\sum_{k=1}^{K}\frac{(TP_k + \lambda)(FP_k + \lambda)}{(TP_k + FP_k + 2\lambda)^2(TP_k + FP_k + 2\lambda + 1)},$$

$$V_r = \left(1 + \frac{K-1}{K}\right)\frac{1}{K^2}\sum_{k=1}^{K}\frac{(TP_k + \lambda)(FN_k + \lambda)}{(TP_k + FN_k + 2\lambda)^2(TP_k + FN_k + 2\lambda + 1)}.$$

Proof. Here, $p^A \triangleq \sum_{k=1}^{K}(p|D_k)/K$, and $p|D_k \sim Be(TP_k + \lambda, FP_k + \lambda)$ for $D_k = (TP_k, FP_k, FN_k, TN_k)$ and $k = 1, \ldots, K$. By equating the first and second moments of p^A and the random variable following beta distribution, we have

$$\begin{cases} E(p^A) = \dfrac{a_p^A}{a_p^A + b_p^A}, \\ \text{Var}(p^A) = \dfrac{a_p^A b_p^A}{(a_p^A + b_p^A)^2(a_p^A + b_p^A + 1)}. \end{cases}$$

On the other hand, we have

$$E(p^A) = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k + \lambda}{TP_k + FP_k + 2\lambda} \triangleq E_p.$$

However, the variance of p^A cannot simply be expressed as the average of the variances of the $p|D_k$s. This is because the correlations between TP_ks and FP_ks from a K-fold cross-validation cannot be negligible, as already noted. Thus, from lemma 4, the variance of p^A is written as

$$\mathrm{Var}(p^A) = \frac{\delta}{K}(1 + (K-1)\rho) \triangleq V_p, \tag{3.7}$$

where $\delta = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Var}(p|D_k)$, ρ denotes the correlation of ps from different D_ks:

$$\mathrm{Var}(p|D_k) = \frac{(TP_k + \lambda)(FP_k + \lambda)}{(TP_k + FP_k + 2\lambda)^2(TP_k + FP_k + 2\lambda + 1)}.$$

According to the recommendation in Nadeau and Bengio (2003), the ratio of the test sample size to the total sample size should be adopted when estimating ρ, that is, $\hat{\rho} = 1/K$.

From this, one can show that

$$\begin{cases} a_p^A = \dfrac{E_p}{V_p}(E_p - E_p^2 - V_p) \\ b_p^A = \dfrac{1 - E_p}{V_p}(E_p - E_p^2 - V_p) \end{cases}.$$

Similarly, we can develop the approximated distribution $B(a_r^A, b_r^A)$ of r^A, where

$$a_r^A = \frac{E_r}{V_r}(E_r - E_r^2 - V_r),\ b_r^A = \frac{1 - E_r}{V_r}(E_r - E_r^2 - V_r),$$

$$E_r = E(r^A) = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k + \lambda}{TP_k + FN_k + 2\lambda}$$

and

$$\begin{aligned} V_r &= \mathrm{Var}(r^A) \\ &= \left(1 + \frac{K-1}{K}\right)\frac{1}{K^2}\sum_{k=1}^{K} \frac{(TP_k + \lambda)(FN_k + \lambda)}{(TP_k + FN_k + 2\lambda)^2(TP_k + FN_k + 2\lambda + 1)}. \end{aligned}$$

Based on the obtained a_p^A, b_p^A, a_r^A, and b_r^A, we have

$$p^A \overset{P}{\approx} Be(a_p^A, b_p^A),\ r^A \overset{P}{\approx} Be(a_r^A, b_r^A).$$

Thus, the credible intervals based on the above beta distribution, defined as CI_{p^A} and CI_{r^A} for precision and recall measures, respectively, have the following forms:

$$CI_{p^A} = \left[Be\left(a_p^A, b_p^A\right)_{\alpha/2}, Be\left(a_p^A, b_p^A\right)_{1-\alpha/2}\right], \tag{3.8}$$

$$CI_{r^A} = \left[Be\left(a_r^A, b_r^A\right)_{\alpha/2}, Be\left(a_r^A, b_r^A\right)_{1-\alpha/2}\right]. \tag{3.9}$$

Remark 3. To further validate the approximate extent of the beta distribution to the true distribution of p^A, the density functions of $Be(a_p^A(\rho = 0), b_p^A(\rho = 0))$, $Be(a_p^A, b_p^A)$ and the true density function of p^A are compared by the following simulated experiment, where $a_p^A(\rho = 0)$ and $b_p^A(\rho = 0)$ refer to a_p^A and b_p^A obtained when $\rho = 0$ from equation 3.7 (i.e., with independent D_ks). A similar comparison is also conducted for r^A.

3.2.1 Simulated Experiment 1. Density Functions of the True and Approximate Distributions for p^A and r^A. Considering a classification problem with two classes, we have $Z = (X, Y)$, with $\text{Prob}(Y = 1) = \text{Prob}(Y = 0) = \frac{1}{2}$, $X|Y = 0 \sim N(\mu_0, \Sigma_0)$, $X|Y = 1 \sim N(\mu_1, \Sigma_1)$. Here, we take $\mu_0 = 0_5$, $\Sigma_0 = I_5$, $\mu_1 = \beta_1 1_5$ and $\Sigma_1 = \beta_2 \Sigma_0$, where 0_5 and 1_5 denote the five-dimensional vector with the elements of all 0 and 1; I_5 denotes the five-order identity matrix, $(\beta_1, \beta_2) = (0.2, 1)$. The sample size is 200.

