Bayesian Performance Comparison of Text Classifiers

Dell Zhang
DCSIS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
dell.z@ieee.org

Jun Wang
Dept of Computing Science
University College London
Gower Street
London WC1E 6BT, UK
j.wang@cs.ucl.ac.uk

Emine Yilmaz
Dept of Computing Science
University College London
Gower Street
London WC1E 6BT, UK
e.yilmaz@cs.ucl.ac.uk

Xiaoling Wang
Software Engineering Institute
East China Normal University
3663 North Zhongshan Road
Shanghai 200062, China
xlwang@sei.ecnu.edu.cn

Yuxin Zhou
Software Engineering Institute
East China Normal University
3663 North Zhongshan Road
Shanghai 200062, China
10122510216@ecnu.cn

ABSTRACT

How can we know whether one classifier is really better than the other? In the area of text classification, since the publication of Yang and Liu's seminal SIGIR-1999 paper, it has become a standard practice for researchers to apply null-hypothesis significance testing (NHST) on their experimental results in order to establish the superiority of a classifier. However, such a frequentist approach has a number of inherent deficiencies and limitations, e.g., the inability to accept the null hypothesis (that the two classifiers perform equally well), the difficulty to compare commonly-used multivariate performance measures like F_1 scores instead of accuracy, and so on. In this paper, we propose a novel Bayesian approach to the performance comparison of text classifiers, and argue its advantages over the traditional frequentist approach based on t-test etc. In contrast to the existing probabilistic model for F_1 scores which is unpaired, our proposed model takes the correlation between classifiers into account and thus achieves greater statistical power. Using several typical text classification algorithms and a benchmark dataset, we demonstrate that the our approach provides rich information about the difference between two classifiers' performances.

Keywords

Bayesian inference; hypothesis testing; performance evaluation; text classification

1. INTRODUCTION

Text classification (aka categorisation) [25] is a fundamental technique in information retrieval (IR) [19]. It has many important applications, including topic categorisation, spam filtering, sentiment analysis, message routing, language identification, genre detection, authorship attribution, and so on. In fact, most modern IR systems for search, recommendation, or advertising contain multiple components that use some form of text classification.

How can we know whether one classifier is really better than the other? Is it possible that they perform equally well? Sure we should be able to evaluate their classification performances on some benchmark datasets using some performance measures. However, given any finite amount of test results, we can never be completely certain that one classifier works better than the other or vice versa: the observed difference between their performance scores do not necessarily reflect their intrinsic qualities. The central question here is how to reliably tell if classifier A indeed outperforms classifier B, given a set of test results. Perhaps the simplest solution is to apply k-fold cross-validation [21] and then calculate the sample variance of performance scores over multiple "folds" of the dataset. This method tends to yield poor estimations though: the sample variance can approximate the true variance well only if we have a large number of folds, but when the dataset is divided into many folds, the size of each fold is likely to be too small to give a meaningful performance score (especially for complex multivariate performance measures like F_1 [26]). Hence it is desirable to derive the uncertainty of performance scores directly from all the atomic document-category classification results.

To address this problem, Yang and Liu defined in their seminal SIGIR-1999 paper [27] a suite of null-hypothesis significance testing (NHST) methods which aim to verify how strongly the experimental results support the claim that one particular classifier is more accurate than another classifier. That paper has been influential within and beyond the realm of text classification. Since its publication, it has received about 3,000 citations (according to Google Scholar). Today, it is almost compulsory for researchers to validate the superiority of their proposed text classification algorithms by means of NHST and report the p-values in their papers.

Although NHST has proven to be useful in assessing text classifiers and its adoption has greatly improved the rigour of performance evaluation in IR, such a frequentist approach has many inherent deficiencies and limitations which we shall elaborate on later. In this paper, we propose a novel approach to performance comparison of text classifiers based

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '16, July 17–21, 2016, Pisa, Italy

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2911547

on Bayesian estimation [15], and argue its advantages over the traditional frequentist approach based on t-test etc. Using a few representative text classification algorithms and a benchmark dataset, we demonstrate that the our approach provides rich information about the difference between two classifiers' performances.

2. RELATED WORK

2.1 Frequentist Performance Comparison

The traditional frequentist approach to comparing classifiers is to use NHST [21]. The usual process of NHST consists of four steps: (1) formulate the null hypothesis H_0 that the observations are the result of pure chance and the alternative hypothesis H_1 that the observations show a real effect combined with a component of chance variation; (2) identify a test statistic that can be used to assess the truth of H_0; (3) compute the p-value, which is the probability that a test statistic equal to or more extreme than the one observed would be obtained under the assumption of hypothesis H_0; (4) if the p-value is less than an acceptable significance level, the observed effect is statistically significant, i.e., H_0 is ruled out and H_1 is valid.

Specifically for performance comparison of text classifiers, the usage of NHST has been presented in detail by Yang and Liu in their SIGIR-1999 paper [27]. In summary, on the document level (micro level), sign-test can be used to compare two classifiers' accuracy scores (called s-test), while unpaired t-test can be used to compare two classifiers' performance measures in the form of proportions, e.g., precision, recall, error, and accuracy (called p-test); on the category level (macro level), sign-test and paired t-test can both be used to compare two classifiers' F_1 scores [26] (which are called S-test and T-test respectively).

In spite of being useful and influential, such a frequentist approach unfortunately has many inherent deficiencies and limitations [14, 15]. First, NHST is only able to tell us whether the experimental data are sufficient to reject the null hypothesis (that the performance difference is zero) or not, but there is no way to accept the null hypothesis. If we fail to reject the null hypothesis, we cannot conclude that it is true, but only recognise that the null hypothesis is a possibility. That is to say, it is impossible for us to use NHST to confidently claim that two classifiers perform equally well . Second, NHST will reject the null hypothesis as long as the experimental data suggest that the performance difference is non-zero, even if the performance difference is too slight to have any real effect in practice. Third, complex performance measures such as the F_1 score can only be compared on the category level but not on the document level, which seriously restricts the statistical power of NHST as the number of categories is usually much much smaller than the number of documents. Fourth, using sign-test, those pairs of identical classification outcomes are completely discarded, which is undesirable because the probability that the two classifiers are essentially equal would be substantially underestimated. Fifth, using unpaired t-test, the correlation between the classifiers in comparison are totally ignored, which is unreasonable because in reality both classifiers are likely to do well on "easy" test documents, and badly on "difficult" test documents, not to mention that those classifiers could be just different versions of the same machine learning algorithm.

The other NHST methods that have been applied to compare classifiers include ANOVA test [10], Friedman test [24], McNemar's test [6], and Wilcoxon signed ranks test [4]. Due to their frequentist nature, no matter which specific test they use, more or less they suffer from the above mentioned perils (especially the first three).

2.2 Bayesian Performance Comparison

It has been loudly advocated in recent years that the Bayesian approach to comparing two groups of data has many advantages over the frequentist NHST [14, 15]. However, to our knowledge, almost all the existing models of Bayesian performance comparison deal with continuous values (that can be described by Gaussian or t distributions) but not discrete classification outcomes, and they produce estimations for simple statistics (such as the average difference between the two given groups) but not complex performance measures (such as the F_1 score).

Probably the most closely related work is that of Goutte and Gaussier [8]. Their F_1 score model constructed using a couple of Gamma variates is not as expressive and flexible as ours. For example, generalising their model to the F_β measure ($\beta \geq 0$) [19, 26] with $\beta \neq 1$ would end up with a complex equation involving three Gamma variates, but that would be trivial in our approach. It seems that their model is restricted to a single F_1 score for binary classification with two classes only, due to its reliance upon the special properties of the Gamma distribution. In contrast, our approach is a *probabilistic graphical model* [13] which opens up many possibilities for adaptation or extension (see Section 5).

