Confidence Interval for F_1 Measure of Algorithm Performance Based on Blocked 3×2 Cross-Validation

Yu Wang, Jihong Li, Yanfang Li, Ruibo Wang, and Xingli Yang

Abstract—In studies on the application of machine learning such as Information Retrieval (IR), the focus is typically on the estimation of the F_1 measure of algorithm performance. Approximate symmetrical confidence intervals constructed by the F_1 value based on cross-validated t distribution are commonly used in the literature. However, theoretical analysis on the distribution of F_1 values shows that such distribution is actually non-symmetrical. Thus, simply using symmetrical distribution to approximate non-symmetrical distribution may be inappropriate and may result in a low degree of confidence and long interval length for the confidence interval. In the present study, a non-symmetrical confidence interval of the F_1 measure based on Beta prime distribution is constructed by using the F_1 value computed based on the average confusion matrix of a blocked 3×2 cross-validation. Experimental results show that in most cases, our method has high degrees of confidence. With an acceptable degree of confidence, our method has a shorter interval length than the approximate symmetrical confidence intervals based on the blocked 3×2 and 5×2 cross-validated t distributions. The approximate symmetrical confidence interval based on the 10-fold cross-validated t distribution has the shortest interval length of the four confidence intervals but with low degrees of confidence in all cases. Taking these two factors into consideration, our method is recommended.

Index Terms—Blocked 3 × 2 cross-validation, F_1 measure, confidence interval, Beta prime distribution

1 Introduction

In applications of machine learning such as Information Retrieval (IR) or natural language processing (NLP), the standard measure for algorithm performance is the F_1 measure, which is defined as the harmonic average of precision and recall. When applying a learning algorithm, the focus is typically on the estimation of the F_1 measure of algorithm performance. The point estimation is rather trivial, and the estimation accuracy is always measured based on the mean square error. Confidence interval is a simpler and more intuitive method than point estimation in terms of measuring estimation accuracy ([1], [2]). Researchers often test the significance of the differences between two classifiers in comparing classifiers based on the F_1 measure by examining whether the corresponding confidence intervals cross ([3], [4]).

Thus, for effectively measuring the performance of algorithm, it is very important to construct a faithful confidence interval with a high degree of confidence (DOC) and short interval length (IL). The degree of confidence of a confidence interval is the probability of the inclusion of the F_1 true value in the confidence interval. Interval length indicates the accuracy of the confidence interval. Approximate symmetrical confidence intervals for the F_1 measure based on the F_1 value computed by using the confusion matrix are commonly used in the literature. However, theoretical analysis on the distribution of F_1 values shows that such distribution is actually non-symmetrical. Thus, the use of symmetrical distribution, such as the commonly used t distribution, may be inappropriate to approximate non-symmetrical distribution.

In practice, to be able to eliminate the effect by chance (e.g., variance due to small changes in the training set), typically, one does training and validation a number of times, possibly by various forms of cross-validation ([3], [5], [6], [7], [8], [9], [10]). Thus, confidence intervals based on cross-validations are often used to estimate algorithm performances for the F_1 measure. In particular, [10] proposed a new blocked 3×2 cross-validation and demonstrated that it is comparable with the test based on the 5×2 cross-validation but with less computation complexity, and the 5×2 cross-validated test is slightly more powerful than the 10-fold cross-validated t-test shown by [5]. Thus, we apply blocked 3×2 cross-validation in this study for the F_1 measure. An exact percentile of the F_1 value based on the average confusion matrix of the blocked 3×2 cross-validation and the corresponding non-symmetrical confidence interval of the F_1 measure based on Beta prime distribution are provided through a theoretical analysis of the F_1 value distribution.

The remainder of this study is organized as follows. Section 2 defines the standard F_1 measure of algorithm performance and the F_1 value estimation based on the confusion matrices of the blocked 3×2 cross-validation. Non-symmetrical confidence interval based on the blocked 3×2 cross-validated Beta prime distribution

- *Y. Wang, J. Li, and R. Wang are with the Computer Center of Shanxi University, Taiyuan 030006, P.R. China. E-mail: {wangyu, lijh, wangruibo}@sxu.edu.cn.*
- *Y. Li and X. Yang are with School of Mathematical Sciences, Shanxi University. E-mail: {liyanfang, yangxingli}@sxu.edu.cn.*

Manuscript received 22 Jan. 2014; revised 22 July 2014; accepted 3 Sept. 2014. Date of publication 21 Sept. 2014; date of current version 28 Jan. 2015.
Recommended for acceptance by G. Cormode.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2014.2359667

proposed in this study and approximate symmetrical confidence intervals based on the blocked 3×2 cross-validated t distribution, the 5×2 cross-validated t distribution, and the 10-fold cross-validated t distribution are described in Sections 3 and 4, respectively. Section 5 discusses the simulated experiments that show how the confidence intervals behaves compare. Section 6 concludes the study.

2 Measures of Performance

In studies on machine learning, multiple measures can be employed to assess the performance of learning algorithms. These measures include accuracy, precision, recall, F-score, Receiver operating characteristics (ROC) and Area under the ROC curve (AUC) ([11], [12], [13]). In this study, we focus on a commonly used performance indicator called F_1 measure, which refers to the harmonic average of precision and recall.

2.1 Standard F_1 Measure

Without loss of generality, we only consider the following simple setting (two class classification problem): each object is associated with a binary label l which accounts for the correctness of the object with respect to the task at hand. In addition, the classification algorithm produces an prediction z indicating whether it believes the object to be correct or not. Then precision may be defined as the probability that an object is relevant given that it is returned by the system, while the recall is the probability that a relevant object is returned (see [14]):

$$p = P(l = +|z = +), \quad r = P(z = +|l = +). \tag{1}$$

In order to summarize these two values, it is common to consider the so-called F_1 measure. It is described as the harmonic average of precision and recall:

$$F_1 = \left(\frac{1}{2}\left[\frac{1}{p} + \frac{1}{r}\right]\right)^{-1} = \frac{2 \cdot p \cdot r}{p + r}. \tag{2}$$

2.2 F_1 Value Based on Confusion Matrix

For a specific two class classification problem, the experimental outcome may be conveniently summarised in a 2×2 confusion matrix:

		Pridicted +	class −
True	+	TP	FN
class	−	FP	TN

where TP (resp. TN) is the number of true positives (resp. negatives) and FP (resp. FN) the number of false positives (resp. negatives). From these counts, one can obtain the precision (p), recall (r) and F_1 value:

$$p = \frac{TP}{TP + FP}, \quad r = \frac{TP}{TP + FN},$$

$$F_1 = \frac{2\ TP}{2\ TP + FP + FN} \tag{3}$$

TABLE 1
Blocked 3×2 Cross-Validation

Group	Training set	Test set	$\hat{\mu}_k^{(i)}$
1	$D_1^{(1)} = (P_1, P_2)$	$T_1^{(1)} = (P_3, P_4)$	$\hat{\mu}_1^{(1)}$
1	$D_2^{(1)} = (P_3, P_4)$	$T_2^{(1)} = (P_1, P_2)$	$\hat{\mu}_2^{(1)}$
2	$D_1^{(2)} = (P_1, P_3)$	$T_1^{(2)} = (P_2, P_4)$	$\hat{\mu}_1^{(2)}$
2	$D_2^{(2)} = (P_2, P_4)$	$T_2^{(2)} = (P_1, P_3)$	$\hat{\mu}_2^{(2)}$
3	$D_1^{(3)} = (P_1, P_4)$	$T_1^{(3)} = (P_2, P_3)$	$\hat{\mu}_1^{(3)}$
3	$D_2^{(3)} = (P_2, P_3)$	$T_2^{(3)} = (P_1, P_4)$	$\hat{\mu}_2^{(3)}$

It is obvious that the F_1 value is an estimation of the theoretical F_1 measure.

2.3 Average F_1 Value Based on the Blocked 3×2 Cross-Validation

To test the significance of the differences between two algorithms, [5], [6], [9] proposed a random 5×2 cross-validation method based on loss function and demonstrated that its performance is (slightly) better than the 10-fold cross-validation by simulated experiments.

However, [10] found that an accurate theoretical expression of variance for a random 5×2 cross-validation can not be obtained, thereby causing difficulty in the corresponding variance estimation. Furthermore, they show that this is due to the different number of overlapped samples between training sets in five replications of 5×2 cross-validation. [8] theoretically proved that the number of overlapped samples from any two training sets follows a hypergeometric distribution and that the mathematical expectation is $\frac{n}{4}$ (n is the sample size). Based on this, a new blocked 3×2 cross-validation was proposed. This method was asserted to mitigate the effect of the number of overlapped samples such that the covariance between different replications becomes theoretically identical. Simulated results demonstrated that it is comparable with the test based on the 5×2 cross-validation, but with less computation complexity.

In detail, the blocked 3×2 cross-validation relies on a preliminary partitioning of data into 4 blocks of approximately equal cardinality to implement three replications of two-fold cross-validation. Formally, the data set D is split into four disjoint and equal-sized blocks, denoted as $P_j, j = 1, 2, 3, 4$. The combination of any two P_js will result in three groups and six different combinations shown in Table 1.

Here, $D_k^{(i)}, i = 1, 2, 3, k = 1, 2$ denotes the training set, $T_k^{(i)}, i = 1, 2, 3, k = 1, 2$ denotes the test set. However, they are as a training or test set with each other, thus $D_1^{(i)} = T_2^{(i)}, D_2^{(i)} = T_1^{(i)}, i = 1, 2, 3$. The average F_1 value based on the blocked 3×2 cross-validation is defined as the average of the F_1 scores on three groups:

$$\hat{\mu}_{3\times 2} = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}\sum_{k=1}^{2}\hat{\mu}_k^{(i)} = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}\sum_{k=1}^{2}F_1(A(D_k^{(i)}), T_k^{(i)}), \tag{4}$$

where $F_1(A(D_k^{(i)}), T_k^{(i)})$ is the F_1 score returned by algorithm A trained on the set $D_k^{(i)}$ and tested on $T_k^{(i)}$ for $i = 1, 2, 3,\ k = 1, 2$.

3 Non-Symmetrical Confidence Interval for the F_1 Measure Based on the Blocked 3×2 Cross-Validated Beta Prime Distribution

In this section, a non-symmetrical confidence interval based on the blocked 3×2 cross-validated Beta prime distribution is provided by studying the distribution of the F_1 value based on the average confusion matrix of the blocked 3×2 cross-validation. [14] revealed that the distributions of precision (p) and recall (r) based on blocked 3×2 cross-validation have the following forms:

Lemma 1.

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

that is, $p \mid D \sim Be(TP+\lambda, FP+\lambda)$*(Beta distribution), where* $P(p)$ *is the prior distribution and* $p \sim Be(\lambda,\lambda), D = (TP, FP, FN, TN)$ *is the average of the corresponding elements of six confusion matrices of the blocked* 3×2 *cross-validation. A similar development yields the posterior distribution for the recall:* $r \mid D \sim Be(TP+\lambda, FN+\lambda)$.

Lemma 2. *Given two variables with Gamma distributions* $X \sim \Gamma(\alpha, h)$ *and* $Y \sim \Gamma(\beta, h)$*, with identical shape parameter* h*, three interesting properties hold:*

1) $\forall c > 0, cX \sim \Gamma(\alpha, ch)$;
2) $X + Y \sim \Gamma(\alpha+\beta, h)$;
3) $\frac{X}{X+Y} \sim Be(\alpha, \beta)$.

Lemma 2 is reported in [14]. For readers' convenience, it is listed here.

Proposition 1. *The density function of the* F_1 *value based on the blocked* 3×2 *cross-validation takes the form*

$$p(t) = \frac{2^a(1-t)^{a-1}(2-t)^{-a-b}t^{b-1}}{B(a,b)}, \; 0 < t < 1, \tag{5}$$

where B *is a Beta function with parameters* $a = FP + FN + 2\lambda$ *and* $b = TP + \lambda$.

Proof. Lemma 1 and 2 enable us to postulate that the posterior distributions of p and r, which are Beta distributions, arise from the combination of independent Gamma variates:

$$p = \frac{X}{X+Y}, \; r = \frac{X}{X+Z}$$

with $X \sim \Gamma(TP+\lambda, 1), Y \sim \Gamma(FP+\lambda, 1)$ and $Z \sim \Gamma(FN+\lambda, 1)$. Combining these in the F_1 value expression, and using the fact that $U = 2X$ is a Gamma variate and that $V = Y + Z$ is also a Gamma variate, we get:

$$F_1 = \frac{U}{U+V}$$

with $U \sim \Gamma(TP+\lambda, 2), V \sim \Gamma(FP+FN+2\lambda, 1)$.