First, we can obtain the observed TP_k, FP_k, FN_k, and TN_k for $k = 1, \ldots, K$ with classification tree and support vector machine classifiers. The parameters $a_p^A(\rho = 0)$, $b_p^A(\rho = 0)$, $a_r^A(\rho = 0)$, $b_r^A(\rho = 0)$, a_p^A, b_p^A, a_r^A, and b_r^A are then computed. Thus, the approximate density functions of p^A and r^A can be obtained based on the distributions $Be(a_p^A(\rho = 0), b_p^A(\rho = 0))$, $Be(a_r^A(\rho = 0), b_r^A(\rho = 0))$, $Be(a_p^A, b_p^A)$, and $Be(a_r^A, b_r^A)$. Their true density function is computed by kernel density estimation with gaussian kernel.

In this experiment, we provide results from the most commonly used case (i.e., $K = 10$). However, under other conditions, such as $K = 2$ or $K = 5$, similar conclusions can be obtained. Next, we compare the difference of f_{p^A} (f_{r^A}), $f_{Be(a_p^A(\rho=0), b_p^A(\rho=0))}$ $(f_{Be(a_r^A(\rho=0), b_r^A(\rho=0))})$ and $f_{Be(a_p^A, b_p^A)}$ $(f_{Be(a_r^A, b_r^A)})$, where f refers to the density function.

From Figures 2 and 3, we can see that each of the three density curves has a similar shape for p^A and r^A regardless of whether a classification tree classifier or a support vector machine classifier is used. However, the density curves of $Be(a_p^A, b_p^A)$ and $Be(a_r^A, b_r^A)$ closely approximate

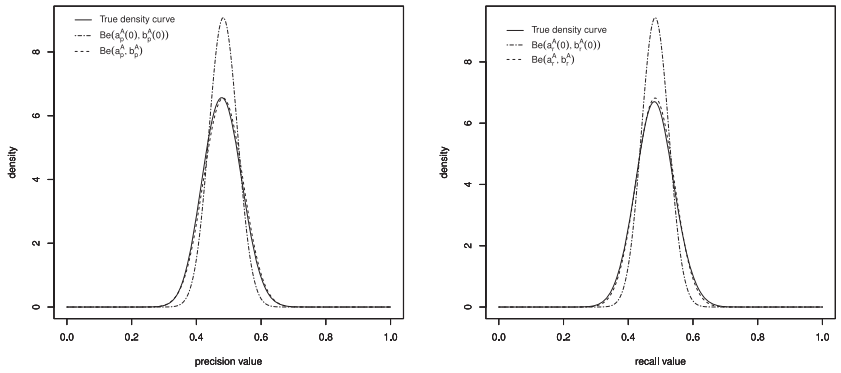

Figure 2: Density curves of true and approximate distributions for p^A and r^A with classification tree classifier.

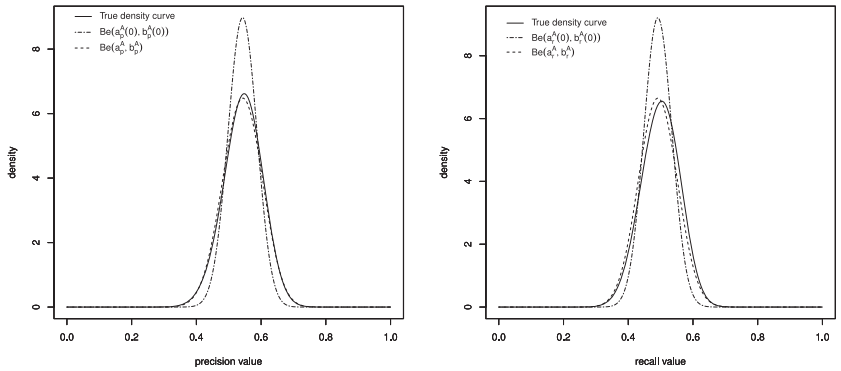

Figure 3: Density curves of true and approximate distributions for p^A and r^A with support vector machine classifier.

the true densities of p^A and r^A. Here, both $Be(a_p^A(\rho = 0), b_p^A(\rho = 0))$ and $Be(a_r^A(\rho = 0), b_r^A(\rho = 0))$ are based on the independence assumption and express a considerable bias at their peak points with respect to the true distributions of p^A and r^A. This again suggests the need to correct the parameters of $Be(a_p^A(\rho = 0), b_p^A(\rho = 0))$, and $Be(a_r^A(\rho = 0), b_r^A(\rho = 0))$. By not correcting these parameters, a liberal credible interval will doubtless obtain. This observation further indicates that the approximate beta distribution is relatively simple and easily adopted when constructing credible intervals compared to the complicated true distributions of p^A and r^A.

3.3 Symmetrical Confidence Intervals Based on the *K*-Fold Cross-Validated *t* Distribution. Symmetrical confidence intervals (statistical test

of significance) based on the normal or t distribution are widely used in the literature (Bisani & Ney, 2004; Keller, Bengio, & Wong, 2006; Nadeau & Bengio, 2003; Yang & Liu, 1999). The symmetrical confidence intervals based on the K-fold cross-validated t distribution at confidence level $1-\alpha$ will look like

$$\left[\hat{\mu} - c\sqrt{\hat{\sigma}^2}, \hat{\mu} + c\sqrt{\hat{\sigma}^2}\right],$$

where $\hat{\mu}$ is a mean estimator based on the average of the K-fold cross-validated estimators, $\hat{\sigma}^2$ is a variance estimator, and c is a percentile from Student's t distribution with a degree of freedom of $K-1$. Then the confidence intervals of precision and recall are written as

$$CI_{p^t} = \left[p^{Macro} - c\sqrt{\hat{\sigma}^2_{p^{Macro}}}, p^{Macro} + c\sqrt{\hat{\sigma}^2_{p^{Macro}}}\right], \tag{3.10}$$

$$CI_{r^t} = \left[r^{Macro} - c\sqrt{\hat{\sigma}^2_{r^{Macro}}}, r^{Macro} + c\sqrt{\hat{\sigma}^2_{r^{Macro}}}\right], \tag{3.11}$$

where $\hat{\sigma}^2_{p^{Macro}} = \frac{1}{K(K-1)}\sum_{k=1}^{K}(p_k - p^{Macro})^2$, $\hat{\sigma}^2_{r^{Macro}} = \frac{1}{K(K-1)}\sum_{k=1}^{K}(r_k - r^{Macro})^2$.

