A Bayesian Hierarchical Model for Comparing Average F1 Scores

Dell Zhang[1], Jun Wang[2], Xiaoxue Zhao[2], **Xiaoling Wang**[3]

[1]Birkbeck, University of London, UK

[2]University College London, UK

[3]East China Normal University, China

17 Nov 2015

Outline

Outline

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-averagedF_1 score (MiF_1):
    Gives equal weight to each classification decision
  - Macro-averaged F_1 score (MaF_1):
    Gives equal weight to each class
- Limitations:
  - Does not tell us how reliable it is on unseen data.

Outline

Problem Statement

- Goal:
  - Assess the uncertainty of a classifier's performance as measured by miF_1 and maF_1

Outline

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' performancce measures in the form of proportions

Related Work - Frequentist Performance Comparison

- 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
   - D. Barber, "Are two classifiers performing equally? a treatment using Bayesian hypothesis testing," IDIAP, Tech. Rep.,2004.
   - ——, Bayesian Reasoning and Machine Learning. Cambridge University Press, 2012.
2. Deficiencies of Beyes 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

Related Work - Bayesian Estimation

1. Bayesian Estimation
   - 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

Models - True Classiication

- Multi-class single-label classification
  - M different classes
  - N labelled test documents
- Documents' true class labels y_i are i.i.d.
  - $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$: the probabilities that a test document truly belongs to each class
  - $\boldsymbol{n} = (n_1, \ldots, n_M)$: the true size of each class
- n follows a multinomial distribution with parameter μ, where $\sum_{j=1}^{M} n_j = N$.

Models - True Classiication

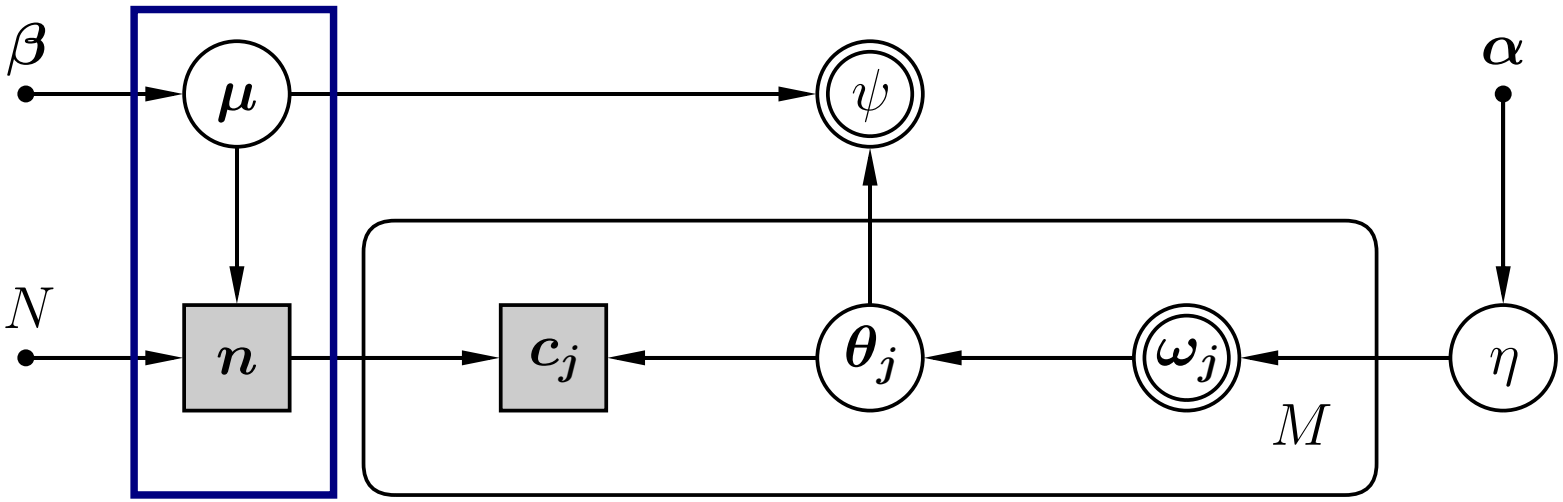


Figure: The probabilistic graphical model for estimating the uncertainty of average F_1 scores.

