Importance Weighted Feature Selection Strategy for Text Classification

Baoli LI（李保利）

Henan University of Technology
河南工业大学

Outline

- Introduction:
  - Text Classification and Feature Selection
  - What's the problem of the state-of-the-art FS metrics
- Importance Weighted Feature Selection Strategy
- Experiments and Discussion
- Conclusions and Future Work

Introduction

- Text Classification:
  - assigning one or more predefined categories to a textual segment
    - one-class, binary, multiclass, hierarchical
    - single label vs. multiple labels
    - balanced vs. imbalanced
    - ......
  - larger feature space, higher time and space cost
    - very challenging at the era of big data

Introduction

- Feature Selection:

  - **Task**: finding the most effective feature subset
  - **Benefits**: a) lower time and space cost; b) less overfitting
  - **Methods**: Wrapper and Filter (widely used)

    Filtering Metrics: information gain, chi-square, bi-normal separation, document frequency, odds ratio, mutual information, power, and so on.

Introduction

- What's the problem of the traditional FS metrics:

– To calculate a metric, many statistics need to be collected. During this process, **the traditional methods treat all features equally**, and do not consider whether a feature is an important one in a sample.

– In a textual sample, **some features usually play more important roles than others**.

Introduction

- What's the problem of the traditional FS metrics:

– To calculate a metric, many statistics need to be collected. During this process, **the traditional methods treat all features equally**, and do not consider whether a feature is an important one in a sample.

– In a textual sample, **some features usually play more important roles than others**.

Important features should give more weights? But how can we do that?

Importance Weighted Feature Selection Strategy

For a feature t and a class CLS_i, we need to collect the following numbers and obtain a contingency table:

	CLS_i	Other Classes
t	A_i	B_i
No t	C_i	D_i

A_i: how many documents belong to class CLS_i and contain the feature t;
B_i: how many documents do not belong to class CLS_i but contain the feature t;
C_i: how many documents belong to class CLS_i but do not contain the feature t;
D_i: how many documents do not belong to class CLS_i and, at the same time, do not contain the feature t.

Importance Weighted Feature Selection Strategy

(1) Chi-Square metric

$$CHI = \sum_{i=1}^{M} CHI_i$$

$$= \sum_{i=1}^{M} \frac{(A_i + B_i + C_i + D_i) \times (A_i \times D_i - C_i \times B_i)}{(A_i + C_i) \times (B_i + D_i) \times (A_i + B_i) \times (C_i + D_i)}$$

(2) Information Gain metric

$$IG = \sum_{i=1}^{M} IG_i$$

$$IG_i = e(A_i + C_i, B_i + D_i) - \frac{A_i + B_i}{A_i + B_i + C_i + D_i} e(A_i, B_i) - \frac{C_i + D_i}{A_i + B_i + C_i + D_i} e(C_i, D_i)$$

$$e(x, y) = -\frac{x}{x + y} \log(\frac{x}{x + y}) - \frac{y}{x + y} \log(\frac{y}{x + y})$$

Importance Weighted Feature Selection Strategy

When deriving the following contingency table, we take some different strategies:

	CLS_i	Other Classes
t	A_i	B_i
No t	C_i	D_i

For A_i **and** B_i , when a sample contains the feature *t*, we will add rather than a constant 1 **a real value** between 0 and 1, which indicates how important *t* is in the sample.

For C_i **and** D_i, three options:

- ***MIN:*** to use the minimal importance value of all features;
- ***AVG:*** to use the average importance value;
- ***MAX:*** to use the maximal importance value.

Importance Weighted Feature Selection Strategy

Importance value:

(1) $$I(t, d_j) = \frac{TF_t}{\max\left(TF_1, TF_2, \ldots, TF_{|d_j|}\right)}$$

(2) $$I(t, d_j) = \frac{TFIDF_t}{\max\left(TFIDF_1, TFIDF_2, \ldots, TFIDF_{|d_j|}\right)}$$

Experiments and Discussion

1. ***Goal***: to verify whether the proposed importance weighted feature selection strategy performs better;
   - ***Metrics***: Chi-Square and information gain;
   - ***Algorithm***: Liblinear (performs best overall);
   - ***Datasets***: 20 newsgroups, Sector, Nlpcc2014
2. Other settings:
   - ***Term weighting***: TFIDF (ltc)
   - ***Evaluation measure***: Micro–Averaging F1 and Macro-Averaging F1