Our previous work on this topic [28, 29] has ignored any possible connection between the predictions from the two classifiers in comparison. Although this is totally fine when those two classifiers are evaluated separately each on a different test dataset, it is not the optimal solution in the common situation when those two classifiers are evaluated on exactly the same test dataset. In this paper, we extend such a simplistic "unpaired" model to the more general "paired" model which takes the correlation between classifiers into account, and demonstrate that the former has much less statistical power than the latter (see Section 4.1).

3. OUR APPROACH

3.1 Probabilistic Models

Let us consider a text classifier which has been tested on a collection of N labelled test documents, $\mathcal{D}$. For each document $\boldsymbol{x}_i$ ($i = 1, \ldots, N$), we have its true class label y_i as well as the predicted class label $\hat{y}_i$.

If this classifier is actually a Bayesian model, in principle there should be a direct way to assess the suitability of model $\mathcal{M}$ in explaining the experimental data by computing $\Pr[\mathcal{M}|\mathcal{D}] \propto \Pr[\mathcal{M}] \int_\Theta \Pr[\mathcal{D}|\Theta, \mathcal{M}] \Pr[\Theta|\mathcal{M}] \mathrm{d}\Theta$. However, here we would like to consider the general situation where the true and predicted class labels are the only information presumed to be available.

In the most basic setting, *binary classification*, a document belongs to either the positive class or the negative class. Without loss of generality, we use integer 1 as the ID of the positive class and integer 0 as the ID of the negative class. Furthermore, for the sake of clarity, we will also denote the true positive and negative classes using notations $+$ and

Table 1: The classification results from one binary classifier.

y_i		$\hat{y}_i$	
$+$	μ	1	ρ^+
		0	$1-\rho^+$
$-$	$1-\mu$	1	ρ^-
		0	$1-\rho^-$

$-$ respectively which should be regarded as interchangeable synonyms of class IDs 1 and 0.

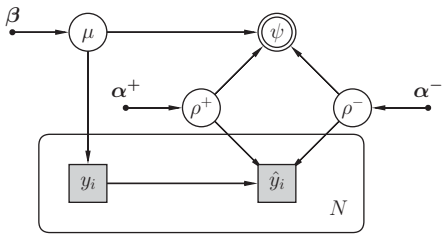
The test documents can usually be considered as "independent trials", so we regard both their true class labels y_i and their predicted class labels $\hat{y}_i$ as independent and identically distributed (i.i.d.) random variables.

Table 1 lists all the possible classification results and their corresponding probabilities for a test document using one binary classifier. It is worth noting that in our model a classifier is allowed to exhibit different prediction accuracies on documents from different true classes. This flexibility is necessary to reflect the reality and facilitate the estimation of complex performance measures that take class imbalance into account.

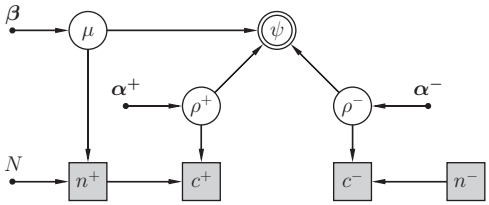
Given a test document $\boldsymbol{x}_i$, we use μ to represent the probability that its true class label y_i is positive. Obviously the probability that y_i is negative would therefore be $1-\mu$. This means that y_i follows a Bernoulli distribution with parameter μ: $y_i \sim \text{Bern}(\mu)$, i.e., $\Pr[y_i|\mu] = \mu^{y_i}(1-\mu)^{1-y_i}$. It would then be convenient to use the Beta distribution (which is conjugate to the Bernoulli distribution) as the prior distribution of parameter μ. More specifically, $\mu \sim \text{Beta}(\boldsymbol{\beta})$, i.e., $\Pr[\mu] = \frac{\Gamma(\beta^+,\beta^-)}{\Gamma(\beta^+)\Gamma(\beta^-)}\mu^{\beta^+-1}(1-\mu)^{\beta^--1}$ where the hyper-parameter $\boldsymbol{\beta} = (\beta^+, \beta^-)$ encodes our prior belief about each class's proportion. If we do not have such knowledge, we can simply set $\boldsymbol{\beta} = (1,1)$ that yields a uniform distribution, as we did in our experiments.

When a test document $\boldsymbol{x}_i$ with true class label y_i is classified, we anticipate that it will be classified as positive with a certain probability ρ^{y_i}, i.e., $\Pr[\hat{y}_i = 1|\rho^{y_i}] = \rho^{y_i}$. For example, ρ^- is the probability that a negative ($-$) document is classified to be positive (1). Hence we can say that $\hat{y}_i$ follows a Bernoulli distribution with parameter ρ^+ when y_i is positive and ρ^- when y_i is negative. In other words, $\hat{y}_i \sim \text{Bern}(\rho^+)$ if $y_i = +$ and $\hat{y}_i \sim \text{Bern}(\rho^-)$ if $y_i = -$. It would then be convenient to use the Beta distribution as the prior distribution of parameter ρ^+ and ρ^-. More specifically, $\rho^+ \sim \text{Beta}(\boldsymbol{\alpha}^+)$, i.e., $\Pr[\rho^+] = \frac{\Gamma(\alpha_1^+,\alpha_0^+)}{\Gamma(\alpha_1^+)\Gamma(\alpha_0^+)}{\rho^+}^{\alpha_1^+-1}(1-\rho^+)^{\alpha_0^+-1}$ where the hyper-parameter $\boldsymbol{\alpha}^+ = \left(\alpha_1^+, \alpha_0^+\right)$ encodes our prior belief about the classifier's prediction accuracy on positive test documents. In the same way, we have $\rho^- \sim \text{Beta}(\boldsymbol{\alpha}^-)$, where $\boldsymbol{\alpha}^- = \left(\alpha_1^-, \alpha_0^-\right)$. If we do not have any prior knowledge, we can simply set $\boldsymbol{\alpha}^+ = \boldsymbol{\alpha}^- = (1,1)$ that yields a uniform distribution, as we did in our experiments.

Once the parameters μ, ρ^+ and ρ^- have been estimated, it will be easy to calculate the contingency table of "expected" classification results: true positive (tp), false positive (fp), true negative (tn), and false negative (fn). For example, the anticipated number of true positive predictions of the classifier should be the number of positive test documents $N\mu$ times the rate of being predicted by the classifier as positive



(a) original



(b) compact

Figure 1: The probabilistic graphical model for a binary text classifier's performance.

ρ^+. The equations to calculate the contingency table for a classifier are listed as follows.

$$tp = N\mu\rho^+ \qquad fp = N(1-\mu)\rho^-$$
$$fn = N\mu(1-\rho^+) \qquad tn = N(1-\mu)(1-\rho^-)$$

With the contingency table for a classifier available, we can compute not only the accuracy, but also more complex performance measures such as the F_1 score for that classifier. The precision P, recall R, and their harmonic mean F_1 score could be computed as follows.

$$P = \frac{tp}{tp+fp} = \frac{\mu\rho^+}{\mu\rho^+ + (1-\mu)\rho^-}$$
$$R = \frac{tp}{tp+fn} = \frac{\mu\rho^+}{\mu\rho^+ + \mu(1-\rho^+)} = \rho^+$$
$$F_1 = \frac{2PR}{P+R}$$

It can be seen that N is cancelled out in the calculation of the precision, the recall, and the F_1 score.

Such a model is quite general to accommodate various performance measures (see Section 5), though in this paper we focus on the F_1 score only to illustrate the usage of our model. Let ψ denote the chosen performance measure, then it is simply a function that depends on μ, ρ^+ and ρ^- only: $\psi = f(\mu, \rho^+, \rho^-)$.