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 θ_j's Dirichlet prior.
- Model level
  - η: the overall tendency of making correct predictions
- $w_{jk} = \begin{cases} \eta & \text{if } k = j \\ (1-\eta)/(M-1) & \text{if } k \neq j \text{ for } k = 1, ..., M \end{cases}$

Models - Predicted Classification

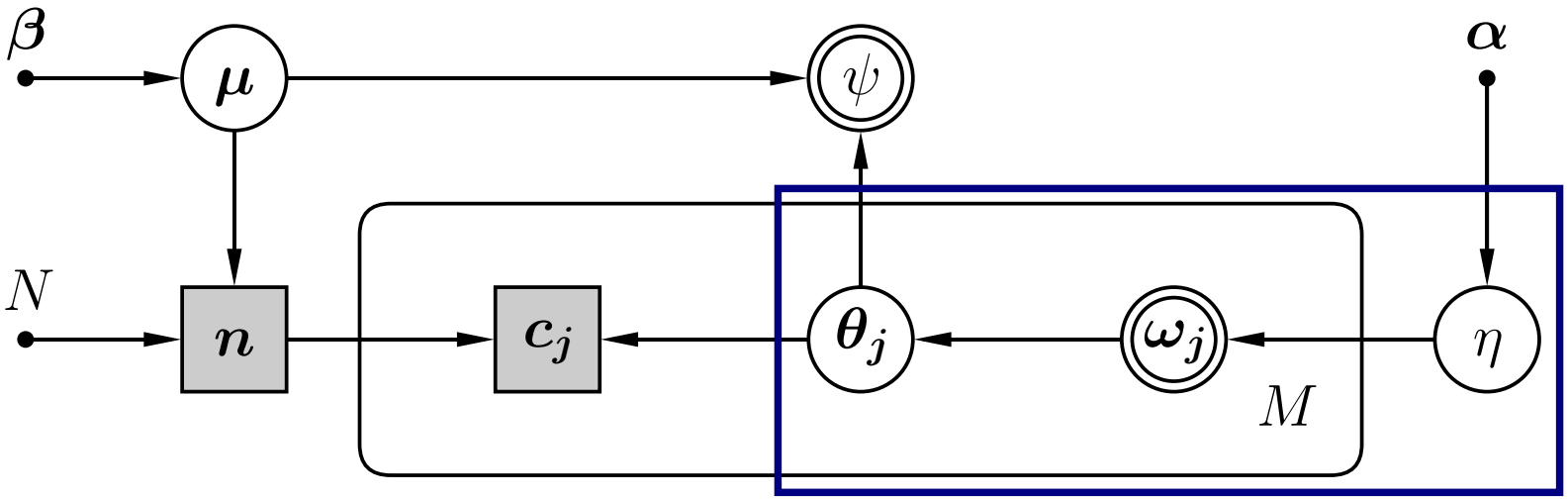
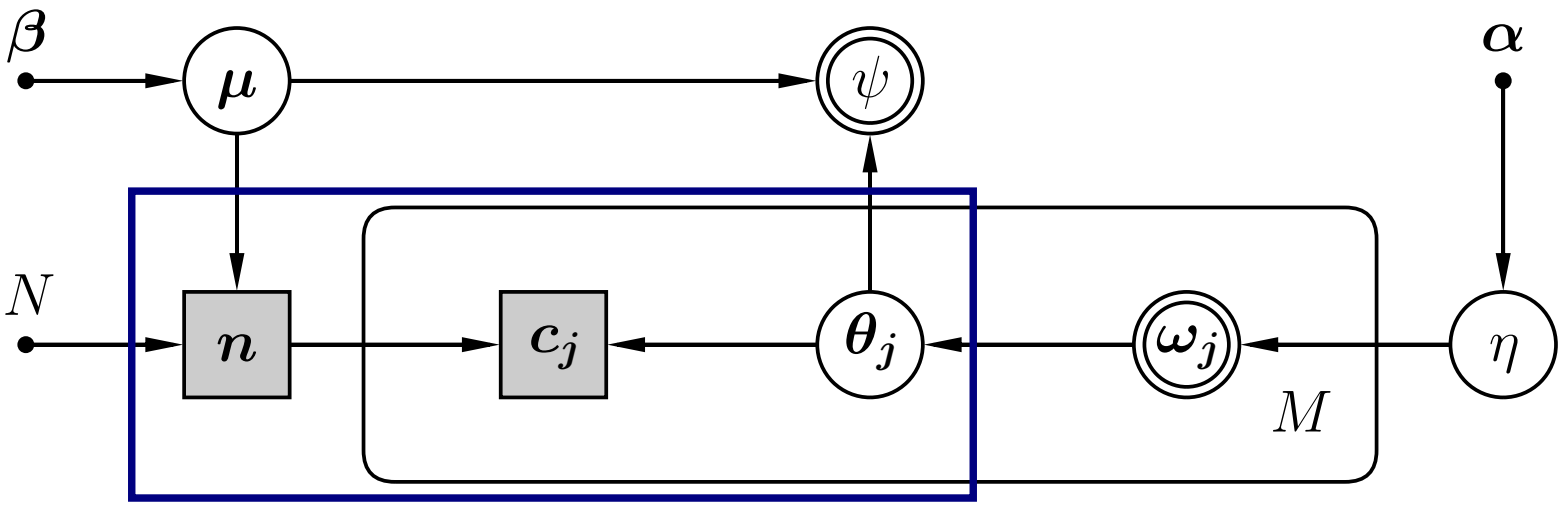