Experiments and Discussion

- More information about the Datasets
  - **20 Newsgroups**: balanced, 20 classes,11,293 training samples and 7,528 test samples.
    - 73,712 candidate features
  - **Sector**: modest imbalanced, 105 categories, 6,412 training samples, and 3,207 test samples. The stop words and rare words (DF=1) are removed in this version.
    - 48,988 candidate features
  - **Nlpcc2014**: imbalanced, 247 categories, 11,385 training samples, and 11,577 test samples.
    - 425,488 candidate features

Experiments and Discussion

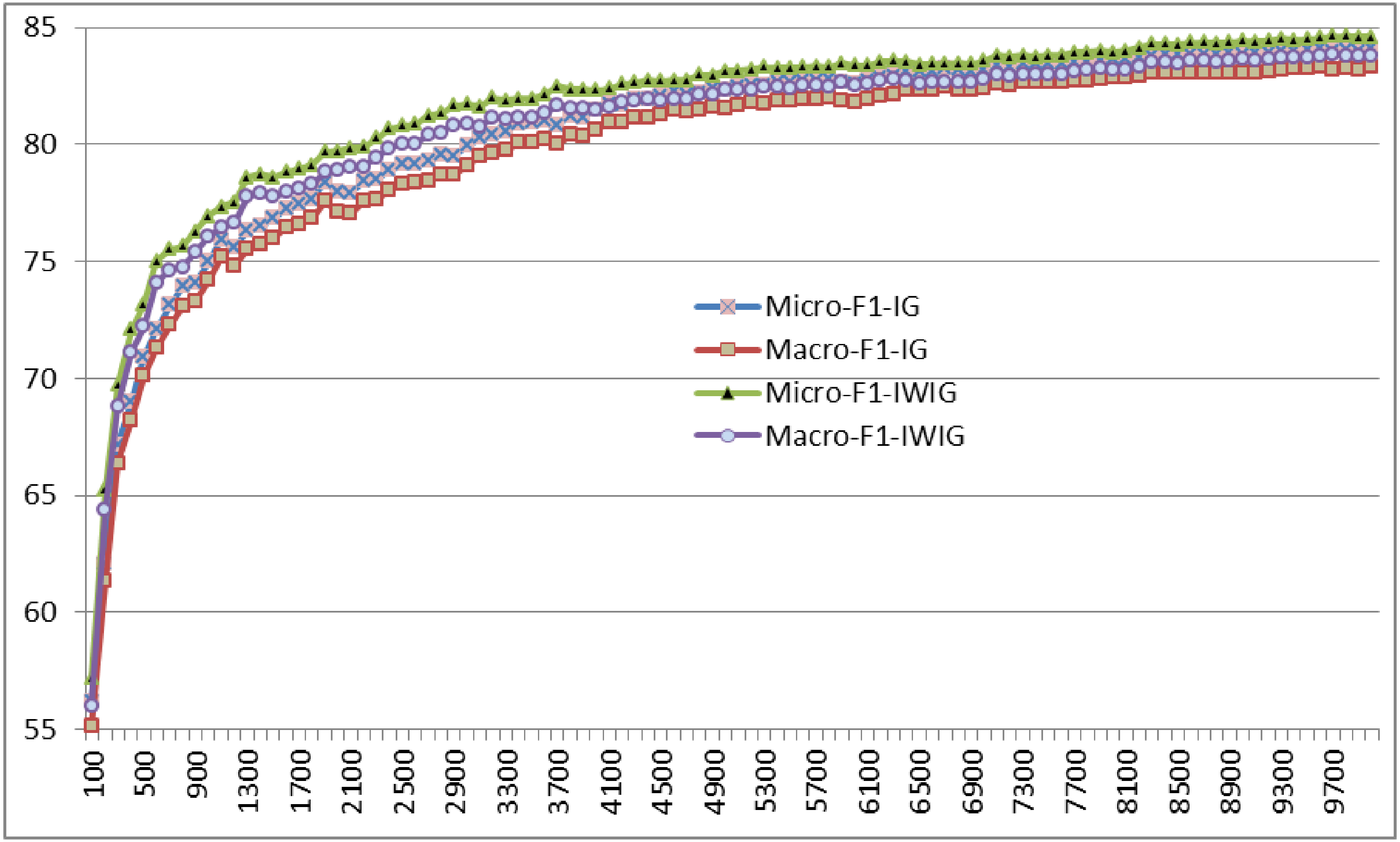

Figure 1. Performance on the 20 newsgroups dataset with Information Gain metric.