3.4 Symmetrical Confidence Intervals Based on the Corrected *K*-Fold Cross-Validated *t* Distribution. Bengio and Grandvalet (2004) showed that the correlation of test blocks cannot be ignored in computing the variance of K-fold cross-validation; otherwise, the variance will be grossly underestimated. Based on this, Grandvalet and Bengio (2006) obtained a corrected K-fold cross-validated t-test by correcting the variance of K-fold cross-validation. If we let $\hat{\mu}$ be p^{Macro} (or r^{Macro}), $\hat{\sigma}^2$ be $\hat{\sigma}^2_{p^{Macro}}/(1-\rho_{p^{Macro}})$ (or $\hat{\sigma}^2_{r^{Macro}}/(1-\rho_{r^{Macro}})$), we can obtain the symmetrical confidence interval based on the corrected K-fold cross-validated t distribution:

$$CI_{p^{Cl}} = \left[p^{Macro} - c\sqrt{\frac{\hat{\sigma}^2_{p^{Macro}}}{(1-\rho_{p^{Macro}})}}, p^{Macro} + c\sqrt{\frac{\hat{\sigma}^2_{p^{Macro}}}{(1-\rho_{p^{Macro}})}}\right], \tag{3.12}$$

$$CI_{r^{Cl}} = \left[r^{Macro} - c\sqrt{\frac{\hat{\sigma}^2_{r^{Macro}}}{(1-\rho_{r^{Macro}})}}, r^{Macro} + c\sqrt{\frac{\hat{\sigma}^2_{r^{Macro}}}{(1-\rho_{r^{Macro}})}}\right], \tag{3.13}$$

where $\rho_{p^{Macro}}$ ($\rho_{r^{Macro}}$) is the ratio of the covariance of p_ks (r_ks) for $k = 1, \ldots, K$ and the variance of p^{Macro} (r^{Macro}). Grandvalet and Bengio (2006) suggested an empirical estimation of $\hat{\rho}_{p^{Macro}} = 0.7$ by conducting a large number of experiments.

Table 3: Single-Point Estimation of Precision and Recall in the Case of $\mu_0 = 0_5$, $\Sigma_0 = I_5$, $\mu_1 = \beta_1 1_5$, and $\Sigma_1 = \beta_2 \Sigma_0$ with Different Combination of (β_1, β_2).

(β_1, β_2)	p^{Mac}	p^{Mic}	$r^{Mic} = r^{Mac}$
(1,2)	0.768	0.759	0.758
(1,0.1)	0.919	0.915	0.929
(0.2,3)	0.689	0.683	0.682

4 Simulated Experiments for Comparison

In this section, we first demonstrate with a simulation that false conclusions proceed from the use of single-point micro- and macroaveraged precision and recall estimations to estimate precision and recall measures. It may be more suitable based on confidence (credible) interval to infer them. We then investigate the degree of confidence and the interval length of the four credible and confidence intervals based on *K*-fold cross-validation presented in this study for multiple classifiers on simulated and real letter recognition and MAGIC gamma telescope data sets. For a given problem, we generated 1000 independent data sets to fully take into account the effect of the randomness of the training set, as well as that of the test examples.

For comparison, we took $K = 10$ (most commonly used in the literature) in *K*-fold cross-validation. We chose $\lambda = 1$, the uniform prior, in the beta distribution. The sample sizes were $n = 200$ and 1000 for simulated and real data sets. The confidence level $1 - \alpha = 0.95$, that is, $\alpha = 0.05$.

4.1 Single-Point Estimations of Precision and Recall Based on Micro- and Macroaverages. The simulated data $Z = (X, Y)$ were generated in a manner similar to simulated experiment 1, but we took $(\beta_1, \beta_2) = (1, 2), (1, 0.1), (0.2, 3)$. The classifier was classification tree. The sample size was 200.

From Table 3, it is clear that the single-point p^{Macro} value is higher than p^{Micro} and that the values of r^{Macro} and r^{Micro} are equivalent. It is always said that the macroaverage is superior to the microaverage in the literature because higher precision and recall values are blindly desirable by the authors. However, as Goutte and Gaussier (2005) and Wang et al. (2015) noted, the point estimation does not consider the variance of the estimation, and thus it is prone to false conclusions. For example, in the case of $(\beta_1, \beta_2) = (1, 2)$, the confidence interval for precision based on the *K*-fold cross-validated *t* distribution inferred from p^{Macro} was $(0.689, 0.846)$, which obviously includes the values of $p^{Micro} = 0.759$. This implies that even with a liberal confidence interval, it was difficult to make a distinction between p^{Macro} and p^{Micro}. In other words, the difference between p^{Macro} and p^{Micro} was not statistically significant, and this difference may result from random error. The

Table 4: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall Based on Perceptron Classifier.