Figure: The probabilistic graphical model for estimating the uncertainty of average F_1 scores.

Models - Performance

- 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 θ_j, where $\sum_{k=1}^{M} c_{jk} = n_j$.



Models - Performance

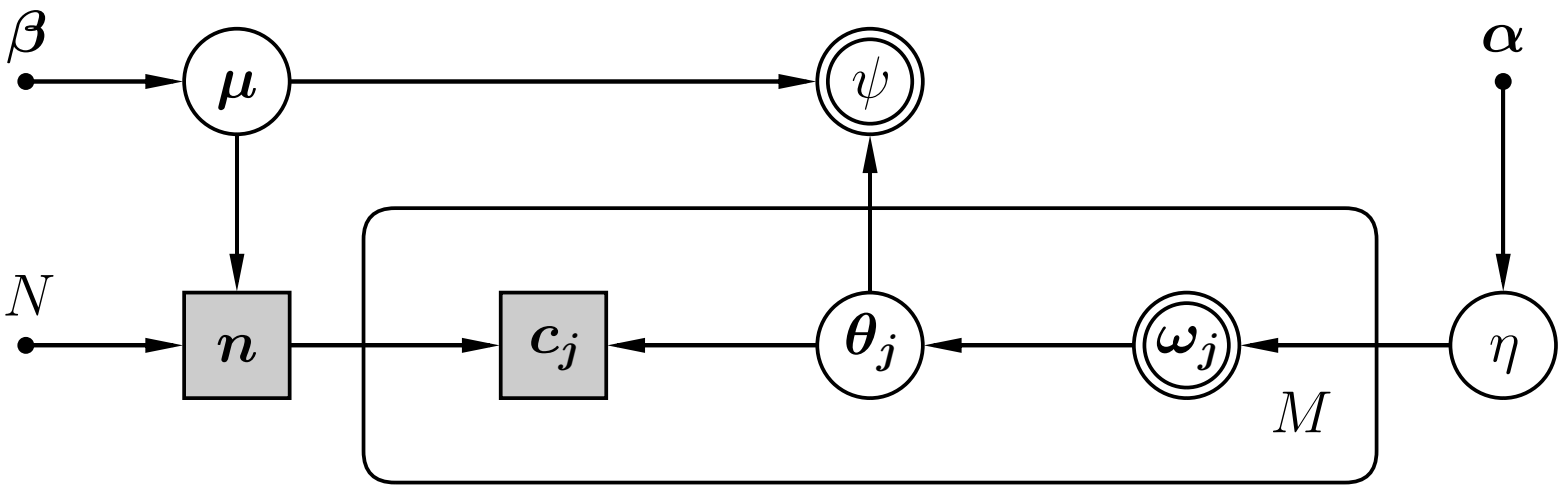
- μ presents the true classification of documents.
- ω presents the predicted classification.
- Treat the performance measure (either miF_1 or maF_1) as a random variable ψ, which is a function of μ and ω. For example, in miF_1

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

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

- In multi-class single-label, mi$F1$ = Precision = Recall.

Models - Performance



- For two models A and B, the difference of the overall performance is represented by δ, where $\delta = \psi_A - \psi_B$.
- Estimate the uncertainty difference of two models by examining the posterior probability distribution of δ.