This model describes a generative mechanism of a classifier's test results. It is summarised as follows, and also depicted in Figure 1a as a *probabilistic graphical model* [13] using common notations.

$$\begin{aligned}
\mu &\sim \text{Beta}(\boldsymbol{\beta}) \\
y_i &\sim \text{Bern}(\mu) \text{ for } i = 1, \ldots, N \\
\rho^+ &\sim \text{Beta}(\boldsymbol{\alpha}^+) \qquad \rho^- \sim \text{Beta}(\boldsymbol{\alpha}^-) \\
\hat{y}_i &\sim \begin{cases} \text{Bern}(\rho^+) \text{ for } i = 1, \ldots, N & \text{if } y_i = + \\ \text{Bern}(\rho^-) \text{ for } i = 1, \ldots, N & \text{if } y_i = - \end{cases} \\
\psi &= f(\mu, \rho^+, \rho^-)
\end{aligned}$$

In the above model, each true class label y_i is regarded as an individual sampling event, and each prediction $\hat{y}_i$ is treated as an individual sampling event too. If we aggregate the occurrences of such individual sampling events into the counts of their occurrences, the model could be greatly simplified.

Let n^+ represent the total number of positive test documents and $n^- = N - n^+$ represent the total number of negative test documents, then n^+ is known to follow the Binomial distribution with parameters N and μ: $n^+ \sim \text{Bin}(N, \mu)$, i.e., $\Pr[n^+|N, \mu] = \binom{N}{n^+}\mu^{n^+}(1-\mu)^{N-n^+}$ where $\binom{N}{n^+} = \frac{N!}{n^+!(N-n^+)!}$ is the Binomial coefficient.

Let c^+ represent the count of positive predictions ($\hat{y}_i = 1$) produced on positive test documents ($y_i = +$), then c^+ is known to follow the Binomial distribution with parameters n^+ and ρ^+: $c^+ \sim \text{Bin}(n^+, \rho^+)$, i.e., $\Pr[c^+|n^+, \rho^+] = \binom{n^+}{c^+}\rho^{+c^+}(1-\rho^+)^{n^+-c^+}$. In the same way, we have $c^- \sim \text{Bin}(n^-, \rho^-)$.

The parameters μ, ρ^+ and ρ^- are the same as before and their prior distributions remain the same. The deterministic variable ψ also stays unchanged.

This compact model is equivalent to the original model, but it will be computationally much more efficient due to the drastic reduction of sampling events. So hereafter the compact model will be used instead of the original model for our work on performance comparison.

The compact model is summarised as follows and depicted in Figure 1b.

$$
\begin{aligned}
\mu &\sim \text{Beta}(\boldsymbol{\beta}) \\
n^+ &\sim \text{Bin}(N, \mu) \qquad & n^- &= N - n^+ \\
\rho^+ &\sim \text{Beta}(\boldsymbol{\alpha}^+) \qquad & \rho^- &\sim \text{Beta}(\boldsymbol{\alpha}^-) \\
c^+ &\sim \text{Bin}(n^+, \rho^+) \qquad & c^- &\sim \text{Bin}(n^-, \rho^-) \\
\psi &= f(\mu, \rho^+, \rho^-)
\end{aligned}
$$

The usage of conjugate priors (e.g., Beta for Bernoulli or Binomial) is not obligatory in our model. Actually any reasonable probability distribution can be used as the prior of μ, ρ^+ or ρ^-. If we insist on using conjugate priors, it is possible to simplify the model even further by computing the posterior probability distributions of our model parameters analytically and then sampling from the posterior probability distributions directly. However, this will only bring moderate improvement to computational efficiency, and more importantly it will make the model less flexible as some extensions to the model (such as hierarchical modelling) will be obstructed. So we shall not go down that direction in this paper.

3.1.1 Unpaired Model

In the unpaired model for performance comparison, the predictions from the two classifiers A and B being compared are assumed to be independent with each other [28, 29]. Actually the two classifiers could be evaluated each on a different test dataset as long as the data come from the same distribution (e.g., with the same proportion of positive test examples). So we can simply pool the two probabilistic models for those two classifiers together, and introduce a deterministic variable δ to capture the difference between their performance scores ψ^{A} and ψ^{B}.

$$\delta = \psi^{\text{A}} - \psi^{\text{B}}$$

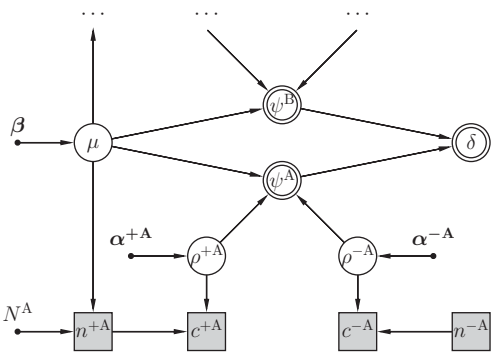


Figure 2: The unpaired model for performance comparison.

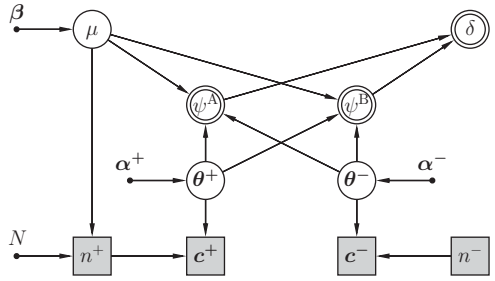


Figure 3: The paired model for performance comparison.

The unpaired model consisting of two separate sub-models for two classifiers A and B is depicted in Figure 2, where most of the sub-model for B is omitted as it is symmetric to that of A.

3.1.2 Paired Model

Although the unpaired model is simple and effective, its underlying assumption that the predictions from two classifiers A and B are independent of each other is unrealistic when those two classifiers are evaluated on the same test dataset. In contrast to the existing work for classification performance comparison (see Section 2), we would like to avoid this unrealistic assumption by modelling the two classifiers' predictions *jointly* as pairs. This is indeed crucial to assessing the real significance of the two classifiers' performance difference, as we demonstrate later in our experiments (see Section 4.1).

Considering two classifiers A and B evaluated on the same document collection, we have for each document $\boldsymbol{x}_i$ ($i = 1, \ldots, N$) a prediction outcome pair $\boldsymbol{o}_i = (\hat{y}_i^{\text{A}}, \hat{y}_i^{\text{B}})$ where $\hat{y}_i^{\text{A}}$ and $\hat{y}_i^{\text{B}}$ are the predicted class labels given by A and B respectively.

Table 2 lists all the possible classification results and their corresponding probabilities for a test document using two binary classifiers. Since for each of the two possible y_i values there are four possible $\boldsymbol{o}_i$ values $\{(1,1),(1,0),(0,1),(0,0)\}$, this table has $2 \times 4 = 8$ entries in total.

When a test document $\boldsymbol{x}_i$ with true class label y_i is classified by the two classifiers A and B, we anticipate that each possible prediction outcome pair $\boldsymbol{o}_i$ will occur with a certain probability $\theta^{y_i}_{\boldsymbol{o}_i}$, i.e., $\Pr[\boldsymbol{o}_i|\boldsymbol{\theta}^{y_i}] = \theta^{y_i}_{\boldsymbol{o}_i}$. For example, $\theta^+_{(0,1)}$ is the probability that a positive (+) document is classified to be negative (0) by the classifier A and positive (1) by

Table 2: The classification results from two binary classifiers.