Notably, $\frac{V}{U/2}$ follows a Beta prime distribution $Be'(FP+FN+2\lambda, TP+\lambda)$. We thus obtain the density function of $F_1 = \frac{U}{U+V} = \frac{1}{1+\frac{1}{2}\frac{V}{U/2}}$:

$$p(t) = \frac{2^a(1-t)^{a-1}(2-t)^{-a-b}t^{b-1}}{B(a,b)}, \; 0 < t < 1,$$

where B is a Beta function with parameters $a = FP + FN + 2\lambda$ and $b = TP + \lambda$. □

We can also deduce the percentile of the distribution of the F_1 value from the percentile of the Beta prime distribution, i.e.,

$$F_{1\alpha} = \frac{1}{1+\frac{1}{2}Be'_{1-\alpha}}. \tag{6}$$

The resulting confidence interval based on the Beta prime distribution is

$$CI_{BP(3\times2CV)} = [F_{1\alpha/2}, F_{11-\alpha/2}]. \tag{7}$$

Corollary 1. *The density function* $p(t)$ *is a unimodal function when* $a > 1$ *and* $b \geq 1$*. This function reaches the maximum at* $t = -(1/2)b - (1/4)a + (5/4) + (1/4)\sqrt{4b^2 + 4ab - 4b + a^2 - 10a + 9}$ *(mode).*

Corollary 2. *The density function* $p(t)$ *is non-symmetrical. For* $mode(F_1) = 0.5$*, the area under the function* $p(t)$ *at* t *between 0 and 0.5 is larger than that at* t *between 0.5 and 1, i.e., the expectation of the* F_1 *value is smaller than its mode at* $mode(F_1) = 0.5$.

Fig. 1 demonstrates the shape of the density curve of the F_1 value with the changes in parameters a and b. This finding further validates the unimodal and non-symmetrical properties of $p(t)$. Thus, Corollaries 1 and 2 and Fig. 1 also show that simply using symmetrical distribution, such as commonly used Normal or t distribution to approximate non-symmetrical distribution may be inappropriate.

The following simulated experiments further validate the extent of the non-symmetry of the density curve. Skewness is a standard measure that generally has the following two forms:

1) The skewness of a random variable X is the third standardized moment, defined as

$$SK_1 = \frac{E(X-E(X))^3}{(E(X-E(X))^2)^{3/2}}. \tag{8}$$

2) Pearson suggested a simple calculation as a measure of skewness:

$$SK_2 = \frac{E(X) - Mode(x)}{\sqrt{Var(X)}}. \tag{9}$$

Notably, the exact expressions of $E(X), E(X-E(X))^2$ and $E(X-E(X))^3$ can not be obtained and are thus replaced by sample central moments. For a sample of n values, we have the sample skewness

$$\widehat{SK_1} = \frac{\frac{n}{(n-2)(n-1)}\sum_{i=1}^{n}(x_i-\overline{x})^3}{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2)^{3/2}} \tag{10}$$

$$\widehat{SK_2} = \frac{\frac{\sum_{i=1}^{n}(x_i-\overline{x})}{n} - Mode(x)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}}. \tag{11}$$

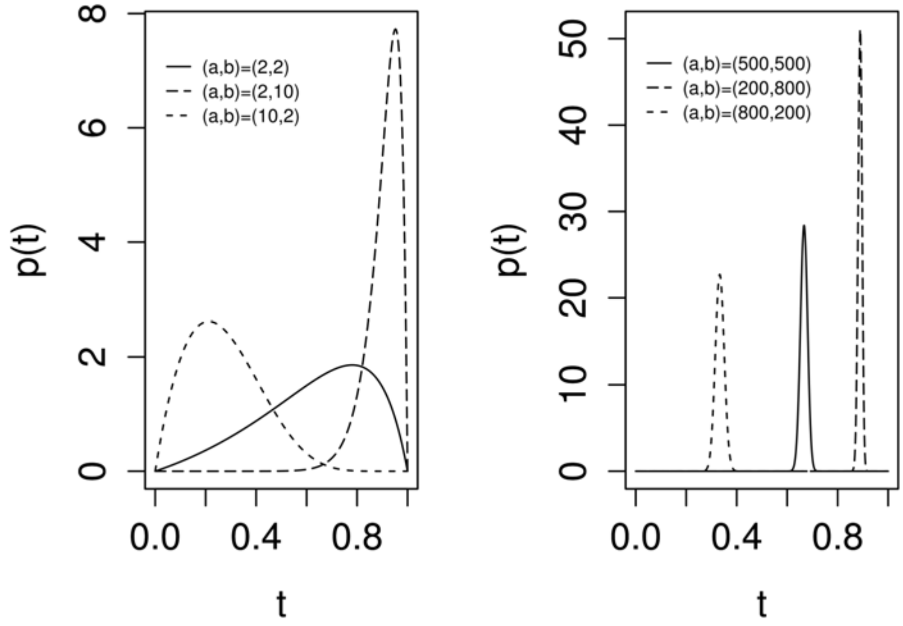

Fig. 1. Density curves of the F_1 value with different combinations of a and b.

Simulated experiment 3.1. Changes in skewness with varying sample sizes

First, we can obtain the observed TP, FP, FN and TN with different values p, r and $accuracy$. The parameters a and b are then obtained. Thus, the sample moments are computed based on 10,000 samples. The replicated times are 1,000. We now look at the changes in skewness for $n = 50, 200, 800, 2{,}000, 6{,}000$.

From Table 2, we know that with increasing sample capacity, the skewness $\widehat{SK}_1$ and $\widehat{SK}_2$ gradually decrease in all cases. This observation indicates that symmetrical distribution may be used to approximate non-symmetrical distribution in case of a large sample. However, large skewness occurs in a small sample.

Simulated experiment 3.2. Changes in skewness with varying modes

This simulation examines the changes of skewness in various cases for $mode = 0.3, 0.4, 0.45, 0.5, 0.6, 0.8, 0.9, 0.95$. Results are shown in Table 3.

Table 3 shows that the initial decrease and subsequent increase in skewness values correspond to an increase in modes. Skewness reaches the minimum when the mode value is between 0.40 and 0.50. In practical applications, we always aim to obtain a high F_1 value. However, the skewness becomes large as the mode of F_1 value increases. This finding also suggests that a large bias may arise from the use of symmetrical distribution to approximate non-symmetrical distribution at a high F_1 value.

