Bayes Test of Precision, Recall, and F_1 Measure for Comparison of Two Natural Language Processing Models

Source code: `https://github.com/RamboWANG/acl2019`

Ruibo Wang & Jihong Li

E-mail: {wangruibo, lijh}@sxu.edu.cn

Shanxi University, Taiyuan, China

Abstract

Direct comparison on point estimation of the precision (P), recall (R), and F_1 measure of two natural language processing (NLP) models on a common test corpus **is unreasonable and results in less replicable conclusions** due to a lack of a statistical test. However, the existing t-tests in cross-validation (CV) for model comparison are inappropriate because the distributions of P, R, F_1 are skewed and an interval estimation of P, R, and F_1 based on a t-test may exceed [0,1]. In this study, we propose to use a **block-regularized** 3×2 **CV** (3×2 **BCV)** in model comparison because it could regularize the difference in certain frequency distributions over linguistic units between training and validation sets and yield stable estimators of P, R, and F_1. On the basis of the 3×2 BCV, we calibrate the **posterior distributions of P, R, and F_1** and derive an **accurate interval estimation of P, R, and F_1**. Furthermore, **we formulate the comparison into a hypothesis testing problem and propose a novel Bayes test**. The test could directly compute the probabilities of the hypotheses on the basis of the posterior distributions and provide more informative decisions than the existing significance t-tests. Three experiments with regard to NLP chunking tasks are conducted, and the results illustrate the validity of the Bayes test.

Comparing two NLP models with P, R and F_1 on a Given Corpus

Given a corpus D_n and two NLP models $\mathcal{A}$ and $\mathcal{B}$, which model produces a higher performance system with a relatively high probability in terms of P, R and F_1?

It corresponds to a hypothesis testing problem:

$$H_0: \nu_{\mathcal{B}} - \nu_{\mathcal{A}} \leq 0 \;\; v.s. \;\; H_1: \nu_{\mathcal{B}} - \nu_{\mathcal{A}} > 0, \tag{1}$$

where $\nu_{\mathcal{A}}$ and $\nu_{\mathcal{B}}$ are the evaluation metrics of $\mathcal{A}$ and $\mathcal{B}$. In this study, P, R and F_1 are considered.

Disadvantages of Previous Model Comparison methods

1. Direct comparison on a test set with the models built based on a hold-out validation.
   - From statistical perspective, it is unscientific due to a lack of the probability $P\{\nu_{\mathcal{B}} > \nu_{\mathcal{A}}\}$ and a lack of interval estimation of performance measures of the models.
   - Many published results are less replicable.
2. A t-test based on K-fold cross-validation.
   - A sample-variance estimator in the t-test based on K-fold cross-validation is an under-estimation of true variance. Thus, the t-test often results in a false positive conclusion.
   - The distributions of P, R and F_1 are skewed. P, R and F_1 follow Beta distributions rather than Normal distributions.

Our Proposed Bayes Test Based on 3×2 BCV

1. A proposed block-regularized 3×2 cross-validation (3×2 BCV):
   - 3 repetitions of two-fold CVs with certain regularized conditions on data partitioning.
   - During data partitioning, the distribution of a training set should be consistent with that of a validation set as much as possible. Thus, 3×2 BCV regularizes empirical distributions of training and validation sets from multiple perspectives, and yields stable estimators of P, R and F_1.
2. Posterior distributions of P, R and F_1.
   - Exact Beta distributions of P, R and F_1 based on 3×2 BCV are obtained.
   - Accurate credible intervals of P, R and F_1 are proposed.
3. A Bayes test of P, R and F_1.
   - It provide how to calculate the probability of $P\{\nu_{\mathcal{B}} > \nu_{\mathcal{A}}\}$ based on 3×2 BCV.
   - The method is more reasonable then conventional null hypothesis significance testing.

Construction of 3×2 BCV

Step (a) Dividing a corpus D_n into four equal-sized blocks B_1, B_2,B_3, B_4, then taking either two blocks as a training set and the other two as a validation set to form a partition set (Table 1).

Step (b) Verifying certain frequency distributions over linguistic units, e.g. entity types in an NER task, between the training and validation sets in each two-fold CV be approximately identical.

Partitions	First fold		Second fold		Confusion matrix	
	Training	Validation	Training	Validation	First fold	Second fold
1st two-fold CV	B_1,B_2	B_3,B_4	B_3,B_4	B_1,B_2	$(\mathrm{TP}_1^{(1)},\mathrm{FP}_1^{(1)},\mathrm{FN}_1^{(1)},\mathrm{TN}_1^{(1)})$	$(\mathrm{TP}_2^{(1)},\mathrm{FP}_2^{(1)},\mathrm{FN}_2^{(1)},\mathrm{TN}_2^{(1)})$
2nd two-fold CV	B_1,B_3	B_2,B_4	B_2,B_4	B_1,B_3	$(\mathrm{TP}_1^{(2)},\mathrm{FP}_1^{(2)},\mathrm{FN}_1^{(2)},\mathrm{TN}_1^{(2)})$	$(\mathrm{TP}_2^{(2)},\mathrm{FP}_2^{(2)},\mathrm{FN}_2^{(2)},\mathrm{TN}_2^{(2)})$
3rd two-fold CV	B_2,B_3	B_1,B_4	B_1,B_4	B_2,B_3	$(\mathrm{TP}_1^{(3)},\mathrm{FP}_1^{(3)},\mathrm{FN}_1^{(3)},\mathrm{TN}_1^{(3)})$	$(\mathrm{TP}_2^{(3)},\mathrm{FP}_2^{(3)},\mathrm{FN}_2^{(3)},\mathrm{TN}_2^{(3)})$

Table 1: Partition set and confusion matrices of 3×2 BCV.