		Case: $n = 200, d = 5$, $\mu_1 = 0.21_5, \Sigma_1 = I_5$	Case: $n = 1000, d = 100$, $\mu_1 = 0.21_{100}, \Sigma_1 = I_{100}$	Case: $n = 200, d = 300$, $\mu_1 = 0.21_{300}, \Sigma_1 = I_{300}$
CI_{p^M}	DOC	99.9%	99.4%	98.3%
	IL	0.256	0.099	0.184
CI_{p^A}	DOC	99.8%	98.7%	63.7%
	IL	0.237	0.099	0.196
CI_{p^t}	DOC	91.7%	93.1%	89.9%
	IL	0.172	0.074	0.144
$CI_{p^{Ct}}$	DOC	99.5%	99.4%	98.2%
	IL	0.314	0.135	0.263
CI_{r^M}	DOC	97.4%	98.7%	98.7%
	IL	0.254	0.101	0.234
CI_{r^A}	DOC	97.7%	97.2%	94.3%
	IL	0.225	0.101	0.213
CI_{r^t}	DOC	93.4%	95.3%	95.3%
	IL	0.247	0.094	0.216
$CI_{r^{Ct}}$	DOC	99.4%	99.7%	99.8%
	IL	0.452	0.172	0.395

fact that the conditional random variables $\sum_{k=1}^{K} TP_k \big| \sum_{k=1}^{K} (TP_k + FP_k)$ and $\sum_{k=1}^{K} (TP_k | (TP_k + FP_k)) \big(\sum_{k=1}^{K} TP_k \big| \sum_{k=1}^{K} (TP_k + FN_k)$ and $\sum_{k=1}^{K} (TP_k | (TP_k + FN_k)) \big)$ have the same distribution also validated this point from a different perspective. Thus, it may be more suitable based on confidence (credible) interval to implement the inference for precision and recall. We next compare the degree of confidence and the interval length of four credible and confidence intervals of precision and recall for multiple classifiers on simulated and real data sets.

4.2 Comparison of Credible and Confidence Intervals on Simulated Data. The experimental setup in this section was similar to that of section 4.1, in which multiple combinations of μ_0, Σ_0, μ_1, and Σ_1 were considered. The classifiers were a perceptron with one hidden layer, a classification tree, and a support vector machine with gaussian kernel.

Tables 4, 5, and 6 show the simulated results of the degree of confidence and interval length of four credible and confidence intervals based on K-fold cross-validation for precision and recall. First, we see that the confidence intervals based on a K-fold cross-validated t distribution exhibited a lower degree of confidence (below 95%) in almost all cases (in 28 of the 30 cases). For example, in six cases in Table 5, the degrees of confidence for this confidence interval of precision were 90.1%, 92.1%, 90.6%, 88.9%, 92.0%, and 87.9% for the classification tree classifier. In contrast, the degrees

Table 5: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall Based on Classification Tree Classifier.

		Case: $n = 200, d = 5$, $\mu_1 = 0.21_5, \Sigma_1 = I_5$	Case: $n = 1000, d = 100$, $\mu_1 = 0.21_{100}, \Sigma_1 = I_{100}$	Case: $n = 200, d = 300$, $\mu_1 = 0.21_{300}, \Sigma_1 = I_{300}$
CI_{p^M}	DOC	99.7%	99.6%	99.7%
	IL	0.256	0.116	0.255
CI_{p^A}	DOC	99.8%	99.6%	99.4%
	IL	0.236	0.116	0.235
CI_{p^t}	DOC	90.1%	92.1%	90.6%
	IL	0.164	0.070	0.165
$CI_{p^{Ct}}$	DOC	99.1%	99.2%	98.8%
	IL	0.299	0.127	0.302
CI_{r^M}	DOC	97.6%	97.1%	97.9%
	IL	0.254	0.116	0.253
CI_{r^A}	DOC	98.1%	97.5%	97.5%
	IL	0.226	0.115	0.226
CI_{r^t}	DOC	92.4%	92.9%	91.8%
	IL	0.232	0.107	0.228
$CI_{r^{Ct}}$	DOC	99.7%	99.5%	99.2%
	IL	0.423	0.195	0.416
		Case: $n = 200, d = 5$, $\mu_1 = 1_5, \Sigma_1 = 2I_5$	Case: $n = 1000, d = 100$, $\mu_1 = 1_{200}, \Sigma_1 = 2I_{200}$	Case: $n = 200, d = 300$, $\mu_1 = 1_{300}, \Sigma_1 = 2I_{300}$
CI_{p^M}	DOC	99.2%	98.7%	99.1%
	IL	0.219	0.090	0.221
CI_{p^A}	DOC	96.5%	99.3%	94.7%
	IL	0.208	0.091	0.209
CI_{p^t}	DOC	88.9%	92.0%	87.9%
	IL	0.157	0.067	0.157
$CI_{p^{Ct}}$	DOC	99.1%	99.2%	99.3%
	IL	0.287	0.122	0.286
CI_{r^M}	DOC	98.6%	98.5%	97.2%
	IL	0.219	0.094	0.220
CI_{r^A}	DOC	93.3%	96.2%	91.9%
	IL	0.204	0.094	0.205
CI_{r^t}	DOC	94.5%	93.8%	91.2%
	IL	0.203	0.083	0.194
$CI_{r^{Ct}}$	DOC	99.6%	99.6%	98.9%
	IL	0.370	0.151	0.355

of confidence for credible intervals constructed based on the beta posterior distribution inferred by the K data sets corresponding to K confusion matrices from K-fold cross-validation all exceeded 95%. The confidence interval based on the corrected K-fold cross-validated t distribution elevated the degrees of confidence of those based on the K-fold cross-validated t distribution by correcting the variance of the t statistic.