4 Approximate Symmetrical Confidence Interval for the F_1 Measure

Approximate symmetrical confidence intervals (statistical test of significance) are widely used in the literature ([3], [15], [16], [17]). We present three different techniques to perform inference in this section. The approximate symmetrical confidence intervals based on Central Limit Theorem for the F_1 measure at confidence level $1 - \alpha$ will look like

$$F_1 \in \left[\hat{\mu} - c\sqrt{\hat{\sigma}^2}, \hat{\mu} + c\sqrt{\hat{\sigma}^2}\right]. \tag{12}$$

Notably, in Eq. (12), $\hat{\mu}$ is an estimator of the F_1 measure, $\hat{\sigma}^2$ is a variance estimator, and c is a percentile from Student's t distribution. The only difference between the three techniques is in the choice of $\hat{\mu}$, $\hat{\sigma}^2$ and c. We are now ready to introduce the approximate symmetrical confidence intervals we will consider in this paper.

Remark. Compared with the non-symmetrical confidence interval based on Beta prime distribution given in the previous section, the approximate symmetrical confidence

TABLE 2
Skewness Values with Different Experimental Setups for $n = 50, 200, 800, 2{,}000, 6{,}000$

	Case I		Case II		Case III		Case IV		Case V		Case VI	
n	$\widehat{SK}_2$	$\widehat{SK}_1$	$\widehat{SK}_2$	$\widehat{SK}_1$	$\widehat{SK}_2$	$\widehat{SK}_1$	$\widehat{SK}_2$	$\widehat{SK}_1$	$\widehat{SK}_2$	$\widehat{SK}_1$	$\widehat{SK}_2$	$\widehat{SK}_1$
50	−0.047	−0.079	−0.074	−0.128	−0.140	−0.253	−0.218	−0.416	−0.275	−0.528	−0.486	−0.950
200	−0.028	−0.052	−0.042	−0.080	−0.073	−0.142	−0.140	−0.225	−0.145	−0.288	−0.295	−0.590
800	−0.014	−0.027	−0.021	−0.043	−0.037	−0.073	−0.057	−0.115	−0.074	−0.148	−0.154	−0.309
2,000	−0.009	−0.019	−0.014	−0.027	−0.024	−0.047	−0.037	−0.075	−0.047	−0.093	−0.098	−0.198
6,000	−0.006	−0.010	−0.008	−0.016	−0.014	−0.027	−0.021	−0.043	−0.027	−0.054	−0.056	−0.116

where Case I: {p = r = accuracy = 0.5}, Case II: {p = 0.6, r = accuracy = 0.5}, Case III: {p = 0.7, r = accuracy = 0.6}, Case IV: {p = 0.8, r = 0.9, accuracy = 0.7}, Case V: {p = 0.9, r = accuracy = 0.8}, Case VI: {p = r = accuracy = 0.95}.

TABLE 3
Skewness Values with Different Experimental Setups for $mode = 0.3, 0.4, 0.45, 0.5, 0.6, 0.8, 0.9, 0.95$

	Case I		Case II		Case III		Case IV		Case V		Case VI	
$mode$	$\widehat{SK_2}$	$\widehat{SK_1}$	$\widehat{SK_2}$	$\widehat{SK_1}$	$\widehat{SK_2}$	$\widehat{SK_1}$	$\widehat{SK_2}$	$\widehat{SK_1}$	$\widehat{SK_2}$	$\widehat{SK_1}$	$\widehat{SK_2}$	$\widehat{SK_1}$
0.30	0.101	0.183	0.031	0.059	0.013	0.027	0.007	0.014	0.055	0.008	0.004	0.001
0.40	0.020	0.043	0.006	0.014	0.002	0.004	0.001	0.002	0.000	0.003	0.000	0.003
0.45	−0.022	−0.027	−0.007	−0.019	−0.000	−0.009	−0.003	−0.004	−0.002	−0.004	−0.001	−0.003
0.50	−0.062	−0.095	−0.019	−0.041	−0.011	−0.019	−0.004	−0.011	−0.003	−0.002	−0.000	−0.008
0.60	−0.153	−0.250	−0.048	−0.094	−0.020	−0.044	−0.012	−0.021	−0.007	−0.013	−0.005	−0.006
0.80	−0.374	−0.647	−0.124	−0.249	−0.057	−0.116	−0.027	−0.057	−0.016	−0.037	−0.014	−0.024
0.90	−0.551	−0.993	−0.202	−0.402	−0.094	−0.181	−0.047	−0.090	−0.025	−0.052	−0.020	−0.042
0.95	−0.703	−1.295	−0.293	−0.578	−0.135	−0.271	−0.068	−0.139	−0.038	−0.080	−0.030	−0.058

where Case I: {b = 10}, Case II: {b = 100}, Case III: {b = 500}, Case IV: {b = 2,000}, Case V: {b = 6,000}, Case VI: {b = 10,000}.

interval may significantly affect the estimation accuracy of the F_1 measure in some cases. Notably, the F_1 value is between 0 and 1, however, the approximate symmetrical confidence interval may exceed the range of $(0, 1)$. For example, the simulated experiment in Section 5 shows that the approximate symmetrical confidence interval based on a blocked 3×2 cross-validated t distribution is $[0.84, 1.03]$, which obviously exceeds the limit value of 1 in the case of $\mu_1 = (1, 1)$, $\Sigma_1 = 0.1 I_2$, $n = 200$, and logistic regression (LR) classifier.

4.1 Approximate Symmetrical Confidence Interval Based on the K-Fold Cross-Validated t Distribution

K-fold cross-validation is probably the simplest and most widely used resampling method. It uses all available examples as training and test examples, mimics K training and test sets by using part of the data to fit the model, and a different part to test it. ([7], [10])

First, the data set D is split into K disjoint and equal-sized blocks, which is denoted as $T_k, k = 1, 2, \ldots, K$. Let D_k be the training set obtained by removing the elements in T_k from D, then the average F_1 value based on the K-fold cross-validation has the following form:

$$\hat{\mu}_K = \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_k = \frac{1}{K}\sum_{k=1}^{K} F_1(A(D_k), T_k), \tag{13}$$

where $F_1(A(D_k), T_k)$ is the F_1 score returned by algorithm A trained on the set D_k and tested on T_k for $k = 1, \ldots, K$.