Outline

Experiments - Dataset

- A standard benchmark dataset for text classification, `20newsgroups`[1].
- 60% subset for training
- 40% subset for testing
- Filtered by striping newsgroup-related metadata

[1] `http://qwone.com/~jason/20Newsgroups/`

Experiments - Classifiers

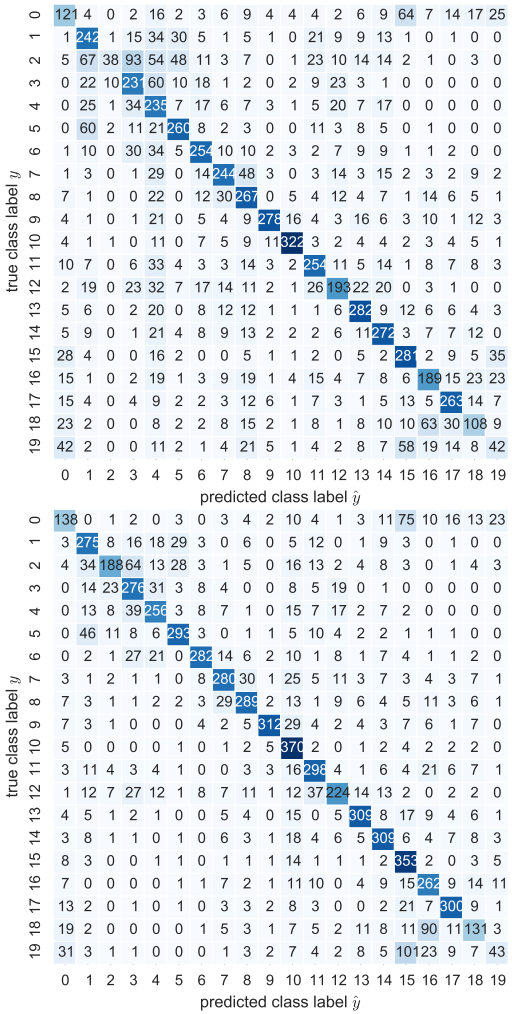
Classification algorithms:

- Naive Bayse (NB)
  - Bernoulli event model (NB_{Bern})
  - Multinomial event model (NB_{Mult})
- linear Support Vector Machine (SVM)
  - *L*1 penalty (SVM_{L1})
  - *L*2 penalty (SVM_{L2})

Implementation of these algorithms:

- Python library `scikit-learn`

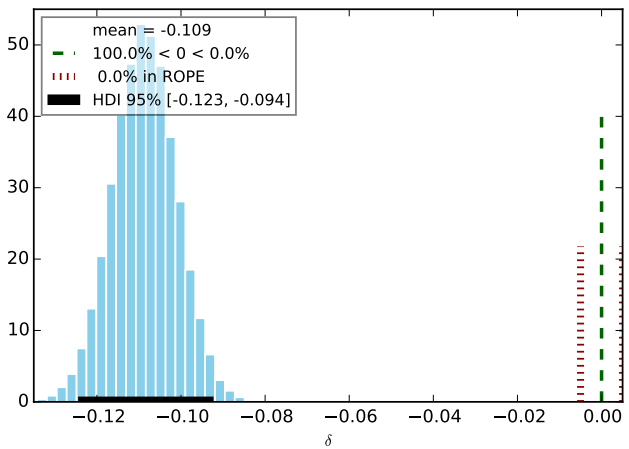
Experiments - Results



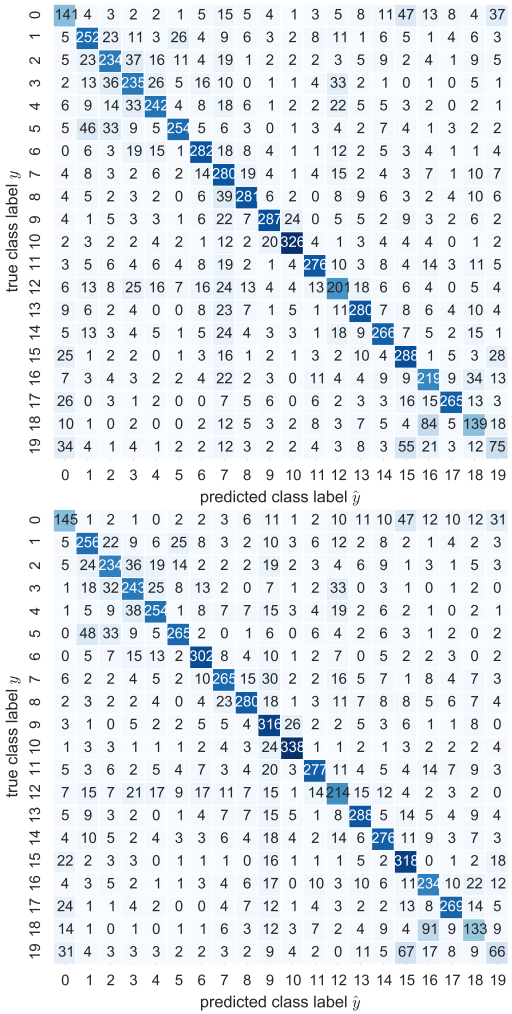
Comparing maF_1 between NB_{Bern} and NB_{Mult}.