Table 6: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall Based on Support Vector Machine Classifier.

		Case: $n = 200, d = 5$, $\mu_1 = 0.21_5, \Sigma_1 = I_5$	Case: $n = 1000, d = 200$, $\mu_1 = 0.21_{200}, \Sigma_1 = I_{200}$	Case: $n = 200, d = 300$ $\mu_1 = 0.21_{300}, \Sigma_1 = I_{300}$
CI_{p^M}	DOC	99.6%	99.2%	98.5%
	IL	0.254	0.073	0.162
CI_{p^A}	DOC	99.6%	95.2%	72.8%
	IL	0.234	0.075	0.172
CI_{p^t}	DOC	90.1%	93.9%	92.2%
	IL	0.163	0.055	0.117
$CI_{p^{Ct}}$	DOC	99.1%	99.4%	99.2%
	IL	0.297	0.100	0.213
CI_{r^M}	DOC	97.9%	98.4%	97.4%
	IL	0.253	0.073	0.162
CI_{r^A}	DOC	97.4%	92.7%	71.7%
	IL	0.226	0.076	0.172
CI_{r^t}	DOC	91.7%	94.9%	91.3%
	IL	0.222	0.062	0.136
$CI_{r^{Ct}}$	DOC	99.1%	99.5%	99.0%
	IL	0.405	0.113	0.248
		Case: $n = 200, d = 5$, $\mu_1 = 1_5, \Sigma_1 = 2I_5$	Case: $n = 1000, d = 200$, $\mu_1 = 0.21_{200}, \Sigma_1 = 3I_{200}$	Case: $n = 200, d = 300$, $\mu_1 = 0.21_{300}, \Sigma_1 = 3I_{300}$
CI_{p^M}	DOC	99.9%	100%	99.8%
	IL	0.204	0.092	0.204
CI_{p^A}	DOC	98.6%	99.7%	98.9%
	IL	0.196	0.092	0.193
CI_{p^t}	DOC	91.1%	93.5%	91.4%
	IL	0.144	0.061	0.127
$CI_{p^{Ct}}$	DOC	99.3%	99.8%	99.1%
	IL	0.262	0.111	0.230
CI_{r^M}	DOC	99.4%	98.6%	98.7%
	IL	0.194	0.078	0.147
CI_{r^A}	DOC	93.9%	95.2%	56.7%
	IL	0.190	0.080	0.165
CI_{r^t}	DOC	93.6%	94.5%	90.7%
	IL	0.169	0.066	0.120
$CI_{r^{Ct}}$	DOC	99.6%	99.7%	98.3%
	IL	0.309	0.121	0.219

However, the credible interval based on the average of the K beta posterior distributions from K-fold cross-validation returned somewhat ambivalent results. In 10 of the 30 cases, its degrees of confidence fell below 95%. One example is the situation represented by the case of $\mu_1 = 0.21_{300}$, $\Sigma_1 = I_{300}$ for a perceptron classifier in Table 4. In this case, its degree of

confidence was only 63.7%, which is far below 95%. This can be explained by the fact that this method merely adopts the average of the K results, and this average is significantly affected by a poor result. That is, this method is less robust than a credible interval constructed based on a beta posterior distribution inferred by the K data sets corresponding to K confusion matrices.

Indeed, the degree of confidence is not the only important consideration when choosing a statistical confidence (credible) interval. Another measure for the confidence (credible) interval is the interval length. From Tables 4, 5, and 6, we can see that the interval length of the confidence interval based on a K-fold cross-validated t distribution was the shortest among the four credible and confidence intervals. At the same time, however, it had the lowest degree of confidence.

It is thus important to consider how these two factors might be compromised. In general, when the degree of confidence is comparative, the confidence (credible) interval is always measured based on the interval length. That is, for a given degree of confidence, a fundamental principle for selecting the confidence (credible) interval is to select the one with the shortest interval length (Mao, Wang, & Pu, 2006; Shi, 2008; Shao, 2003). With an acceptable degree of confidence (above 95%), credible intervals of precision based on an average of the K beta posterior distributions had a shorter or comparable interval length compared to those based on the beta posterior distribution inferred by K data sets. Moreover, the interval lengths of these two credible intervals were both shorter than those based on the corrected K-fold cross-validated t distribution. Consider, for example, the case of $\mu_1 = 1_5$, $\Sigma_1 = 2I_5$ in Table 6, classified using a support vector machine classifier. In this case, the interval length for the confidence interval of precision based on the corrected K-fold cross-validated t distribution was 0.262. However, the interval lengths for credible intervals based on the beta posterior distribution inferred by K data sets and the average of posterior distributions were 0.204 and 0.196, respectively.

In particular, when the sample size increased, there was little change to the degree of confidence for the credible and confidence intervals. However, their interval lengths decreased by approximately half.

Remark 4. In the extreme case where precision and recall were 1, the degree of confidence was 0 with the proposed credible intervals based on a K-fold cross-validated beta distribution with a confidence level of $\alpha = 0.05$. This was demonstrated in the case where $\mu_1 = 1_5$, $\Sigma_1 = 0.1I_5$, and $n = 1000$ for a support vector machine classifier. Such a situation obtained because the $1 - \alpha/2$ quantile of the beta distribution does not exceed 1 when $\alpha = 0.05$. Thus, the credible intervals do not include the true value of 1.

In fact, in this special case, precision and recall are fixed, not random variables, and thus the credible interval has degenerated into the confidence interval of frequentist statistics. Furthermore, because the precision

Table 7: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall at $n = 200$ for Letter Recognition Data.