Posterior Distributions of P, R and F_1 based on 3×2 BCV

Effective confusion matrix $\mathcal{M} = (\mathbf{TP}_e, \mathbf{FP}_e, \mathbf{FN}_e, \mathbf{TN}_e)$

$$\mathrm{TP}_e = \frac{1}{1+\rho_1+4\rho_2}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)},\ \mathrm{FP}_e = \frac{1}{1+\rho_1+4\rho_2}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{FP}_k^{(j)},\ \mathrm{FN}_e = \frac{1}{1+\rho_1+4\rho_2}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{FN}_k^{(j)} \tag{2}$$

where ρ_1 , ρ_2 are intergroup, intragroup correlation coefficients in 3×2 BCV, and they satisfy that $0 < \rho_1 < 0.5$, $0.25 < \rho_2 < 0.5$ approximately.

Posterior distributions:

Precision:

$$P(p = t|\boldsymbol{\mathcal{M}}) = \frac{t^{\mathrm{TP}_e+1}(1-t)^{\mathrm{FP}_e+1}}{Beta(\mathrm{TP}_e+1, \mathrm{FP}_e+1)}, \tag{3}$$

Recall:

$$P(r = t|\boldsymbol{\mathcal{M}}) = \frac{t^{\mathrm{TP}_e+1}(1-t)^{\mathrm{FN}_e+1}}{Beta(\mathrm{TP}_e+1, \mathrm{FN}_e+1)}, \tag{4}$$

F_1 measure:

$$P(f_1 = t|\boldsymbol{\mathcal{M}}) = \frac{2^{\mathrm{FP}_e+\mathrm{FN}_e+2}(1-t)^{\mathrm{FP}_e+\mathrm{FN}_e+1}(2-t)^{-\mathrm{FP}_e-\mathrm{FN}_e-\mathrm{TP}_e-3}t^{\mathrm{TP}_e}}{Beta(\mathrm{FP}_e+\mathrm{FN}_e+2, \mathrm{TP}_e+1)}. \tag{5}$$

Credible Intervals of P, R and F_1 based on 3×2 BCV

Precision:

$$\mathrm{CI}_p = [Be_{\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FP}_e+\lambda), Be_{1-\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FP}_e+\lambda)]. \tag{6}$$

Recall:

$$\mathrm{CI}_r = [Be_{\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FN}_e+\lambda), Be_{1-\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FN}_e+\lambda)]. \tag{7}$$

F_1 measure:

$$\mathrm{CI}_{f_1} = \left[\frac{2}{2+Be'_{1-\frac{\alpha}{2}}}, \frac{2}{2+Be'_{\frac{\alpha}{2}}}\right], \tag{8}$$

Bayes Test based on 3×2 BCV for Hypothesis Testing (1)

Input: Text corpus, D_n; NLP models, $\mathcal{A}$ and $\mathcal{B}$;
Output: Probabilities $P(H_0)$ and $P(H_1)$, and a decision between “Accept H_0” and “Accept H_1”;
$\boxed{Step\ (1)}$: Construct a partition set $\mathbb{P}$ on D_n according to Table 1;
$\boxed{Step\ (2)}$: Train and validate models $\mathcal{A}$ and $\mathcal{B}$ on $\mathbb{P}$, and summarize the results as a set of confusion matrices for $\mathcal{A}$ and $\mathcal{B}$, respectively;
$\boxed{Step\ (3)}$: Apply Eq. (2) on the set of confusion matrices in Step (2) to get effective matrices $(\mathrm{TP}_{e,\mathcal{A}}, \mathrm{FN}_{e,\mathcal{A}}, \mathrm{FP}_{e,\mathcal{A}})$ and $(\mathrm{TP}_{e,\mathcal{B}}, \mathrm{FN}_{e,\mathcal{B}}, \mathrm{FP}_{e,\mathcal{B}})$;
$\boxed{Step\ (4)}$: Compute $P(\nu_{\mathcal{A}}|\boldsymbol{\mathcal{M}}_{\mathcal{A}})$ and $P(\nu_{\mathcal{B}}|\boldsymbol{\mathcal{M}}_{\mathcal{B}})$ by employing Eqs. (3), (4) and (5) on $(\mathrm{TP}_{e,\mathcal{A}}, \mathrm{FP}_{e,\mathcal{A}}, \mathrm{FN}_{e,\mathcal{A}})$ and $(\mathrm{TP}_{e,\mathcal{B}}, \mathrm{FP}_{e,\mathcal{B}}, \mathrm{FN}_{e,\mathcal{B}})$ for P, R and F_1, respectively;
$\boxed{Step\ (5)}$: Approximate $P(\nu_{\mathcal{A}} - \nu_{\mathcal{B}} \leq 0|\boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$ with 10^7 Monte-Carlo simulations.
$\boxed{Step\ (6)}$: Compute $P(H_0) \leftarrow P(\nu_{\mathcal{A}} - \nu_{\mathcal{B}} \leq 0|\boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$ and $P(H_1) \leftarrow 1 - P(\nu_{\mathcal{A}} - \nu_{\mathcal{B}} \leq 0|\boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$;
$\boxed{Step\ (7)}$: If $P(H_0) \geq P(H_1)$ **return** $(P(H_0), P(H_1)$, “Accept H_0”); else **return**$(P(H_0), P(H_1)$, “Accept H_1”);