Experiments and Discussion

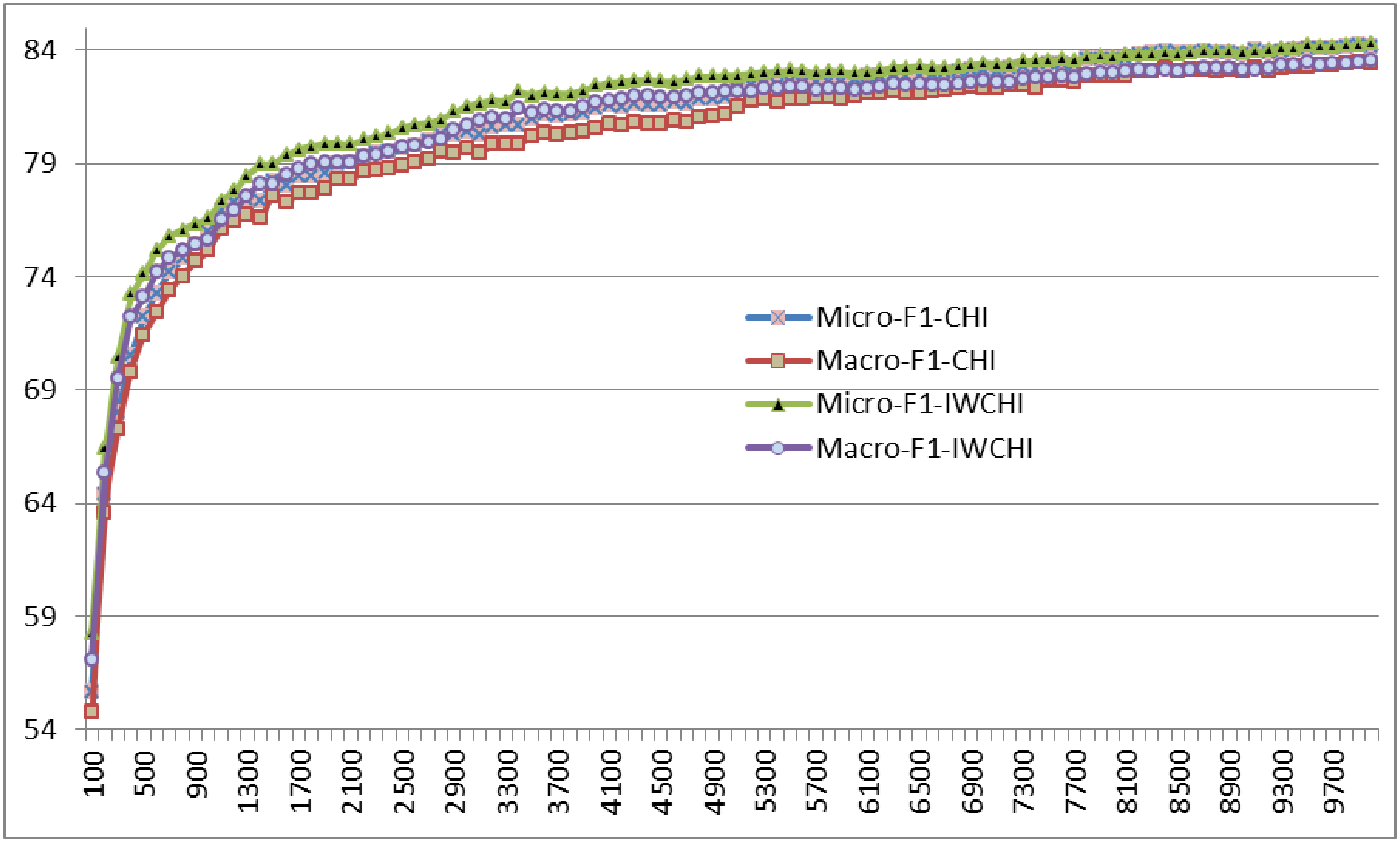

Figure 2. Performance on the 20 newsgroups dataset with Chi-Square metric.