		CI_{p^M}	CI_{p^A}	CI_{p^t}	$CI_{p^{Ct}}$	CI_{r^M}	CI_{r^A}	CI_{r^t}	$CI_{r^{Ct}}$
Classification tree classifier	DOC	99.1%	98.0%	90.6%	98.9%	98.2%	95.4%	93.6%	99.5%
	IL	0.238	0.221	0.168	0.306	0.237	0.215	0.220	0.403
Perceptron classifier	DOC	99.2%	97.7%	94.4%	99.8%	98.3%	93.9%	95.6%	99.9%
	IL	0.234	0.218	0.168	0.306	0.234	0.213	0.221	0.404
Support vector machine classifier	DOC	99.7%	99.5%	90.6%	99.2%	95.7%	94.1%	90.5%	99.4%
	IL	0.237	0.221	0.154	0.282	0.235	0.215	0.212	0.388

and recall values were all equal to 1 in all replicated experiments, the variance of the estimation was zero. Thus, the traditional symmetrical interval estimation will actually degenerate into a point estimation.

4.3 Comparison of Credible and Confidence Intervals on Real Data. Two data sets from the UCI database, letter recognition data and MAGIC gamma telescope data, were considered in this section (Frey & Slate, 1991; Heck, Knapp, Capdevielle, & Thouw, 1998). Letter recognition data for identifying the letters of the roman alphabet comprise 20,000 examples described by 16 features. The 26 letters represent 26 categories, similar to Nadeau and Bengio (2003) and Wang et al. (2014), who turned it into a two-class (A–M versus N–Z) classification problem. In the MAGIC gamma telescope data, depending on the energy of the primary gamma, 10 features are allowed to discriminate statistically those caused by primary gammas (signal) from the images of hadronic showers initiated by cosmic rays in the upper atmosphere (background). We sampled, with replacement, 200 (1000) examples from the 20,000 (13,376) examples available in the letter recognition and the MAGIC gamma telescope data, respectively. Repeating this 1000 times, we then computed the degrees of confidence and interval lengths of the four credible and confidence intervals based on 1000 sets of data obtained.

As with the simulated data, the credible intervals constructed based on the beta posterior distribution inferred by the K data sets corresponding to K confusion matrices from K-fold cross-validation for precision and recall resulted in a considerable degree of confidence in almost all cases, as shown in Tables 7, 8, 9, and 10. One exceptional to this obtained when $n = 1000$ with the perceptron classifier. In this case, their degrees of confidence for recall were merely 83.5% and 87.7% for the letter recognition and the MAGIC gamma telescope data, respectively.

For a precision measure, the credible interval based on the average of the K beta posterior distributions from K-fold cross-validation all had

Table 8: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall at $n = 1000$ for Letter Recognition Data.

		CI_{pM}	CI_{pA}	CI_{pt}	CI_{pCt}	CI_{rM}	CI_{rA}	CI_{rt}	CI_{rCt}
Classification tree classifier	DOC	97.2%	96.6%	90.9%	98.7%	98.0%	89.7%	85.9%	99.1%
	IL	0.100	0.100	0.077	0.140	0.103	0.102	0.102	0.186
Perceptron classifier	DOC	96.0%	96.3%	93.3%	99.6%	83.5%	88.3%	87.3%	97.6%
	IL	0.096	0.096	0.086	0.157	0.095	0.094	0.131	0.240
Support vector machine classifier	DOC	98.4%	99.1%	83.5%	97.7%	96.7%	97.7%	91.7%	99.1%
	IL	0.113	0.113	0.069	0.126	0.114	0.113	0.099	0.181

Table 9: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall at $n = 200$ for MAGIC Gamma Telescope Data.

		CI_{pM}	CI_{pA}	CI_{pt}	CI_{pCt}	CI_{rM}	CI_{rA}	CI_{rt}	CI_{rCt}
Classification tree classifier	DOC	98.9%	98.5%	87.8%	99.2%	98.6%	91.0%	94.2%	99.7%
	IL	0.226	0.213	0.165	0.302	0.229	0.211	0.209	0.381
Perceptron classifier	DOC	99.3%	98.0%	90.9%	99.0%	95.3%	84.5%	91.1%	99.1%
	IL	0.245	0.229	0.182	0.332	0.244	0.217	0.254	0.463
Support vector machine classifier	DOC	97.3%	97.9%	81.3%	96.2%	98.1%	95.2%	92.2%	99.2%
	IL	0.240	0.224	0.164	0.300	0.241	0.219	0.208	0.379

Table 10: Degrees of Confidence and Interval Lengths of Credible and Confidence Intervals for Precision and Recall at $n = 1000$ for MAGIC Gamma Telescope Data.

		CI_{pM}	CI_{pA}	CI_{pt}	CI_{pCt}	CI_{rM}	CI_{rA}	CI_{rt}	CI_{rCt}
Classification tree classifier	DOC	97.6%	97.0%	88.4%	98.6%	95.2%	95.5%	91.9%	99.2%
	IL	0.096	0.097	0.074	0.135	0.103	0.103	0.095	0.174
Perceptron classifier	DOC	98.4%	98.6%	93.5%	99.6%	87.7%	85.2%	93.3%	99.5%
	IL	0.105	0.106	0.083	0.152	0.105	0.103	0.139	0.253
Support vector machine classifier	DOC	99.6%	99.5%	90.3%	99.4%	99.3%	98.7%	93.3%	99.9%
	IL	0.114	0.114	0.072	0.132	0.114	0.113	0.101	0.185

acceptable degrees of confidence with the two sample sizes in two real data sets. For recall, however, in 7 of 12 cases, the degree of confidence fell below 95%. Similarly, the confidence interval based on the *K*-fold cross-validated *t* distribution exhibited a degraded degree of confidence.