y_i		$\hat{y}_i^{\mathrm{A}}$	$\hat{y}_i^{\mathrm{B}}$	$\boldsymbol{o}_i$	
$+$	μ	1	1	(1,1)	$\theta^+_{(1,1)}$
		1	0	(1,0)	$\theta^+_{(1,0)}$
		0	1	(0,1)	$\theta^+_{(0,1)}$
		0	0	(0,0)	$\theta^+_{(0,0)}$
$-$	$1-\mu$	1	1	(1,1)	$\theta^-_{(1,1)}$
		1	0	(1,0)	$\theta^-_{(1,0)}$
		0	1	(0,1)	$\theta^-_{(0,1)}$
		0	0	(0,0)	$\theta^-_{(0,0)}$

the classifier B. If we let $\boldsymbol{\theta}^+$ denote the vector of parameters $\theta^+_{\boldsymbol{o}_i}$ and similarly let $\boldsymbol{\theta}^-$ denote the vector of parameters $\theta^-_{\boldsymbol{o}_i}$, then we can say that $\boldsymbol{o}_i$ follows a Categorical distribution with parameter $\boldsymbol{\theta}^+$ when y_i is positive and $\boldsymbol{\theta}^-$ when y_i is negative. In other words, $\boldsymbol{o}_i \sim \mathrm{Cat}(\boldsymbol{\theta}^+)$ if $y_i = +$ and $\boldsymbol{o}_i \sim \mathrm{Cat}(\boldsymbol{\theta}^-)$ if $y_i = -$. It would then be convenient to use the Dirichlet distribution (which is conjugate to the Categorical distribution) as the prior distribution of parameter $\boldsymbol{\theta}^+$ or $\boldsymbol{\theta}^-$. More specifically, $\boldsymbol{\theta}^+ \sim \mathrm{Dir}(\boldsymbol{\alpha}^+)$, i.e., $\Pr[\boldsymbol{\theta}^+] = \frac{\Gamma(\sum_k \alpha_k^+)}{\prod_k \Gamma(\alpha_k^+)} \prod_k \theta_k^{\alpha_k^+ - 1}$ where the hyper-parameter $\boldsymbol{\alpha}^+ = \left(\alpha^+_{(1,1)}, \ldots, \alpha^+_{(0,0)}\right)$ encodes our prior belief about the classifier's prediction accuracy on positive test documents. In the same way, we have $\boldsymbol{\theta}^- \sim \mathrm{Dir}(\boldsymbol{\alpha}^-)$, where $\boldsymbol{\alpha}^- = \left(\alpha^-_{(1,1)}, \ldots, \alpha^-_{(0,0)}\right)$. If we do not have any prior knowledge, we can simply set $\boldsymbol{\alpha}^+ = \boldsymbol{\alpha}^- = (1, \ldots, 1)$ that yields a uniform distribution, as we did in our experiments.

Let $\boldsymbol{c}^+ = \left(c^+_{(1,1)}, \ldots, c^+_{(0,0)}\right)$ represent the counts of different types of prediction outcome pairs produced on positive test documents, then $\boldsymbol{c}^+$ is known to follow the Multinomial distribution with parameters n^+ and $\boldsymbol{\theta}^+$: $\boldsymbol{c}^+ \sim \mathrm{Mult}(n^+, \boldsymbol{\theta}^+)$, i.e., $\Pr[\boldsymbol{c}^+|N, \boldsymbol{\theta}^+] = \frac{n^+!}{c^+_{(1,1)}! \ldots c^+_{(0,0)}!} \prod_k \theta_k^{c_k^+} = \frac{\Gamma\left(\left(\sum_k c_k^+\right)+1\right)}{\prod_k \Gamma(c_k^+ + 1)} \prod_k \theta_k^{c_k^+}$. In the same way, we have $\boldsymbol{c}^- \sim \mathrm{Mult}(n^-, \boldsymbol{\theta}^-)$.

Once the parameters μ, $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ have been estimated, it will be easy to calculate, for each classifier, the contingency table of "expected" classification results as before by noticing the following facts:

$$\rho^{+\mathrm{A}} = \theta^+_{(1,1)} + \theta^+_{(1,0)} \qquad \rho^{-\mathrm{A}} = \theta^-_{(1,1)} + \theta^-_{(1,0)}$$
$$\rho^{+\mathrm{B}} = \theta^+_{(1,1)} + \theta^+_{(0,1)} \qquad \rho^{-\mathrm{B}} = \theta^-_{(1,1)} + \theta^-_{(0,1)}$$

Thus the performance scores ψ^{A} and ψ^{B}, as well as their difference δ could be estimated.

The paired model is summarised as follows and depicted in Figure 3.

$$\begin{aligned}
\mu &\sim \mathrm{Beta}(\boldsymbol{\beta}) \\
n^+ &\sim \mathrm{Bin}(N, \mu) & n^- &= N - n^+ \\
\boldsymbol{\theta}^+ &\sim \mathrm{Dir}(\boldsymbol{\alpha}^+) & \boldsymbol{\theta}^- &\sim \mathrm{Dir}(\boldsymbol{\alpha}^-) \\
\boldsymbol{c}^+ &\sim \mathrm{Mult}(n^+, \boldsymbol{\theta}^+) & \boldsymbol{c}^- &\sim \mathrm{Mult}(n^-, \boldsymbol{\theta}^-) \\
\psi^{\mathrm{A}} &= f(\mu, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-) & \psi^{\mathrm{B}} &= f'(\mu, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-) \\
\delta &= \psi^{\mathrm{A}} - \psi^{\mathrm{B}}
\end{aligned}$$

3.2 Decision Making

3.2.1 Bayes Factor

Given a probabilistic model of the chosen performance measure, we can consider the comparison of two classifiers as a *model selection* problem and utilise the Bayes factor to address it [1, 2].

In our context, the Bayes factor is the marginal likelihood of classification results data for the null model $\Pr[\mathcal{D}|\mathcal{M}_0]$ (where two classifiers perform equally well) relative to the marginal likelihood of classification results data for the alternative model $\Pr[\mathcal{D}|\mathcal{M}_1]$ (where one classifier works better than the other classifier): $\mathrm{BF} = \Pr[\mathcal{D}|\mathcal{M}_0] / \Pr[\mathcal{D}|\mathcal{M}_1]$. As the BF becomes larger, the evidence increases in favour of model $\mathcal{M}_0$ over model $\mathcal{M}_1$. The rule of thumb for interpreting the magnitude of the BF is that there is "substantial" evidence for the null model $\mathcal{M}_0$ when the BF exceeds 3, and similarly, "substantial" evidence for the alternative model $\mathcal{M}_1$ when the BF is less than $\frac{1}{3}$ [11].

Although for simple models the value of Bayes factor can be derived analytically as shown by [1, 2], for complex models it can only be computed numerically using for example the Savage-Dickey (SD) method [5]. The SD method assumes that the prior on the variance in the null model equals the prior on the variance in the alternative model at the null value: $\Pr[\sigma^2|\mathcal{M}_0] = \Pr[\sigma^2|\mathcal{M}_1, \delta = 0]$. From this it follows that the likelihood of the data in the null model equals the likelihood of the data in the alternative model at the null value: $\Pr[\mathcal{D}|\mathcal{M}_0] = \Pr[\mathcal{D}|\mathcal{M}_1, \delta = 0]$. Thus, the Bayes factor can be determined by considering the alternative hypothesis alone, because it is just the ratio of the probability density at $\delta = 0$ in the posterior relative to the probability density at $\delta = 0$ in the prior: $\mathrm{BF} = \Pr[\delta = 0|\mathcal{M}_1, \mathcal{D}] / \Pr[\delta = 0|\mathcal{M}_1]$.

3.2.2 Bayesian Estimation

Instead of relying on the Bayes factor which is a single value, we can make use of the entire posterior probability distribution of δ, the performance difference between two classifiers, for their comparison. This Bayesian (parameter) estimation approach to performance comparison is said to be more informative and more robust than using the Bayes factor [14, 15].