Experiments and Discussion

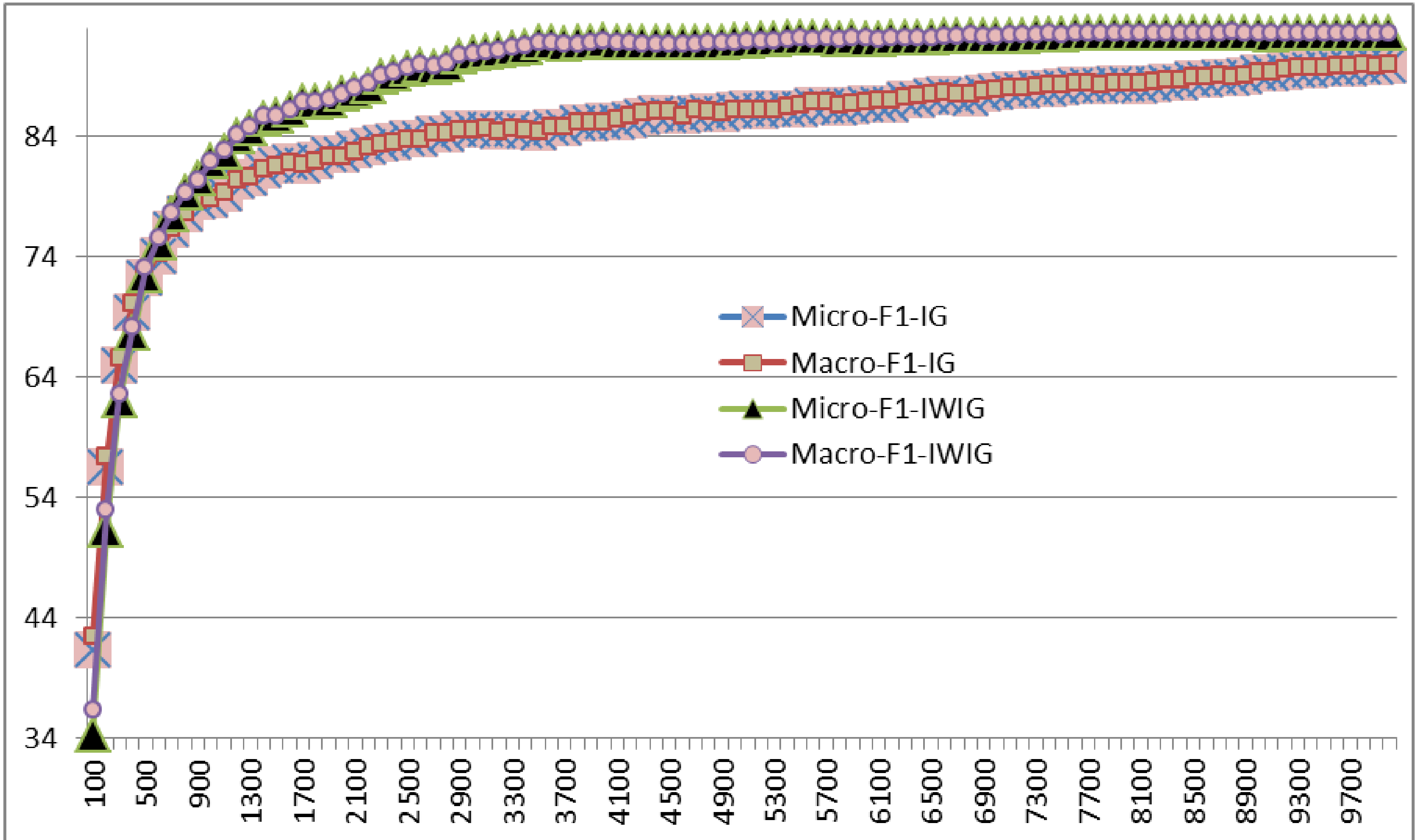

Figure 3. Performance on the Sector dataset with Information Gain metric.