With an acceptable degree of confidence (above 95%), the credible interval based on the average of the K beta posterior distributions had the shortest interval length compared with the other confidence and credible intervals. In particular, when $n = 1000$, with an acceptable degree of confidence, the intervals were of comparable length to credible intervals based on the beta posterior distribution inferred by the K data sets and based on the average of the K beta posterior distributions whether for the letter recognition data or for the MAGIC gamma telescope data. Specifically, the intervals based on the beta distribution were 71.4%, 61.1%, 89.6%, 71.1%, 69.1%, and 86.4% of the interval length of confidence intervals based on the corrected K-fold cross-validated t distribution for the classification tree, perceptron, and support vector machine classifiers in the letter recognition and the MAGIC gamma telescope data, respectively.

The results in Tables 7 to 10 from two real data sets showed that the interval length also decreased by half as the sample size changed from 200 to 1000. This implies that the sample size had a significant impact on the interval length of the confidence (credible) interval.

4.4 Average Ranks of Four Credible and Confidence Intervals. To further investigate this problem, we compared the average ranks of four credible and confidence intervals with regard to their degree of confidence and interval length in all 27 cases of simulated and real data sets. Table 11 showed the results based on the simulated data sets with 15 cases and real data sets with 12 cases.

The average rank of the confidence interval based on the K-fold cross-validated t distribution was ranked first for interval length, but it ranked last among all four methods for the degree of confidence. By contrast, the confidence interval based on the corrected K-fold cross-validated t distribution ranked first for degree of confidence, but it ranked last for interval length. The two credible intervals proposed in this letter lay between the confidence intervals based on the K-fold and the corrected K-fold cross-validated t distributions. With an acceptable degree of confidence, the average ranks of our methods were first and second, and they were superior to the confidence interval based on the corrected K-fold cross-validated t distribution. The reason for this occurrence was the fact that the degrees of confidence of the confidence interval based on the K-fold cross-validated t distribution were all less than 95%.

4.5 Choice of ω. In the construction of the credible interval based on the beta posterior distribution inferred by the K data sets corresponding to K confusion matrices from K-fold cross-validation for precision and recall, the choice of ω is very important. Poor ω may affect the degree of confidence and interval length of the credible interval. Thus, in this section, we experimentally studied the changes in the degree of confidence and interval length as the changes of the values of the ω. Experimental results are in Table 12.

Table 11: Average Ranks of Four Credible and Confidence Intervals.

	Rank			
	CI_{p^M}	CI_{p^A}	CI_{p^t}	$CI_{p^{Ct}}$
Case			DOC	
1	1	2	4	3
2	1	3	4	1
3	1	4	3	2
4	2	1	4	3
5	1	1	4	3
6	1	2	4	3
7	1	3	4	2
8	3	1	4	2
9	2	3	4	1
10	1	1	4	3
11	2	3	4	1
12	2	4	3	1
13	1	3	4	2
14	1	3	4	2
15	1	3	4	2
16	1	3	4	2
17	2	3	4	1
18	1	2	4	3
19	2	3	4	1
20	3	2	4	1
21	2	1	4	3
22	2	3	4	1
23	1	3	4	2
24	2	1	4	3
25	2	3	4	1
26	3	2	4	1
27	1	2	4	3
Average rank	1.6(1)	2.4(3)	3.9(4)	2(2)
			IL	
Average rank	2.5(3)	2.2(2)	1(1)	4(4)
	CI_{r^M}	CI_{r^A}	CI_{r^t}	$CI_{r^{Ct}}$
			DOC	
Average rank	2.3(2)	3.1(3)	3.6(4)	1.0(1)
			IL	
Average rank	2.6(3)	1.8(2)	1.4(1)	4(4)

Table 12: Changes of the Degree of Confidence and Interval Length as the Changes of the Values of the ω.

		Letter Recognition Data, $n = 200$ Classification Tree Classifier		Letter Recognition Data, $n = 200$ Support Vector Machine Classifier		MAGIC Gamma Telescope Data, $n = 200$ Classification Tree Classifier	
ω		DOC	IL	DOC	IL	DOC	IL
$1/K = 0.10$	CI_{pM}	100%	0.504	100%	0.501	100%	0.488
	CI_{rM}	100%	0.503	100%	0.499	100%	0.490
$(K-1)/2K = 0.45$	CI_{pM}	99.8%	0.262	99.9%	0.261	99.8%	0.250
	CI_{rM}	98.8%	0.261	96.4%	0.259	98.7%	0.252
$(K+1)/2K = 0.55$	CI_{pM}	99.1%	0.238	99.7%	0.237	98.9%	0.226
	CI_{rM}	98.2%	0.237	95.7%	0.235	98.6%	0.229
0.65	CI_{pM}	98.6%	0.220	99.3%	0.220	98.4%	0.210
	CI_{rM}	96.8%	0.219	94.0%	0.218	97.4%	0.213
0.75	CI_{pM}	98.2%	0.205	98.5%	0.205	97.5%	0.196
	CI_{rM}	95.6%	0.204	92.1%	0.204	94.5%	0.198
0.85	CI_{pM}	97.6%	0.193	97.1%	0.193	96.8%	0.185
	CI_{rM}	92.5%	0.192	89.6%	0.191	91.6%	0.187
0.95	CI_{pM}	96.3%	0.183	97.0%	0.183	95.8%	0.175
	CI_{rM}	92.7%	0.182	84.9%	0.182	92.8%	0.177
1	CI_{pM}	95.5%	0.178	96.6%	0.178	93.8%	0.170
	CI_{rM}	91.5%	0.178	87.6%	0.177	92.2%	0.172