Given the posterior probability distribution of δ, we can then reach a discrete judgement (decision) about how those two classifiers A and B compare with each other by examining the relationship between the 95% Highest Density Interval (HDI) of δ and the user-defined Region of Practical Equivalence (ROPE) of δ [14, 15]. The 95% HDI is a useful summary of where the bulk of the most credible values of δ falls: by definition, every value inside the HDI has higher probability density than any value outside the HDI, and the total mass of points inside the 95% HDI is 95% of the distribution. The ROPE of δ, e.g., $[-0.05, +0.05]$, encloses those values of δ deemed to be negligibly different from its null value for practical purposes. Using the HDI together with the ROPE, the performance comparison decisions could be made as follows:

- if the HDI sits fully within the ROPE (as illustrated in Figure 6), A is practically equivalent ($\approx$) to B;
- if the HDI sits fully at the left or right side of the ROPE, A is *significantly* worse ($\ll$) or better ($\gg$) than B respectively;

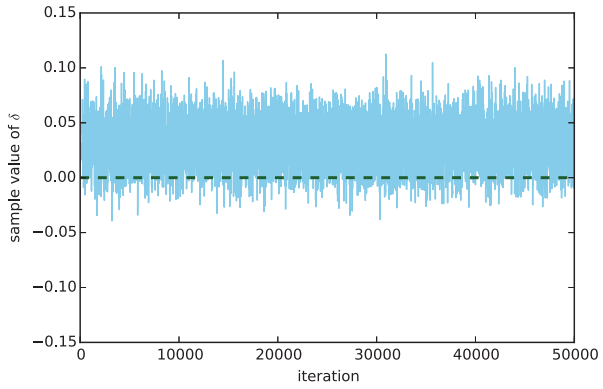


Figure 4: An example trace plot.

- if the HDI sits mainly though not fully at the left or right side of the ROPE, A is *slightly* worse ($<$) or better ($>$) than B respectively, but more experimental data would be needed to make a reliable judgement.

The need to specify the ROPE may sound like an extra burden on users compared to NHST, but in fact it is only making a hidden problem — how much performance difference would really matter for practical purposes (such as customer satisfaction and business profit) — explicit. The determination of the ROPE requires only knowledge about the application domain but not expertise in statistics. When the HDI is far away from or tightly surrounding the null value, the exact ROPE is inconsequential as any reasonable ROPE would lead to the same decision. Furthermore, in many situations, the exact ROPE can be left indeterminate. By reporting the HDI and other summary information about the full posterior distribution of δ, readers can apply whatever ROPE appropriate for them to make their own decisions.

3.3 Software Implementation

The purpose of building these models for classification results is to assess the Bayesian posterior probability of δ — the performance difference between two classifiers A and B. An approximate estimation of δ can be obtained by sampling from its posterior probability distribution via Markov Chain Monte Carlo (MCMC) [15] techniques.

We have implemented our models with an MCMC method *Metropolis-Hastings sampling* [15]. The default configuration is to generate 50,000 samples, with no "burn-in", "lag", or "multiple-chains". It has been argued in the MCMC literature that those tricks are often unnecessary: it is perfectly right to do a single long sampling run and keep all samples [13, 18]. In fact, the approximation accuracy of our program is very high: its Monte Carlo error (MC error) was usually close to 0 and never went beyond 0.002 in all our experiments (see Section 4). Figure 4 shows an example MCMC trace of our program in the experiments which clearly demonstrates the convergence of MCMC sampling.

In order to calculate the Bayes factor using the SD method (see Section 3.2.1), we approximate the posterior density $\Pr[\delta = 0|\mathcal{M}_1, \mathcal{D}]$ and the prior density $\Pr[\delta = 0|\mathcal{M}_1]$ by fitting a smooth function to the corresponding MCMC samples via kernel density estimation (KDE).

The program is written in Python 3 utilising the module `PyMC3`[1] [22] for MCMC based Bayesian model fitting. The source code is made open to the research community as online supplementary material[2]. It is free, easy to use, and extensible to more sophisticated models (see Section 5).

We should mention that this program for Bayesian performance comparison runs much slower than standard frequentist NHST techniques. On a machine with Intel x64 Core i7 CPU 2.30GHz, a sign-test or t-test would normally finish in less than 0.02 seconds, but our program could take up to 20 seconds for one comparison. Most of the time is spent on the computationally expensive MCMC sampling as it does require a decent number of samples to achieve high-fidelity approximation of probability distributions. Nevertheless, such a speed should be perfectly acceptable for the purpose of comparing classifiers because the classification experiments would usually take much longer time. Therefore the program is still very practical. Moreover, the program would be greatly accelerated if GPUs could be used by `Theano`, the underlying computational engine for `PyMC3`.

4. EXPERIMENTS

4.1 Synthetic Data

To demonstrate the advantage of our paired model over unpaired model, we perform power analysis using simulations. The statistical power is the probability of achieving the goal of a planned empirical study, if a suspected underlying state of the world is true [15]. As the power increases, there are decreasing chances of a Type II error aka the false negative rate β since the power is equal to $1 - \beta$.

We consider the following two scenarios where the two hypothetical classifiers A and B are somewhat correlated.
The scenario (a):
$\Pr[+] = \mu \quad = 0.5$
$\Pr[(1,1)|+] = \theta^{+}_{(1,1)} = 0.3$, $\Pr[(1,0)|+] = \theta^{+}_{(1,0)} = 0.3$,
$\Pr[(0,1)|+] = \theta^{+}_{(0,1)} = 0.2$, $\Pr[(0,0)|+] = \theta^{+}_{(0,0)} = 0.2$,
$\Pr[-] = 1 - \mu = 0.5$
$\Pr[(1,1)|-] = \theta^{-}_{(1,1)} = 0.2$, $\Pr[(1,0)|-] = \theta^{-}_{(1,0)} = 0.2$,
$\Pr[(0,1)|-] = \theta^{-}_{(0,1)} = 0.3$, $\Pr[(0,0)|-] = \theta^{-}_{(0,0)} = 0.3$,
It is easy to see that $F_1^{\mathrm{A}} = 0.6$ while $F_1^{\mathrm{B}} = 0.5$, so the goal here is to detect "A $\gg$ B".
The scenario (b):
$\Pr[+] = \mu \quad = 0.5$
$\Pr[(1,1)|+] = \theta^{+}_{(1,1)} = 0.3$, $\Pr[(1,0)|+] = \theta^{+}_{(1,0)} = 0.2$,
$\Pr[(0,1)|+] = \theta^{+}_{(0,1)} = 0.2$, $\Pr[(0,0)|+] = \theta^{+}_{(0,0)} = 0.3$,
$\Pr[-] = 1 - \mu = 0.5$
$\Pr[(1,1)|-] = \theta^{-}_{(1,1)} = 0.3$, $\Pr[(1,0)|-] = \theta^{-}_{(1,0)} = 0.2$,
$\Pr[(0,1)|-] = \theta^{-}_{(0,1)} = 0.2$, $\Pr[(0,0)|-] = \theta^{-}_{(0,0)} = 0.3$,
It is easy to see that $F_1^{\mathrm{A}} = 0.5$ and $F_1^{\mathrm{B}} = 0.5$, so the goal here is to detect "A $\approx$ B". Please note that this goal is infeasible using the frequentist NHST.

The power analysis results are shown in Table 3 and also Figure 5, which clearly indicate the superiority of the paired model to the unpaired model in terms of statistical power.

The reason why the unpaired model does not have as much statistical power as the paired model is because the former cannot tell whether the prediction differences (or

[1] http://pymc-devs.github.io/pymc3/
[2] http://www.dcs.bbk.ac.uk/~dell/publications/dellzhang_sigir2016_supp.html

Table 3: The power analysis of our models.

scenario	goal	dataset-size	power *unpaired*	power *paired*
(a)	A $\gg$ B	500	0.26	0.30
		1000	0.41	0.52
		1500	0.70	0.76
		2000	0.79	0.84
		2500	0.87	0.90
		3000	0.92	0.94
		3500	0.96	0.97
(b)	A $\approx$ B	500	0.00	0.00
		1000	0.01	0.22
		1500	0.26	0.58
		2000	0.63	0.81
		2500	0.72	0.87
		3000	0.88	0.96
		3500	0.92	0.99

agreements) between the two classifiers are consistent or not while the latter can. Inconsistent prediction differences yield a larger variability than consistent ones, and consequently more are required to exhibit statistical significance. Suppose that classifier A has a higher F_1 score than classifier B. If A almost always makes better predictions than B when they disagree, a relatively small amount of such consistent differences could give us enough confidence to assert statistical significance, which is recognised by the paired model but not the unpaired model.