An Illustrative Experiment

Task: Organization entity recognition task.
Data set: CoNLL 2003 English NER training set.
Model $\mathcal{A}$: CRF+IOB2 versus **Model** $\mathcal{B}$: CRF+IOBES.
Research question: Between IOB2 and IOBES, which tagging set could yield a better organization entity recognition model?
Interpretations of TP, FP and FN:
- TP indicates the count of the correctly predicted organization entities;
- FN is the count of the golden organization entities that are incorrectly predicted;
- FP is the count of the predicted organization entities that are not correct.

ν	Credible interval		Outputs of the Bayes test		
	IOB2 ($\mathcal{A}$)	IOBES ($\mathcal{B}$)	$P(H_0)$	$P(H_1)$	Decision
Precision	[91.37,92.86]	[91.85,93.31]	0.191	0.809	Accept H_1
Recall	[64.89,67.11]	[64.45,66.68]	0.706	0.294	Accept H_0
F_1 measure	[76.06,77.74]	[75.93,77.61]	0.587	0.413	Accept H_0

Table 2: Credible intervals and decisions of the Bayes test for the organization entity recognition task.

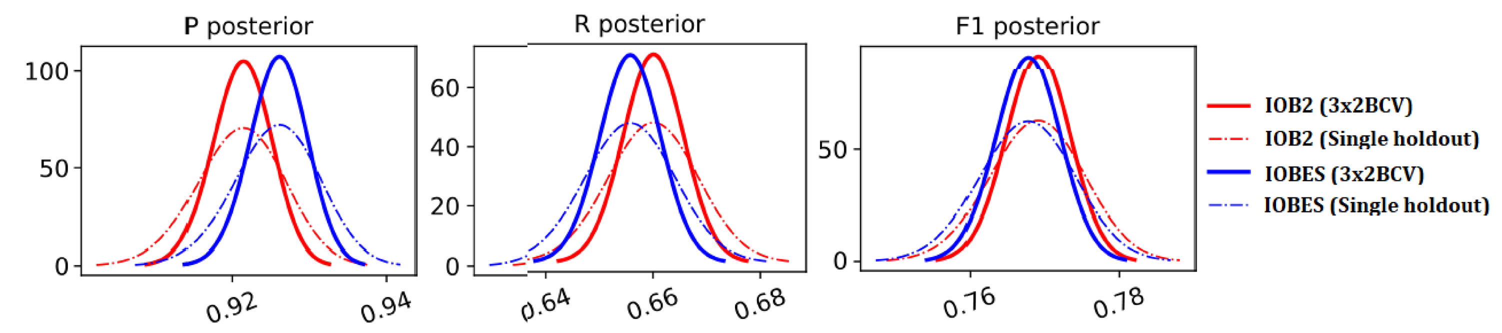

Figure 1: Posterior density curves of Precision, Recall and F_1 measure on the organization entity recognition task.

Analysis

1. Tagging set “IOBES” improves precision but deteriorates recall and F_1 measure in the organization entity recognition task.
2. Our proposed posterior distributions, which yield more accurate CIs, are taller and thinner than those in a single hold-out.
3. The results provided by the Bayes test are with more informative interpretability and help to make a reliable decision

Guidlines for NLP Practitioners

- ▶ A t-test should be avoided in a comparison of two NLP models on the basis of the precision, recall and F_1 measure.
- ▶ The 3×2 BCV could be preferred to evaluate the performance of an NLP model in the task of model comparison.
- ▶ The Bayes test on the basis of the 3×2 BCV could provide informative and fine-grained measures of the differences of precisions, recalls and F_1 measures of two NLP models, and the measures could help practitioners to make a reasonable decision.

Forthcoming Research

- ▶ Refine the Bayes test of P, R, and F_1 in an $m \times 2$ BCV with $m \geq 3$.
- ▶ Provide sequential Bayes test for model comparison.
- ▶ Verify our proposed method in several NLP tasks, such as chunking and semantic role labeling.