Table 12 shows that when ω increased, the degree of confidence and the interval length of the credible interval gradually decreased. In general cases, we opt to select an ω such that the credible interval has an accepted degree of confidence (larger than 95%) and a short interval length. However, the best ω cannot express a closed form because the correlations of the TPs, FPs, and FNs vary in different cases with different classifiers and data sets. For example, the best ω was $(K+1)/2K = 0.55$ for $K = 10$ in the case of letter recognition data, $n = 200$, support vector machine classifier. However, the best ωs were 0.65 and 0.75 in the cases of letter recognition data, $n = 200$, classification tree classifier, and MAGIC gamma telescope data, $n = 200$, classification tree classifier, respectively. To determine the best ω, the entire interval from $1/K$ to 1 should be searched, an expensive computation. Considering this condition, we suggested the computation of ω through $(K+1)/2K$. Although this selection method may not select the best ω, it provides a solution that is close to the best ω with a closed form and greatly saves on computational costs.

5 Conclusion

Considering that the commonly used confidence interval based on a K-fold cross-validated t distribution suffers from a lower degree of confidence, we presented a novel way to construct credible intervals indirectly, based on the posterior distributions of precisions and recall. Two credible intervals based on a K-fold cross-validated beta posterior distribution were thus proposed.

Furthermore, we compared our proposed credible intervals with existing confidence intervals for precision and recall through simulated and real data experiments. With an acceptable degree of confidence, our methods outperformed these existing methods. Specifically, they exhibited shorter interval lengths in all cases. The first proposed credible interval is particularly recommended, given that it displayed high degrees of confidence and short interval lengths in almost all experiments.

One of the key uses of performance metrics is model (algorithm) selection, which is traditionally straightforward to do based on point estimations, but how would this be done based on the performance intervals proposed? When the credible interval is used to select the models A and B, if their credible intervals are uncrossed, the model with high precision (recall) should be selected. However, if the credible interval of precision of A completely contains that of B, we cannot directly provide a definitive conclusion and need further analysis. For example, we can select models by directly comparing their right or left intervals. However, is this appropriate? The use of the proposed credible interval in comparing models is currently being investigated.

In practical applications, we always need to take into consideration the two factors of precision and recall. This enables the construction of a utility function that directly captures the value of true positives and negatives

versus the cost of false positives and negatives. The ROC curve is a useful tool that facilitates choosing an optimal classification threshold for a given application. For quantitatively evaluating the model performance, an AUC measure obtained based on the ROC curve is often used. However, the AUC measure remains a point estimation. How the credible interval of this measure can be constructed by analyzing the distribution of AUC is meaningful research work and our future research direction.

Acknowledgments

This work was supported by National Natural Science Fund of China (61503228, 71503151) and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund. Experiments were supported by High Performance Computing System of Shanxi University.

References

Alpaydin, E. (1999). Combined 5×2 cv F test for comparing supervised classification learning algorithms. *Neural Computation*, *11*, 1885–1892.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of *K*-fold cross-validation. *Journal of Machine Learning Research*, *5*, 1089–1105.

Bisani, M., & Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Piscataway, NJ: IEEE.

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1924.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. London: Chapman and Hall.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Flach, P. A. (2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings of 20th International Conference on Machine Learning* (pp. 194–201). Menlo Park, CA: AAAI Press.

Frey, P. W., & Slate, D. J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, *6*, 161.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proceedings of European Colloquium on IR Resarch* (pp. 345–359). New York: Springer.

Grandvalet, Y., & Bengio, Y. (2006). *Hypothesis testing for cross-validation* (Technical Report). Montreal: University of Montreal.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Heck, D., Knapp, J., Capdevielle, J. N., & Thouw, T. (1998). *CORSIKA: A Monte Carlo code to simulate extensive air showers*. Karlsruhe: Forschungszentrum Karlsruhe GmbH.

Keller, M., Bengio, S., & Wong, S. Y. (2006). Benchmarking non-parametric statistical tests. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems*, *18*. Cambridge, MA: MIT Press.

Lobo, J. M., Jimenez, V. A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*, 145–151.

Mao, S., Wang, J., & Pu, X. (2006). *Advanced mathematical statistics*. Beijing: Higher Education Press.

Markatou, M., Tian, H., Biswas, S., & Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, *6*, 1127–1168.

Moreno-Torres, J., Saez, J.& Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*, 1304–1312.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*, 239–281.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and Correlation. *Journal of Machine Learning Technologies*, *2*, 37–63.

Shao, J. (2003). *Mathematical statistics*. New York: Springer.

Shi, N. (2008). *Statistical test theory and method*. Beijing: Science Press.

Wang, Y., Li, J., Li, Y., Wang, R., & Yang, X. (2015). Confidence interval for F1 measure of algorithm performance based on blocked 3×2 cross-validation. *IEEE Transactions on Knowledge and Data Engineering*, *27*, 651–659.

Wang, Y., Wang, R., Jia, H., & Li, J. (2014). Blocked 3×2 cross-validated *t*-test for comparing supervised classification learning algorithms. *Neural Computation*, *26*, 208–235.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42–49). New York: ACM.

Yildiz, O. T. (2013). Omnivariate rule induction using a novel pairwise statistical test. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 2105–2118.

Yildiz, O. T., Aslan, O., & Alpaydin, E. (2011). Multivariate statistical tests for comparing classification algorithms. *Lecture Notes in Computer Science*: Vol. *6683*. *Learning and Intelligent Optimization* (pp. 1–15). New York: Springer.

Received October 9, 2015; accepted March 29, 2016.