4.2 Real-World Data

We have conducted experiments on a standard benchmark dataset for text classification, `20newsgroups` [16], of which the results are reported here. In order to ensure the reproducibility of our experimental results, we choose to use not the raw document collection, but a publicly-available ready-made "vectorised" version[3], as in [28, 29]. We have also done experiments on other "vectorised" datasets including the classic `Reuters-21578` [27], but due to the space limit those experimental results are reported only as online supplementary material together with our program's source code (see Section 3.3).

In the experiments, we have applied our proposed approach to carefully analyse the performances of two well-known supervised machine learning algorithms that are widely used for real-world text classification tasks: Naive Bayes (NB) and linear Support Vector Machine (SVM) [19]. For the former, we consider its two common variations: one with the Bernoulli event model (NB_{Bern}) and the other with the Multinomial event model (NB_{Mult}) [20]. For the latter, we consider its two common variations: one with the $L1$ norm penalty (SVM_{L1}) and the other with the $L2$ norm penalty (SVM_{L2}) [7, 30]. Thus we have four different classifiers in total. Obviously, the classification results of NB_{Bern} and NB_{Mult} would be highly correlated, and those of SVM_{L1} and SVM_{L2} as well. Among them, SVM_{L2} is widely regarded as the state-of-the-art text classifier [17, 25, 27]. It is also worth to notice that the NB algorithms will be applied not to the raw bag-of-words text datasets as people usually do, but on the vectorised `20newsgroups` dataset which has already been transformed by TF-IDF term weighting and document length normalisation.

We have used the off-the-shelf implementation of these classification algorithms provided by a Python machine learning library `scikit-learn`[4] in our experiments, again for the reproducibility reasons. The smoothing parameter α for the NB algorithm and the regularisation parameter C for the linear SVM algorithm have been tuned via grid search with 5-fold cross-validation on the training data for the macro-averaged F_1 score. The optimal parameters found are: NB_{Bern} with $\alpha = 10^{-14}$, NB_{Mult} with $\alpha = 10^{-3}$, SVM_{L1} with $C = 2^2$, SVM_{L2} with $C = 2^1$.

Table 4 shows the results of performance comparison between NB_{Bern} and NB_{Mult}, based on which we can confidently say that for most of the target categories, NB_{Bern} is outperformed by NB_{Mult}. Such results confirm the finding of [20] on this harder dataset.

Table 5 shows the results of performance comparison between SVM_{L1} and SVM_{L2}, based on which we can confidently say that for most of the target categories, SVM_{L1} and SVM_{L2} have no practical difference on classification effectiveness as measured by the F_1 score (given the ROPE $[-0.05, +0.05]$), though the former may have its advantages in terms of sparsity. Such results are complementary to those reported in [30].

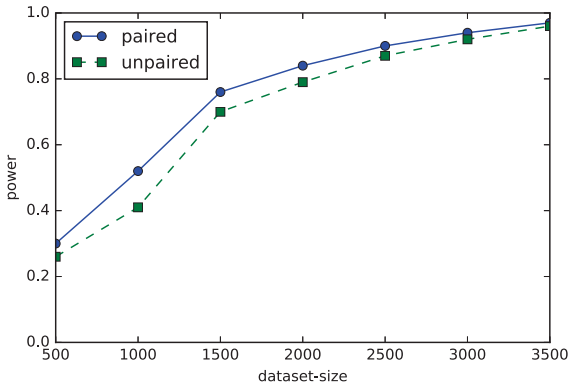
Table 6 shows the results of performance comparison between NB_{Mult} and SVM_{L2} — the better performing classifiers from the NB and SVM camps. It can be clearly seen that for most of the target categories, the competition between NB_{Mult} and SVM_{L2} is too close to call: more test data would be needed to make a reliable judgement which one works better. Nevertheless, for six out of the eight target categories on which we can indeed make reliable judgements, NB_{Mult} and SVM_{L2} are practically equivalent (given the ROPE $[-0.05, +0.05]$). This phenomenon somewhat supports the claim of [23] that NB_{Mult}, if properly enhanced by TF-IDF term weighting and document length normalisation, can reach a comparable performance as SVM_{L2}.

On the micro (document) level, no NHST method exists for the comparison of F_1 scores. So in the above tables we show the results of using NHST to compare classification accuracies instead: the column "sign-test" and "t-test" contain the two-sided p-values of micro level sign-test (called s-test in [27]) and unpaired t-test (called p-test in [27]) respectively. The symbol $\star$ indicates that the accuracy difference between A and B is statistically significant ($p < 0.05$) according to NHST. When NHST fails to reject the null hypothesis that the two classifiers work equally well, no conclusion can be drawn from the comparison.

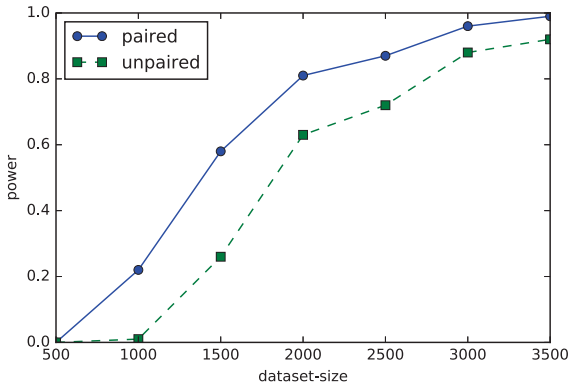
In all those tables, our proposed Bayesian performance comparison method has offered rich information about the difference between two classifiers' F_1 scores: in addition to the final judgement ("decision"), we have shown the posterior "mean", standard deviation ("std"), the Bayes factor estimated by the SD method ("BF_{SD}"), the percentage lower or greater than the null value 0 ("LG pct"), the percentage covered by the ROPE ("ROPE pct"), and the 95% "HDI". By contrast, the frequentist NHST would lead to a far less complete picture: it has only the p-values (and maybe also the confidence intervals) to offer. Furthermore, note that the judgements made by the Bayesian estimation on sev-

[3] http://scikit-learn.org/stable/datasets/twenty_newsgroups.html

[4] http://scikit-learn.org/stable/



(a) example scenario: A $\gg$ B.



(b) example scenario: A $\approx$ B.

Figure 5: Comparing the statistical power of paired and unpaired models.

Table 4: The results of performance comparison between NB_{Bern} and NB_{Mult}.