Experiments and Discussion

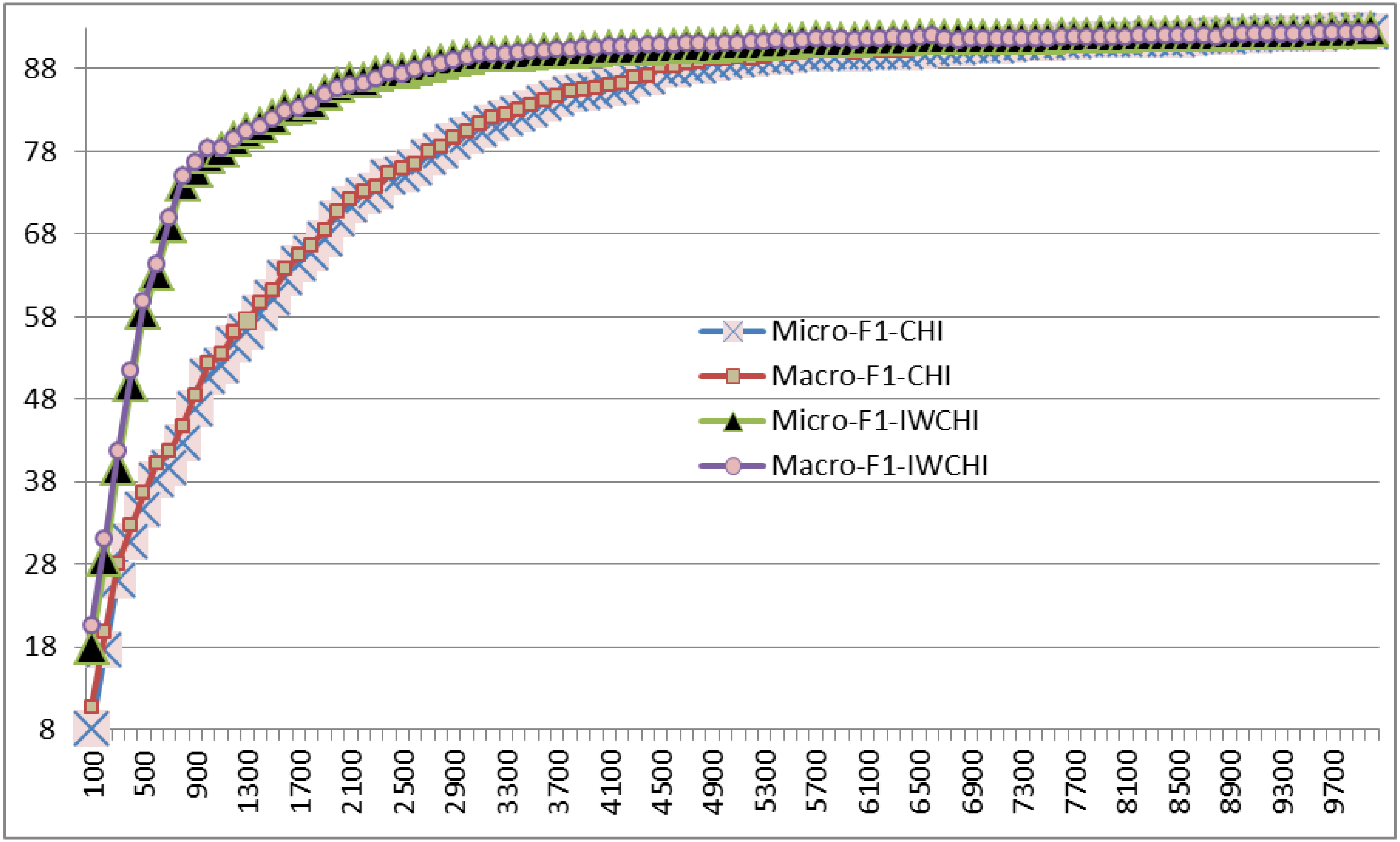

Figure 4. Performance on the Sector dataset with Chi-Square metric.