The K-fold cross-validated t statistic considers $\hat{\mu} = \hat{\mu}_K$ to estimate the F_1 measure and $\hat{\sigma}^2 = \frac{S^2_{\hat{\mu}_k}}{K}$, where $S^2_{\hat{\mu}_k}$ is the sample variance of $\hat{\mu}_k$. If we assume that $\hat{\mu}_k$ were drawn independently from a normal distribution, then the confidence interval based on the K-fold cross-validation can be written as

$$CI_{t(KCV)} = \Big[\hat{\mu}_K - t_{K-1,1-\alpha/2}\sqrt{S^2_{\hat{\mu}_k}/K},\; \hat{\mu}_K + t_{K-1,1-\alpha/2}\sqrt{S^2_{\hat{\mu}_k}/K}\Big], \tag{14}$$

where $c = t_{K-1,1-\alpha/2}$ is the percentile from Student's t distribution with degree of freedom $K - 1$.

Remak. The difference between the variances of the K-fold cross-validation based on the F_1 measure and the loss function is that the variance of the K-fold cross-validation based on the F_1 measure can be only decomposed to block (T_k). However, the variance of the K-fold cross-validation can be decomposed to sample for the loss function.

4.2 Approximate Symmetrical Confidence Interval Based on the 5×2 Cross-Validated t Distribution

Dietterich [5] pointed out that the variance of the K-fold cross-validation can be underestimated due to the overlapping of training sets. Thus, [5] proposed a random 5×2 cross-validation method based on five replications of two-fold cross-validation and show that it is slightly more powerful than the 10-fold cross-validated t-test. This method is further studied by [6], [9]. In each replication of the 5×2 cross-validation, the available data are randomly partitioned into two equal-sized sets $T_1^{(i)}$ and $T_2^{(i)}, i = 1, \ldots, 5$. Each learning algorithm is trained on each set and tested on the other set. This produces cross-validated estimators based on the F_1 measure: $F_1\ (A(T_1^{(i)}), T_2^{(i)})$ and $F_1(A(T_2^{(i)}), T_1^{(i)}), i = 1, \ldots, 5$. The average F_1 value of five group replications has the following form

$$\hat{\mu}_{5\times2} = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{2}\big(F_1\big(A\big(T_1^{(i)}\big), T_2^{(i)}\big) + F_1\big(A\big(T_2^{(i)}\big), T_1^{(i)}\big)\big). \tag{15}$$

Let $S_i^2 = (F_1(A(T_1^{(i)}), T_2^{(i)}) - F_1^{(i)})^2 + (F_1(A(T_2^{(i)}), T_1^{(i)}) - F_1^{(i)})^2$ be the sample variance computed from the ith replication, where $F_1^{(i)} = \frac{1}{2}(F_1(A(T_1^{(i)}), T_2^{(i)}) + F_1(A(T_2^{(i)}), T_1^{(i)}))$. If let $\hat{\mu} = \hat{\mu}_{5\times2}, \hat{\sigma}^2 = \frac{\sum_{i=1}^{5} S_i^2}{5}$, under the assumption of normality the resulting confidence interval is,

$$CI_{t(5\times2CV)} = \Bigg[\hat{\mu}_{5\times2} - t_{5,1-\alpha/2}\sqrt{\sum_{i=1}^{5} S_i^2/5},\; \hat{\mu}_{5\times2} + t_{5,1-\alpha/2}\sqrt{\sum_{i=1}^{5} S_i^2/5}\Bigg]. \tag{16}$$

4.3 Approximate Symmetrical Confidence Interval Based on the Blocked 3×2 Cross-Validated t Distribution

Wang et al. [10] proposed a new test method called blocked 3×2 cross-validated t-test based on the 5×2 cross-validated tests given by [5], [6], [9] and demonstrated that its performance is comparable with the test based on the 5×2 cross-validation, but with less computation complexity. They used $\hat{\mu} = \hat{\mu}_{3\times 2}$, $\hat{\sigma}^2 = \frac{\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}_{3\times 2})^2}{6}$, leading to corresponding confidence interval of

$$CI_{t(3\times 2CV)} = \Bigg[\hat{\mu}_{3\times 2} - t_{5,1-\alpha/2}\sqrt{\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}_{3\times 2})^2/6}, \quad \hat{\mu}_{3\times 2} + t_{5,1-\alpha/2}\sqrt{\sum_{i=1}^{3}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}_{3\times 2})^2/6}\Bigg]. \tag{17}$$

5 Simulated Experiments for Comparison

This section performs a simulation study based on simulated and real Letter Recognition data sets to investigate the degree of confidence and the interval length of the four confidence intervals considered in the previous two sections for multiple classifiers, such as logistic regression, k nearest neighbor (KNN), naive Bayes (NB), classification tree (CT) and support vector machine (SVM). For a given problem, we generate 1,000 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 take $K = 10$ in the approximate symmetrical confidence interval based on the K-fold cross-validated t distribution. We choose $\lambda = 1$, the uniform prior, in the non-symmetrical confidence interval based on the blocked 3×2 cross-validated Beta prime distribution. The number of training samples is $n = 200$, 600, 6,000 for simulated data and $n = 200$, 600 for real Letter Recognition data. The number of test samples is five times that of the training samples. The confidence level $1 - \alpha = 0.95$, i.e., $\alpha = 0.05$.

5.1 Simulated Data Comparison

Considering a classification problem with two classes, we have $Z = (X, Y)$, with $Prob(Y = 1) = Prob(Y = 0) = \frac{1}{2}$, $X \mid Y = 0 \sim N(\mu_0, \Sigma_0), X \mid Y = 1 \sim N(\mu_1, \Sigma_1)$. Similar to [3], we take $\mu_0 = (0, 0), \Sigma_0 = I_2$, but take multiple μ_1 and Σ_1.

Tables 4, 5, and 6 show the simulated results of the degree of confidence and interval length. The non-symmetrical confidence interval based on the blocked 3×2 cross-validated Beta prime distribution has high degrees of confidence in most cases, i.e., the degree of confidence far exceeds 0.95. For example, in case I of Table 4, the degrees of confidence of non-symmetrical confidence interval based on the blocked 3×2 cross-validation are 99.5, 99.6, 98.2, 99.5, and 99.7 percent for the CT, LR, SVM, NB, and KNN classifiers, respectively.

For the CT classifier, the approximate symmetrical confidence interval based on the blocked 3×2 cross-validated t distribution has acceptable degrees of confidence in three cases and with three sample sizes. For the four other classifiers, it exhibits (somewhat) degraded degrees of confidence in 27 of the 36 cases for three sample sizes.