category	frequentist		Bayesian						
	sign-test	t-test	mean	std	BF_{SD}	LG pct	ROPE pct	HDI	**decision**
0	$\star$ 0.000	$\star$ 0.008	−0.081	0.021	$\star$ 0.003	100.0%<0<0.0%	6.6%	[−0.125, −0.041]	$<$
1	$\star$ 0.000	$\star$ 0.000	−0.114	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.148, −0.080]	$\lll$
2	$\star$ 0.000	$\star$ 0.000	−0.400	0.028	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.456, −0.345]	$\lll$
3	$\star$ 0.000	$\star$ 0.003	−0.095	0.016	$\star$ 0.000	100.0%<0<0.0%	0.2%	[−0.126, −0.062]	$\lll$
4	$\star$ 0.000	$\star$ 0.000	−0.249	0.019	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.286, −0.211]	$\lll$
5	$\star$ 0.000	$\star$ 0.000	−0.101	0.017	$\star$ 0.000	100.0%<0<0.0%	0.1%	[−0.135, −0.069]	$\lll$
6	$\star$ 0.000	$\star$ 0.000	−0.136	0.016	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.168, −0.105]	$\lll$
7	$\star$ 0.000	$\star$ 0.001	−0.092	0.016	$\star$ 0.000	100.0%<0<0.0%	0.4%	[−0.123, −0.061]	$\lll$
8	$\star$ 0.000	$\star$ 0.000	−0.144	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.178, −0.111]	$\lll$
9	$\star$ 0.000	$\star$ 0.001	−0.080	0.013	$\star$ 0.000	100.0%<0<0.0%	0.8%	[−0.105, −0.055]	$\lll$
10	$\star$ 0.000	$\star$ 0.000	+0.108	0.017	$\star$ 0.000	0.0%<0<100.0%	0.0%	[+0.074, +0.142]	$\ggg$
11	$\star$ 0.000	$\star$ 0.002	−0.092	0.016	$\star$ 0.000	100.0%<0<0.0%	0.4%	[−0.125, −0.061]	$\lll$
12	$\star$ 0.000	$\star$ 0.002	−0.097	0.020	$\star$ 0.000	100.0%<0<0.0%	0.8%	[−0.137, −0.058]	$\lll$
13	$\star$ 0.000	$\star$ 0.000	−0.115	0.016	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.147, −0.084]	$\lll$
14	$\star$ 0.000	$\star$ 0.000	−0.109	0.016	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.139, −0.079]	$\lll$
15	0.064	0.178	−0.027	0.016	♯ 3.132	95.7%<0<4.3%	92.2%	[−0.059, +0.004]	$<$
16	$\star$ 0.024	0.151	−0.105	0.019	$\star$ 0.000	100.0%<0<0.0%	0.1%	[−0.141, −0.068]	$\lll$
17	$\star$ 0.000	$\star$ 0.000	−0.102	0.016	$\star$ 0.000	100.0%<0<0.0%	0.1%	[−0.134, −0.070]	$\lll$
18	$\star$ 0.000	$\star$ 0.034	−0.088	0.020	$\star$ 0.007	100.0%<0<0.0%	2.9%	[−0.128, −0.049]	$<$
19	$\star$ 0.000	$\star$ 0.011	−0.040	0.030	2.389	91.3%<0<8.7%	62.9%	[−0.097, +0.022]	$<$

eral cases are different from those made by the frequentist NHST (e.g., at the significance level 0.05). So even if in some researchers' opinion the superiority of the former over the latter is still debatable, there is no doubt that the former can at least be complementary to the latter.

Figure 6 illustrates the visualisation of Bayesian performance comparison results produced by our program: the "posterior plot" sub-graph shows the posterior probability distribution of the performance difference variable δ; and the "factor plot" sub-graph shows the estimation of the Bayes factor by the SD method.

5. EXTENSIONS

The proposed Bayesian approach to performance comparison has been described above in the most basic setting for concreteness and simplicity, but it is in fact readily extensible to the following more general scenarios.

Multiple classes. It would be straightforward to extend our model to multi-class classification (either single-label or multi-label): we will need one μ parameter and a pair of $\boldsymbol{\theta}$ parameters for each class. Thus we are able to measure each classifier's overall performance using micro-averaged or macro-averaged F_1 scores [27], and compute their difference as the deterministic variable δ in the model [28]. Note that here δ is estimated using a large number of prediction outcomes for all test documents, rather than just a small number of F_1 scores for test categories as in [27] (see Section 2.1). It would be promising to go further to develop a Bayesian hierarchical model [15] where the classifier's parameters $\boldsymbol{\theta}_j$ for different classes are governed by a higher-level overarching hyper-parameter η (e.g., representing the overall probability of making correct predictions) and thus able to "share statistical strength" [29]. A potential problem, though, is the explosive growth of possible prediction outcome combinations along with the increase of class numbers, which in the worst situation may force us into backing off to the assumption of independence between classifiers so as to keep the model computationally tractable.

Other performance measures. To compare classifiers

Table 5: The results of performance comparison between SVM_{L1} and SVM_{L2}.

category	frequentist		Bayesian						
	sign-test	t-test	mean	std	BF_{SD}	LG pct	ROPE pct	HDI	**decision**
0	$\star$ 0.049	0.299	−0.027	0.018	♯ 3.399	93.6%<0<6.4%	89.9%	[−0.063, +0.008]	$\prec$
1	0.523	0.706	−0.011	0.013	♯ 9.427	80.6%<0<19.4%	99.9%	[−0.038, +0.014]	$\approx$
2	0.632	0.774	−0.007	0.014	♯ 12.058	68.3%<0<31.7%	99.8%	[−0.035, +0.020]	$\approx$
3	0.270	0.542	−0.020	0.014	♯ 4.593	92.3%<0<7.7%	98.5%	[−0.047, +0.007]	$\approx$
4	0.247	0.524	−0.022	0.014	♯ 3.781	94.8%<0<5.2%	97.4%	[−0.049, +0.005]	$\approx$
5	1.000	0.960	−0.009	0.014	♯ 12.206	73.2%<0<26.8%	99.8%	[−0.035, +0.019]	$\approx$
6	$\star$ 0.000	$\star$ 0.031	−0.053	0.013	$\star$ 0.011	100.0%<0<0.0%	41.9%	[−0.078, −0.029]	$\prec$
7	$\star$ 0.000	$\star$ 0.000	−0.133	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.166, −0.098]	$\lll$
8	0.221	0.457	−0.015	0.014	♯ 8.165	84.4%<0<15.6%	99.3%	[−0.042, +0.014]	$\approx$
9	$\star$ 0.000	$\star$ 0.000	+0.126	0.017	$\star$ 0.000	0.0%<0<100.0%	0.0%	[+0.094, +0.160]	$\ggg$
10	0.515	0.657	−0.013	0.011	♯ 9.261	87.7%<0<12.3%	99.9%	[−0.035, +0.009]	$\approx$
11	0.771	0.837	−0.005	0.013	♯ 12.937	64.0%<0<36.0%	100.0%	[−0.030, +0.021]	$\approx$
12	0.061	0.298	−0.033	0.016	1.209	98.2%<0<1.8%	84.7%	[−0.066, −0.003]	$\prec$
13	0.264	0.463	−0.019	0.015	♯ 5.513	90.2%<0<9.8%	97.9%	[−0.050, +0.009]	$\approx$
14	0.733	0.814	−0.011	0.014	♯ 10.671	79.3%<0<20.7%	99.7%	[−0.036, +0.018]	$\approx$
15	0.192	0.448	−0.035	0.013	$\star$ 0.330	99.6%<0<0.4%	86.9%	[−0.062, −0.011]	$\prec$
16	0.065	0.355	−0.030	0.014	1.225	98.7%<0<1.3%	92.1%	[−0.057, −0.003]	$\prec$
17	0.105	0.333	+0.014	0.014	♯ 6.589	15.7%<0<84.3%	99.3%	[−0.015, +0.041]	$\approx$
18	$\star$ 0.014	0.192	−0.014	0.016	♯ 8.065	82.0%<0<18.0%	98.7%	[−0.045, +0.017]	$\approx$
19	$\star$ 0.033	0.265	+0.008	0.022	♯ 7.488	36.5%<0<63.5%	96.9%	[−0.035, +0.051]	$\succ$

Table 6: The results of performance comparison between NB_{Mult} and SVM_{L2}.