Experiments and Discussion

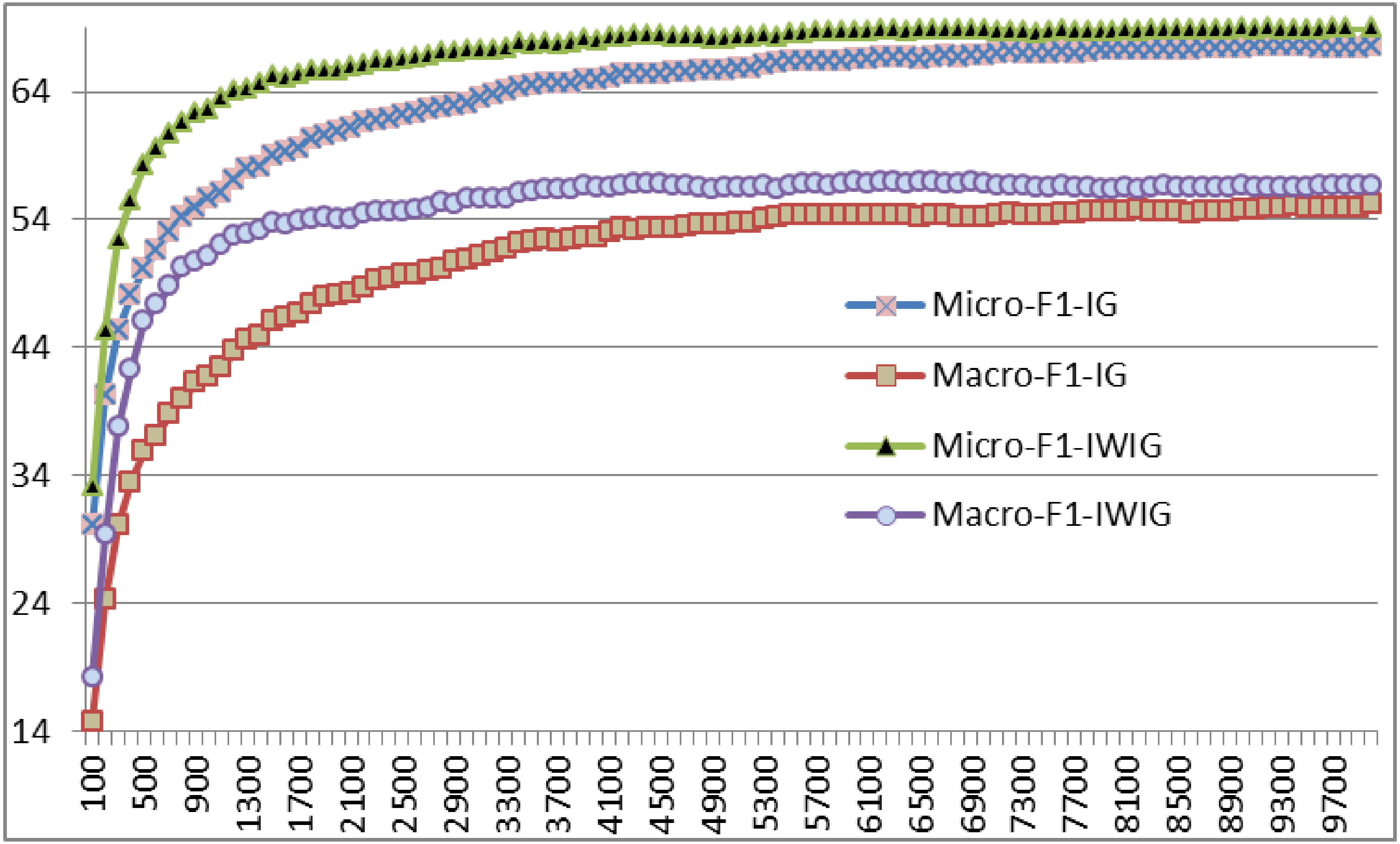

Figure 5. Performance on the Nlpcc2014 dataset with Information Gain metric.