One example is the situation represented by Case II for LR classifier in Table 6. In this case, the degree of confidence of the approximate confidence interval based on the blocked 3×2 cross-validated t distribution is only 88.6 percent, which is far below 95 percent.

TABLE 4
Degrees of Confidence and Interval Lengths of the Four Confidence Intervals at $n = 200$ for Simulated Data

		Case I: $\mu_1 = (0.5, 0.5),\ \Sigma_1 = I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	90.9%	93.5%	98.2%	99.5%
	IL	0.167	0.299	0.276	0.219
LR	DOC	91.8%	95.0%	92.9%	99.6%
	IL	0.153	0.234	0.184	0.203
SVM	DOC	90.8%	94.4%	94.9%	98.2%
	IL	0.154	0.230	0.188	0.205
NB	DOC	83.1%	91.6%	97.4%	99.5%
	IL	0.153	0.227	0.185	0.204
KNN	DOC	87.3%	89.7%	94.6%	99.7%
	IL	0.161	0.217	0.209	0.219
		Case II: $\mu_1 = (1.5, 1.5),\ \Sigma_1 = 2I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	91.9%	93.8%	97.4%	97.4%
	IL	0.123	0.236	0.208	0.173
LR	DOC	94.3%	96.4%	97.2%	98.9%
	IL	0.087	0.146	0.115	0.124
SVM	DOC	90.3%	94.6%	94.7%	98.8%
	IL	0.086	0.127	0.102	0.122
NB	DOC	94.2%	96.6%	95.3%	98.8%
	IL	0.079	0.127	0.096	0.114
KNN	DOC	88.1%	91.9%	93.5%	99.5%
	IL	0.086	0.114	0.113	0.139
		Case III: $\mu_1 = (1, 1),\ \Sigma_1 = 2I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	90.1%	94.2%	97.3%	97.1%
	IL	0.140	0.273	0.236	0.193
LR	DOC	93.8%	96.2%	95.3%	99.5%
	IL	0.120	0.188	0.145	0.165
SVM	DOC	91.7%	95.2%	92.1%	99.9%
	IL	0.121	0.177	0.143	0.164
NB	DOC	94.8%	96.0%	94.0%	97.0%
	IL	0.096	0.169	0.134	0.152
KNN	DOC	88.1%	89.0%	94.5%	98.8%
	IL	0.112	0.143	0.148	0.175

The approximate symmetrical confidence interval based on the 5×2 cross-validated t distribution has acceptable degrees of confidence in half of the cases. The approximate symmetrical confidence interval based on the 10-fold cross-validated t distribution has low degrees of confidence in all cases.

The degree of confidence is not the only important consideration in choosing a statistical confidence interval. When the degree of confidence is comparative, the confidence interval is always measured based on the interval length, i.e., for a given the degree of confidence, a fundamental principle for selecting the confidence interval is to select one with the shortest interval length. Tables 4, 5 and 6 show that with an acceptable degree of confidence, our method (the non-symmetrical confidence interval based on the blocked 3×2 cross-validated Beta prime distribution) has shorter interval lengths than the approximate symmetrical confidence intervals based on the blocked 3×2 and 5×2 cross-validated t distributions in most cases.

TABLE 5
Degrees of Confidence and Interval Lengths of the Four Confidence Intervals at $n = 600$ for Simulated Data

		Case I: $\mu_1 = (0.5, 0.5),\ \Sigma_1 = I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	89.4%	94.7%	97.4%	98.0%
	IL	0.093	0.171	0.150	0.126
LR	DOC	92.2%	96.2%	95.4%	99.6%
	IL	0.075	0.139	0.102	0.117
SVM	DOC	79.6%	78.2%	72.8%	90.6%
	IL	0.089	0.132	0.108	0.121
NB	DOC	93.9%	95.8%	94.2%	98.6%
	IL	0.076	0.137	0.105	0.117
KNN	DOC	87.1%	89.6%	93.4%	98.5%
	IL	0.090	0.121	0.119	0.128
		Case II: $\mu_1 = (1.5, 1.5),\ \Sigma_1 = 2I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	92.2%	93.9%	97.9%	98.4%
	IL	0.064	0.127	0.109	0.089
LR	DOC	94.6%	96.7%	90.9%	99.7%
	IL	0.044	0.077	0.057	0.066
SVM	DOC	91.5%	95.1%	92.4%	99.3%
	IL	0.049	0.073	0.055	0.067
NB	DOC	92.5%	96.4%	90.6%	99.9%
	IL	0.045	0.071	0.052	0.061
KNN	DOC	82.7%	88.9%	92.8%	99.0%
	IL	0.049	0.064	0.066	0.078
		Case III: $\mu_1 = (1, 1),\ \Sigma_1 = 2I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	90.5%	92.9%	96.2%	95.5%
	IL	0.076	0.144	0.125	0.103
LR	DOC	93.2%	96.9%	93.4%	98.5%
	IL	0.067	0.109	0.080	0.093
SVM	DOC	87.7%	90.5%	89.8%	96.3%
	IL	0.071	0.104	0.081	0.096
NB	DOC	93.9%	96.5%	94.6%	99.4%
	IL	0.053	0.097	0.070	0.084
KNN	DOC	87.1%	89.2%	94.1%	99.6%
	IL	0.065	0.082	0.083	0.100

TABLE 6
Degrees of Confidence and Interval Lengths of the Four Confidence Intervals at $n = 6{,}000$ for Simulated Data