Conclusion:

NB_{Bern} is significantly outperformed by NB_{Mult}



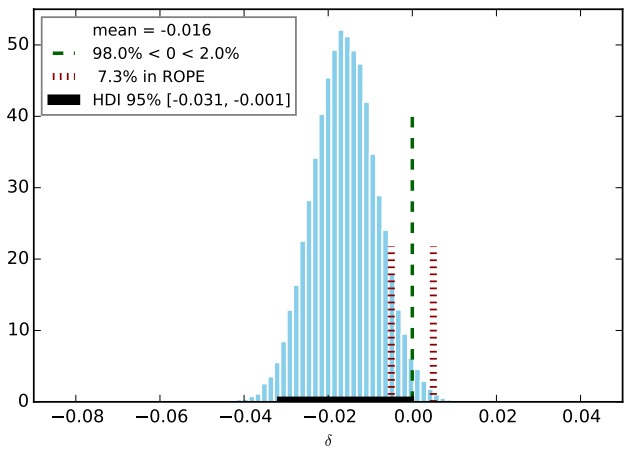
Experiments - Results



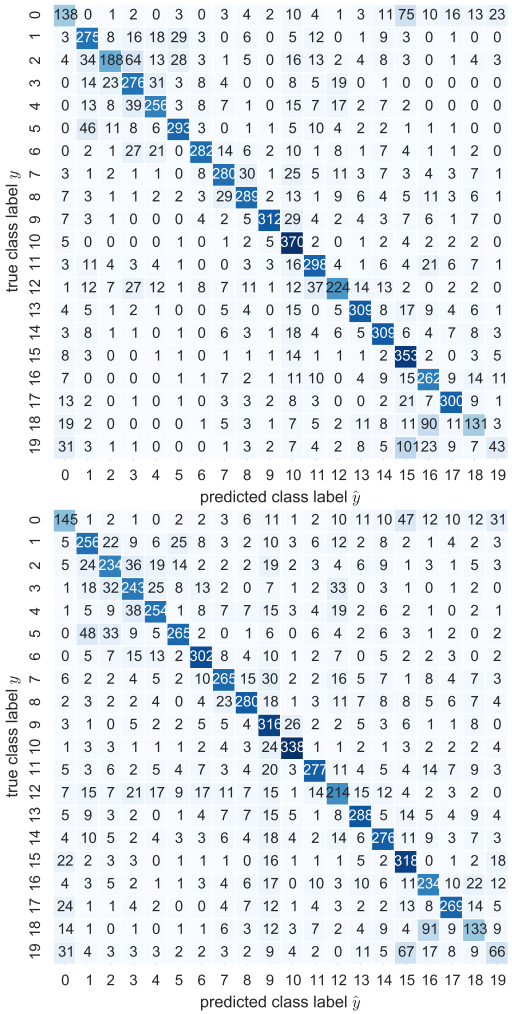
Comparing maF_1 between SVM_{L1} and SVM_{L2}.

Conclusion:

SVM_{L1} is only slightly outperformed by SVM_{L2}



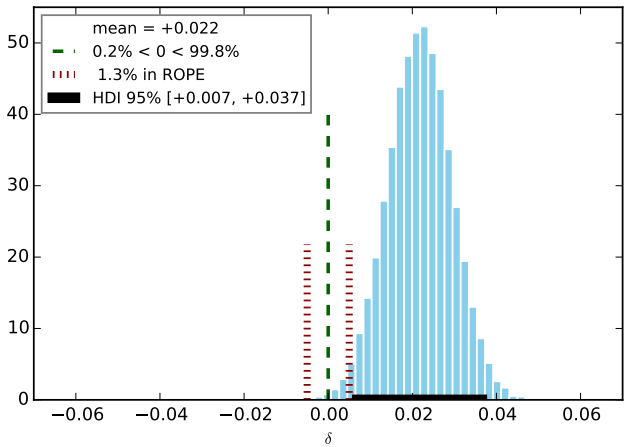
Experiments - Results



Comparing maF_1 between NB_{Mult} and SVM_{L2}.

Conclusion:

NB_{Mult} works a lot better than SVM_{L2}



Summary

- 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.