Experiments and Discussion

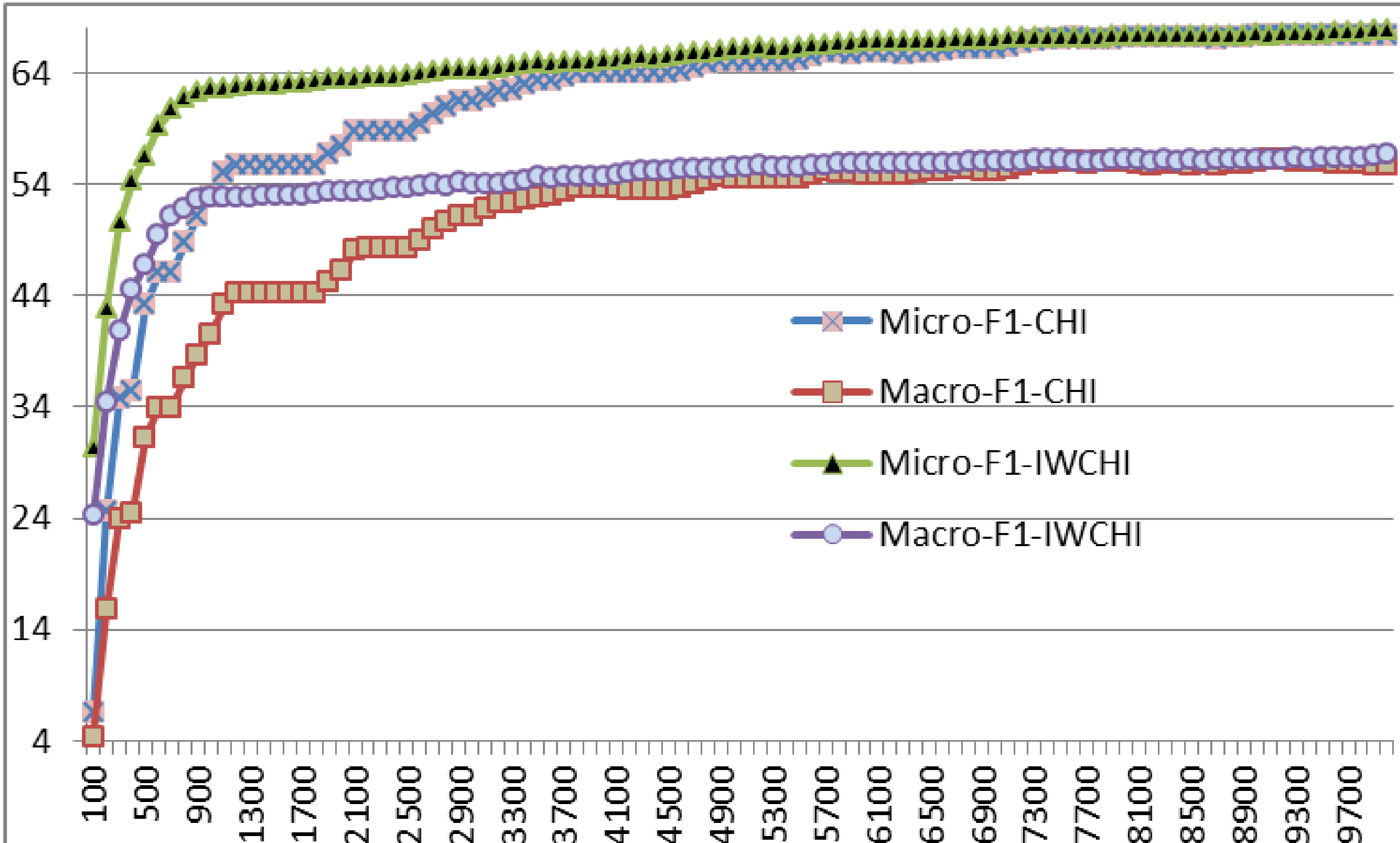

Figure 6. Performance on the Nlpcc2014 dataset with Chi-Square metric.

Conclusions and Future Work

- **The traditional FS metrics do not care about how important a feature is in a sample**, and may introduce much noise.
- A general **importance weighted feature selection strategy** is then proposed. Experiments with two popular FS metrics on three text classification problems demonstrate its effectiveness. The strategy **performs much better on imbalanced datasets**.
- Experiment with more datasets on more text mining applications.
- Apply the strategy into revising other existing feature selection metrics.
- Explore how to better determine the importance of a feature in a sample.

Thanks for your attention!

Questions & Discussion