		Case I: $\mu_1 = (0.5, 0.5),\ \Sigma_1 = I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	87.0%	96.3%	99.4%	97.0%
	IL	0.034	0.090	0.080	0.040
LR	DOC	94.3%	97.6%	94.2%	99.8%
	IL	0.023	0.045	0.032	0.036
SVM	DOC	87.5%	95.1%	96.9%	98.9%
	IL	0.029	0.056	0.044	0.040
NB	DOC	88.6%	95.6%	92.4%	99.4%
	IL	0.024	0.044	0.031	0.037
KNN	DOC	86.8%	88.2%	95.0%	99.2%
	IL	0.028	0.037	0.037	0.041
		Case II: $\mu_1 = (1.5, 1.5),\ \Sigma_1 = 2I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	86.7%	97.0%	99.5%	98.0%
	IL	0.022	0.054	0.047	0.026
LR	DOC	92.9%	96.4%	88.6%	98.3%
	IL	0.014	0.024	0.017	0.020
SVM	DOC	91.0%	96.5%	96.9%	99.3%
	IL	0.017	0.031	0.025	0.023
NB	DOC	91.2%	96.3%	91.8%	99.6%
	IL	0.012	0.022	0.016	0.018
KNN	DOC	88.3%	88.8%	92.6%	98.2%
	IL	0.015	0.019	0.021	0.024
		Case III: $\mu_1 = (1, 1),\ \Sigma_1 = 2I_2$			
		$CI_{t(10CV)}$	$CI_{t(5\times 2CV)}$	$CI_{t(3\times 2CV)}$	$CI_{BP(3\times 2CV)}$
CT	DOC	84.2%	97.1%	98.6%	98.4%
	IL	0.031	0.082	0.070	0.032
LR	DOC	91.2%	96.1%	93.9%	99.7%
	IL	0.019	0.035	0.025	0.028
SVM	DOC	91.2%	96.8%	97.4%	99.4%
	IL	0.024	0.048	0.037	0.033
NB	DOC	82.0%	92.5%	92.5%	99.7%
	IL	0.017	0.030	0.021	0.026
KNN	DOC	87.0%	86.3%	93.7%	99.1%
	IL	0.021	0.026	0.027	0.032

If only the interval length is considered, the approximate confidence interval based on the 10-fold cross-validated t distribution is better than the three other methods. Hence, if the goal is to be confident that the shorter interval length is the better confidence interval, then the approximate confidence interval based on the 10-fold cross-validation may be the best choice, even though its degree of confidence is unacceptable.

5.2 Real Letter Recognition Data Comparison

A data set from UCI database for identifying the letters of the roman alphabet comprises 20,000 examples described by 16 features. The 26 letters represent 26 categories, similar to [3], [10], who turned it into a two-class (A-M vs. N-Z) classification problem. We sample, with replacement, 200 (600) examples from the 20,000 examples available in the Letter Recognition data. Repeating this 1,000 times, we then compute the degrees of confidence and interval lengths of the four confidence intervals based on 1,000 sets of data obtained.

Similar to the simulated data situation, the non-symmetrical confidence interval based on the blocked 3×2 cross-validated Beta prime distribution has high degree of confidence except for the case with $n = 600$ and KNN classifier, as shown in Table 8. The approximate confidence interval based on the 10-fold cross-validated t distribution and 5×2 cross-validated t distribution have degrees of confidence with less than 95 percent in most cases. Tables 7 and 8 exhibit that of the two confidence intervals based on the blocked 3×2 cross-validation with acceptable degrees of confidence, our method has the shortest interval length. Similarly, if degree of confidence is not considered, the approximate confidence interval based on the 10-fold cross-validated t distribution has the shortest interval length of the four confidence intervals.

6 Conclusions

We presented a new perspective on the confidence interval for the F_1 measure. This view, grounded on a Beta prime distribution rather than on the traditional t distribution,

TABLE 7
Degrees of Confidence and Interval Lengths of the Four Confidence Intervals at $n = 200$ for Real Data

		$CI_{t(10CV)}$	$CI_{t(5\times2CV)}$	$CI_{t(3\times2CV)}$	$CI_{BP(3\times2CV)}$
CT	DOC	89.8%	91.5%	97.9%	97.1%
	IL	0.158	0.312	0.279	0.214
LR	DOC	92.2%	91.6%	95.6%	97.9%
	IL	0.150	0.230	0.204	0.204
SVM	DOC	87.4%	94.2%	96.7%	98.9%
	IL	0.150	0.249	0.210	0.205
NB	DOC	93.7%	94.1%	97.4%	99.0%
	IL	0.153	0.241	0.212	0.206
KNN	DOC	88.9%	87.7%	93.7%	96.4%
	IL	0.140	0.213	0.198	0.200

TABLE 8
Degrees of Confidence and Interval Lengths of the Four Confidence Intervals at $n = 600$ for Real Data

		$CI_{t(10CV)}$	$CI_{t(5\times2CV)}$	$CI_{t(3\times2CV)}$	$CI_{BP(3\times2CV)}$
CT	DOC	85.9%	94.2%	97.7%	97.3%
	IL	0.085	0.168	0.150	0.115
LR	DOC	93.1%	94.4%	95.2%	98.9%
	IL	0.084	0.137	0.110	0.114
SVM	DOC	86.0%	86.7%	85.4%	96.2%
	IL	0.090	0.141	0.117	0.121
NB	DOC	94.5%	96.6%	98.0%	99.6%
	IL	0.086	0.156	0.123	0.116
KNN	DOC	81.1%	77.4%	81.3%	82.0%
	IL	0.066	0.109	0.100	0.099

enables the consideration of the construction of a new non-symmetrical confidence interval under the principle of selecting that with the shortest interval length for a given degree of confidence. Such interval may be more reasonable than the commonly used symmetrical approximate confidence intervals.

To develop this view, we further showed how the proposed confidence interval outperformed the other approximate symmetrical confidence intervals through simulated experiments. In studies on machine learning, researchers are also interested in testing the significance of the difference of algorithm performances. ([3], [5], [6], [10], [18], [19]) Thus, further study on the use of the proposed confidence interval in comparisons of algorithms is being conducted.

Appendix A
Proof of Corollary 1

To prove the unimodal property of $p(t)$, we only prove that $p(t)$ is monotonically increasing for $0 < t < mode(F_1)$ and monotonically decreasing for $mode(F_1) < t < 1$ when $a > 1$ and $b \geq 1$.

Notably, the derivative of $p(t)$ has the form

$$p'(t) = \frac{2^a(1-t)^{a-2}(2-t)^{-a-b-1}t^{b-2}}{B(a,b)} \\ [-2t^2 + (-a-2b+5)t + 2(b-1)],$$

and obviously, the $\frac{2^a(1-t)^{a-2}(2-t)^{-a-b-1}t^{b-2}}{B(a,b)} > 0$ for $0 < t < 1$. Thus we only need to consider the function $-2t^2 + (-a-2b+5)t + 2(b-1)$, which is denoted as $f(t)$. When $b \geq 1$, $(-a-2b+5)^2 - 4(-2)2(b-1) \geq 0$, and the solutions of $f(t)$ exist. Obviously, we have the solutions

$$\frac{(-a-2b+5) - \sqrt{(-a-2b+5)^2 + 16(b-1)}}{4} \\ < 0 < \frac{(-a-2b+5) + \sqrt{(-a-2b+5)^2 + 16(b-1)}}{4}.$$

If $\frac{(-a-2b+5)+\sqrt{(-a-2b+5)^2+16(b-1)}}{4} < 1$, then the proof is completed. When $a > 1$,$\frac{(-a-2b+5)+\sqrt{(-a-2b+5)^2+16(b-1)}}{4} < 1$, i.e., $p(t)$ is monotonically increasing for $0 < t < mode(F_1)$ and monotonically decreasing for $mode(F_1) < t < 1$ when $a > 1$ and $b \geq 1$.

