A Bayesian Hierarchical Model for Comparing Average F1 Scores

Dell Zhang\textsuperscript{1}, Jun Wang\textsuperscript{2}, Xiaoxue Zhao\textsuperscript{2}, \textbf{Xiaoling Wang}\textsuperscript{3}

\textsuperscript{1}Birkbeck, University of London, UK

\textsuperscript{2}University College London, UK

\textsuperscript{3}East China Normal University, China

17 Nov 2015
Outline

1. Background
   - Introduction
   - Problem Statement
   - Related Work

2. Our Approach
   - Models
   - Experiments

3. Summary
Outline

1. **Background**
   - Introduction
   - Problem Statement
   - Related Work

2. **Our Approach**
   - Models
   - Experiments

3. **Summary**
Introduction - Text Classification

Definition:
- Automatic text classification is a fundamental technique in information retrieval

Applications:
- Topic categorisation, spam filtering, sentiment analysis, message routing...

Performance measure:
- $F_1$ Score
Introduction - $F_1$ Score

- **Definition:**
  - The harmonic mean of precision (P) and recall (R).

- **Two methods:**
  - **Micro-averaged $F_1$ score** ($\text{MiF}_1$):
    - Gives equal weight to each classification decision
  - **Macro-averaged $F_1$ score** ($\text{MaF}_1$):
    - Gives equal weight to each class

- **Limitations:**
  - Does not tell us how reliable it is on unseen data.
Outline

1. Background
   - Introduction
   - Problem Statement
   - Related Work

2. Our Approach
   - Models
   - Experiments

3. Summary
Goal:
- Assess the uncertainty of a classifier’s performance as measured by $\text{miF}_1$ and $\text{maF}_1$
Related Work - Frequentist Performance Comparison

- NHST
  - 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)
  - use $s$-test to compare two classifiers’ accuracy scores
  - use $t$-test to compare two classifiers’ performance measures in the form of proportions
Deficiencies of NHST

- Can only reject the null hypothesis, can never accept the null hypothesis.
- Will reject the null hypothesis even the performance difference is very close to zero.
- Can only be compared on the category level but not on the document level for complex performance measures.
Related Work - Bayes Factor

1. **Bayes Factor**

2. **Deficiencies of Bayes Factor**
   - Sensitive to the choice of prior distribution in the alternative model.
   - The null hypothesis can be strongly preferred even with very few data and very large uncertainty in the estimate of the performance difference.
Bayesian Estimation

1. 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)*,

2. It is restricted to a single $F_1$ score for binary classification with two classes only.

3. In contrast, our proposed approach opens up many possibilities for adaptation or extension.
Outline

1. Background
   - Introduction
   - Problem Statement
   - Related Work

2. Our Approach
   - Models
   - Experiments

3. Summary
Multi-class single-label classification
- $M$ different classes
- $N$ labelled test documents
Documents’ true class labels $y_i$ are i.i.d.
- $\mu = (\mu_1, \ldots, \mu_M)$: the probabilities that a test document truly belongs to each class
- $n = (n_1, \ldots, n_M)$: the true size of each class
- $n$ follows a multinomial distribution with parameter $\mu$, where $\sum_{j=1}^{M} n_j = N$. 
Figure: The probabilistic graphical model for estimating the uncertainty of average $F_1$ scores.
Models - Predicted Classification

- **Class level**
  - $\theta_j = (\theta_1, ..., \theta_M)$: the probabilities that a document of true class label $j$ is classified into different classes.
  - $\omega_j = (\omega_1, ..., \omega_M)$: the parameters of the $\theta_j$’s Dirichlet prior.

- **Model level**
  - $\eta$: the overall tendency of making correct predictions
  - $w_{jk} = \begin{cases} 
  \eta & \text{if } k = j \\
  (1 - \eta)/(M - 1) & \text{if } k \neq j \text{ for } k = 1, ..., M
  \end{cases}$
Figure: The probabilistic graphical model for estimating the uncertainty of average $F_1$ scores.
Confusion matrix $C$ presents the classification results.

- $C$ is a $M \times M$ matrix.
- $c_{jk}$ represents the number of documents with true class label $j$ but predicted class label $k$.
- $c_j$ follows a multinomial distribution with parameter $\theta_j$, where $\sum_{k=1}^{M} c_{jk} = n_j$. 

\[
\begin{align*}
\beta &\quad \mu \quad \psi \\
N &\quad n \quad c_j \quad \theta_j \quad \omega_j \quad M \quad \eta \\
\alpha &
\end{align*}
\]
μ presents the true classification of documents.

ω presents the predicted classification.

Treat the performance measure (either miF₁ or maF₁) as a random variable ψ, which is a function of μ and ω. For example, in miF₁

Precision = \frac{\sum_{j=1}^{M} tp_j}{\sum_{j=1}^{M} tp_j + fp_j} = \sum_{j=1}^{M} \mu_j \theta_{jj}

Recall = \frac{\sum_{j=1}^{M} tp_j}{\sum_{j=1}^{M} tp_j + fn_j} = \sum_{j=1}^{M} \mu_j \theta_{jj}.

In multi-class single-label, miF₁ = Precision = Recall.
For two models A and B, the difference of the overall performance is represented by $\delta$, where $\delta = \psi_A - \psi_B$.

Estimate the uncertainty difference of two models by examining the posterior probability distribution of $\delta$. 
Outline

1. Background
   - Introduction
   - Problem Statement
   - Related Work

2. Our Approach
   - Models
   - Experiments

3. Summary
A standard benchmark dataset for text classification, 20newsgroups\(^1\).
- 60% subset for training
- 40% subset for testing
- Filtered by stripping newsgroup-related metadata

\(^1\)http://qwone.com/~jason/20Newsgroups/
Experiments - Classifiers

Classification algorithms:
- Naive Bayse (NB)
  - Bernoulli event model (NB\textsubscript{Bern})
  - Multinomial event model (NB\textsubscript{Mult})
- linear Support Vector Machine (SVM)
  - $L1$ penalty (SVM\textsubscript{$L1$})
  - $L2$ penalty (SVM\textsubscript{$L2$})

Implementation of these algorithms:
- Python library scikit-learn
Comparing $maF_1$ between $NB_{Bern}$ and $NB_{Mult}$.

Conclusion:
$NB_{Bern}$ is significantly outperformed by $NB_{Mult}$.
Comparing maF1 between SVM_L1 and SVM_L2.

Conclusion:
SVM_L1 is only slightly outperformed by SVM_L2

8
7
6
5
4
3
2
1
0
mean = -0.016
98.0% < 0 < 2.0%
7.3% in ROPE
HDI 95% [-0.031, -0.001]
Comparing $\text{maF}_1$ between $\text{NB}_{\text{Mult}}$ and $\text{SVM}_{L2}$.

Conclusion: $\text{NB}_{\text{Mult}}$ works a lot better than $\text{SVM}_{L2}$.
The main contribution of this paper is a Bayesian estimation approach to assessing the uncertainty of average \( F_1 \) scores in multi-class text classification. We make *interval estimation* instead of simplistic *point estimation* of a text classifier’s future performance on unseen data.

**Extension**
- To be used in the multi-class multi-label classification.
- To compare classifiers on any type of data, e.g., images.