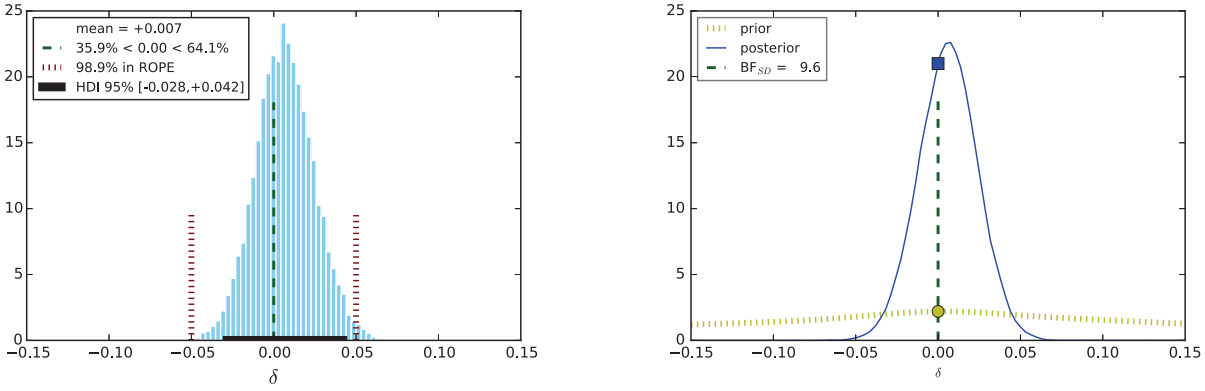
category	frequentist		Bayesian						
	sign-test	t-test	mean	std	BF_{SD}	LG pct	ROPE pct	HDI	**decision**
0	0.314	0.536	−0.001	0.022	♯ 8.542	50.5%<0<49.5%	97.7%	[−0.043, +0.042]	$\approx$
1	0.386	0.521	+0.028	0.018	♯ 3.128	6.2%<0<93.8%	89.0%	[−0.007, +0.063]	$\succ$
2	0.056	0.205	−0.028	0.021	♯ 3.739	91.2%<0<8.8%	84.5%	[−0.069, +0.013]	$\prec$
3	1.000	1.000	+0.030	0.018	2.966	5.1%<0<94.9%	86.8%	[−0.006, +0.066]	$\succ$
4	0.840	0.853	+0.006	0.018	♯ 9.533	37.6%<0<62.4%	99.0%	[−0.031, +0.040]	$\approx$
5	$\star$ 0.017	0.089	+0.051	0.017	$\star$ 0.075	0.2%<0<99.8%	48.0%	[+0.017, +0.083]	$\succ$
6	0.057	0.188	+0.013	0.016	♯ 8.805	20.6%<0<79.4%	99.1%	[−0.017, +0.046]	$\approx$
7	0.180	0.312	+0.031	0.018	2.566	4.4%<0<95.6%	85.1%	[−0.005, +0.067]	$\succ$
8	1.000	1.000	+0.007	0.018	♯ 9.551	35.9%<0<64.1%	98.9%	[−0.028, +0.042]	$\approx$
9	$\star$ 0.000	$\star$ 0.000	+0.215	0.019	$\star$ 0.000	0.0%<0<100.0%	0.0%	[+0.180, +0.253]	$\ggg$
10	$\star$ 0.000	$\star$ 0.000	−0.115	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.149, −0.082]	$\lll$
11	$\star$ 0.024	0.102	−0.016	0.017	♯ 7.016	84.0%<0<16.0%	98.1%	[−0.048, +0.017]	$\approx$
12	$\star$ 0.000	$\star$ 0.001	+0.073	0.020	$\star$ 0.008	0.0%<0<100.0%	12.4%	[+0.034, +0.113]	$\succ$
13	$\star$ 0.000	$\star$ 0.009	+0.060	0.016	$\star$ 0.004	0.0%<0<100.0%	26.1%	[+0.029, +0.090]	$\succ$
14	$\star$ 0.038	0.129	+0.050	0.017	$\star$ 0.194	0.2%<0<99.8%	50.7%	[+0.016, +0.082]	$\succ$
15	$\star$ 0.009	0.064	−0.009	0.014	♯ 10.473	74.6%<0<25.4%	99.8%	[−0.037, +0.021]	$\approx$
16	0.213	0.430	+0.041	0.017	0.650	0.8%<0<99.2%	71.1%	[+0.007, +0.074]	$\succ$
17	$\star$ 0.010	0.070	+0.053	0.016	$\star$ 0.058	0.0%<0<100.0%	43.2%	[+0.023, +0.085]	$\succ$
18	0.119	0.349	+0.015	0.019	♯ 6.927	22.2%<0<77.8%	96.5%	[−0.023, +0.053]	$\succ$
19	$\star$ 0.021	0.179	−0.061	0.031	0.825	97.6%<0<2.4%	36.0%	[−0.119, +0.001]	$\prec$

using a performance measure different from the F_1 score, we would only need to replace the function $f(\mu, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ for computing ψ, as long as that performance measure could be calculated based on the classification contingency table alone [12]. For example, it would be straightforward to extend our model to handle the more general F_β measure ($\beta \geq 0$) [19,26] with $\beta \neq 1$: we just need to substitute the F_β formula for the F_1 formula in the function of ψ. For another example, the Area Under the ROC Curve (AUC) is essentially the proportion of correctly ranked document pairs [9, 12], so it could be modelled in a similar way.

Other tasks. More generally, the idea of building a Bayesian probabilistic graphical model to make comprehensive performance comparison could be applied to not just classifiers, but also search systems (see the ICTIR-2015 best paper [3]), recommender systems, and advertising systems.

6. CONCLUSIONS

This paper tries to address the problem of comparing text classifiers' performances by appealing to Bayesian reasoning. Although we ourselves believe that Bayesian statistics is "the way it should be", we understand that not everyone is a Bayesian or wants to become a Bayesian. Our argument is not whether being a Bayesian is philosophically better than being a frequentist, but that our Bayesian estimation based approach to performance comparison of text classifiers avoids all the aforementioned practical weaknesses of NHST (see Section 2.1) and it provides much richer information about the difference between two classifiers' performances than NHST does, therefore it can supersede or at least complement the currently popular frequentist approach.



(a) posterior plot

(b) factor plot

Figure 6: A ≈ B — NB$_{Mult}$ is practically equivalent to SVM$_{L2}$ for target category 8.

7. ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their very helpful comments. This work was partly supported by the NSFC grants (61472141 and 61321064) as well as the Shanghai Knowledge Service Platform Project (ZF1213).

8. REFERENCES

[1] D. Barber. Are two classifiers performing equally? A treatment using Bayesian hypothesis testing. Technical report, IDIAP, 2004.

[2] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[3] B. Carterette. Bayesian inference for information retrieval evaluation. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval (ICTIR)*, pages 31–40, Northampton, MA, USA, 2015.

[4] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research (JMLR)*, 7:1–30, 2006.

[5] J. M. Dickey and B. P. Lientz. The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics*, 41(1):214–226, 1970.

[6] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. In *Proceedings of the 27th European Conference on IR Research (ECIR)*, pages 345–359, Santiago de Compostela, Spain, 2005.

[9] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[10] D. A. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 282–291, Dublin, Ireland, 1994.

[11] H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 2000.

[12] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 377–384, Bonn, Germany, 2005.

[13] D. Koller and N. Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[14] J. K. Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

[15] J. K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2nd edition, 2014.

[16] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pages 331–339, Tahoe City, CA, USA, 1995.

[17] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research (JMLR)*, 5:361–397, 2004.

[18] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[19] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[20] A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, Madison, WI, 1998.

[21] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[22] A. Patil, D. Huard, and C. J. Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4):1–81, 2010.

[23] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of Naive Bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 616–623, Washington, DC, USA, 2003.

[24] H. Schutze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 229–237, Seattle, WA, USA, 1995.

[25] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[26] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 2nd edition, 1979.

[27] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 42–49, Berkeley, CA, USA, 1999.

[28] D. Zhang, J. Wang, and X. Zhao. Estimating the uncertainty of average F1 scores. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval (ICTIR)*, pages 317–320, Northampton, MA, USA, 2015.

[29] D. Zhang, J. Wang, X. Zhao, and X. Wang. A bayesian hierarchical model for comparing average F1 scores. In *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM)*, pages 589–598, Atlantic City, NJ, USA, 2015.

[30] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In *Advances in Neural Information Processing Systems (NIPS) 16*, volume 16, pages 49–56, Vancouver and Whistler, Canada, 2003.