Appendix B
Proof of Corollary 2

If $p(t)$ is symmetrical, $p(t) = p(2t_0 - t)$ for arbitrary $0 < t < t_0 = mode(F_1)$. Notably, $p(t)$ does not include a constant term. However,

$$p(2t_0 - t) = \frac{2^a(1-(2t_0-t))^{a-1}(2-(2t_0-t))^{-a-b}(2t_0-t)^{b-1}}{B(a,b)}$$

for $t_0 \neq 1/2$ includes a nonzero constant term. Obviously $p(t) = p(2t_0 - t)$ can not be obtained. Thus $p(t)$ is non-symmetrical for $t_0 \neq 1/2$. When $t_0 = 1/2$, we have $a = 2b$. Thus

$$p(t) = \frac{2^{2b}(1-t)^{b-1}t^{b-1}}{B(2b,b)}(2-t)^{-3b}(1-t)^b,$$
$$p(1-t) = \frac{2^{2b}(1-t)^{b-1}t^{b-1}}{B(2b,b)}(1+t)^{-3b}t^b.$$

Similarly, $(2-t)^{-3b}(1-t)^b$ includes a nonzero constant term, but $(1+t)^{-3b}t^b$ does not include a constant term. Thus $p(t) \neq p(1-t)$ for $t_0 = 1/2$.

Noting that

$$\frac{p(t)}{p(1-t)} = \left[\frac{1-t}{t}\frac{(1+t)^3}{(2-t)^3}\right]^b = \left[\frac{-t^4 - 2t^3 + 2t + 1}{-t^4 + 6t^3 - 12t^2 + 8t}\right]^b$$

the proof is completed if we can prove $\frac{-t^4-2t^3+2t+1}{-t^4+6t^3-12t^2+8t} \geq 1$ for arbitrary $0 < t \leq 0.5$. Let $g(t) = -t^4 - 2t^3 + 2t + 1 - (-t^4 + 6t^3 - 12t^2 + 8t) = -8t^3 + 12t^2 - 6t + 1$. We then have $g'(t) = -24t^2 + 24t - 6 = -6(2t-1)^2 \leq 0$, i.e., $g(t)$ is a monotonically decreasing function. Along with the fact that $g(0.5) = 0$, we have $g(t) \geq 0$ for $0 < t \leq 0.5$. Consequently, $\frac{-t^4-2t^3+2t+1}{-t^4+6t^3-12t^2+8t} \geq 1$.

Acknowledgments

Jihong Li is the corresponding author. This work was supported by National Natural Science Fund of China (NNSFC)(60873128). Experiments were supported by High Performance Computing System of Shanxi University.

REFERENCES

[1] S. Mao, J. Wang, and X. Pu, *Advanced Mathematical Statistics*. Beijing, China: Higher Education Press, 2006.

[2] B. P. Dominic, "Confidence Intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects," *Tuts. Quant. Methods Psychol.*, vol. 2, no. 1, pp. 11–19, 2006.

[3] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, 2003.

[4] D. Berrar and J. A. Lozano, "Significance tests or confidence intervals: Which are preferable for the comparison of classifiers?" *J. Exp. Theor. Artif. Intell.*, vol. 25, no. 2, pp. 189–206, 2013.

[5] T. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1924, 1998.

[6] E. Alpaydin, "Combined 5×2 CV F test for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 1885–1892, 1999.

[7] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, 2004.

[8] M. Markatou, H. Tian, S. Biswas, and G. Hripcsak, "Analysis of variance of cross-validation estimators of the generalization error," *J. Mach. Learn. Res.*, vol. 6, pp. 1127–1168, 2005.

[9] O. T. Yildiz, "Omnivariate rule induction using a novel pairwise statistical test," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 9, pp. 2105–2118, Sep. 2013.

[10] Y. Wang, R. Wang, H. Jia, and J. Li, "Blocked 3×2 cross-validated t-test for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 26, no. 1, pp. 208–235, 2014.

[11] P. A. Flach, "The Geometry of ROC Space: Understanding machine learning metrics through ROC isometrics," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 194–201.

[12] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, 2006.

[13] J. M. Lobo, V. A. Jimenez, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Global Ecol. Biogeography*, vol. 17, pp. 145–151, 2008.

[14] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Colloq. IR Res.*, 2005, pp. 345–359.

[15] Y. Yang and X. Liu, "A Re-examination of text categorization methods," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 42–49.

[16] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2004, pp. I-409–I-412.

[17] M. Keller, S. Bengio, and S. Y. Wong, "Benchmarking nonparametric statistical tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 651–658.

[18] O. T. Yildiz and E. Alpaydin, "Ordering and finding the best of K > 2 supervised learning algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 392–402, Mar. 2006.

[19] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

Yu Wang received the BS and MS degrees in applied mathematics and probability and mathematical statistics from Shanxi University, Taiyuan, China, in 2003 and 2006, respectively. He is currently a lecturer at the Computer Center, Shanxi University. His current research focus is on statistical machine learning and data processing.

Jihong Li received the MS and PhD degrees in statistics from the Chinese Academy of Science, Beijing, China and computer application from Shanxi University, Taiyuan, China, respectively. He joined Shanxi University, China, in 1988, where he is currently a full professor. Up to now, he has authored and coauthored more than 30 journal papers on statistical learning, natural language processing, statistics and etc. His current research interests include deep learning, statistical machine learning, and natural language processing.

Yanfang Li received the BS degree in statistics from Shanxi University, Taiyuan, China. She is currently working toward the MS degree with the School of Mathematical Sciences, Shanxi University. Her research interest is focused on statistical learning.

Ruibo Wang is currently a lecturer at the Computer Center, Shanxi University. His current research focus is on natural language processing.

Xingli Yang is currently a lecturer at the School of Mathematical Sciences, Shanxi University. Her current research focus is on statistical machine learning.