Weakly-supervised Occupation Detection for Micro-blogging Users

Ying Chen and Bei Pei
China Agricultural University
2014.12.8

Outline

- Background
- Related Work
- Methodology
- Experiments
- Conclusions

Background

- Personal information detection is very important personalization business applications.

- Our task: the occupation detection for the users in Micro-blogging platforms.

Related Work

- Occupation Detection
  - Can be considered as a sub-problem of Information Extraction (IE)
  - The properties of texts determine the approaches of IE (Sarawagi, 2004).

The two types of texts: personal descriptions and tweets

- Imbalanced Classification
  - Sampling: over-sampling methods and under-sampling methods.

A sampling method for the data imbalance and the data noise

Outline

- Background
- Related Work
- Methodology & Experiments
  - The overall architecture
  - The rule-based user occupation detection
  - The MCS-based user occupation detection (MCS: a multiple classifier system)
- Conclusions

The Overall Architecture

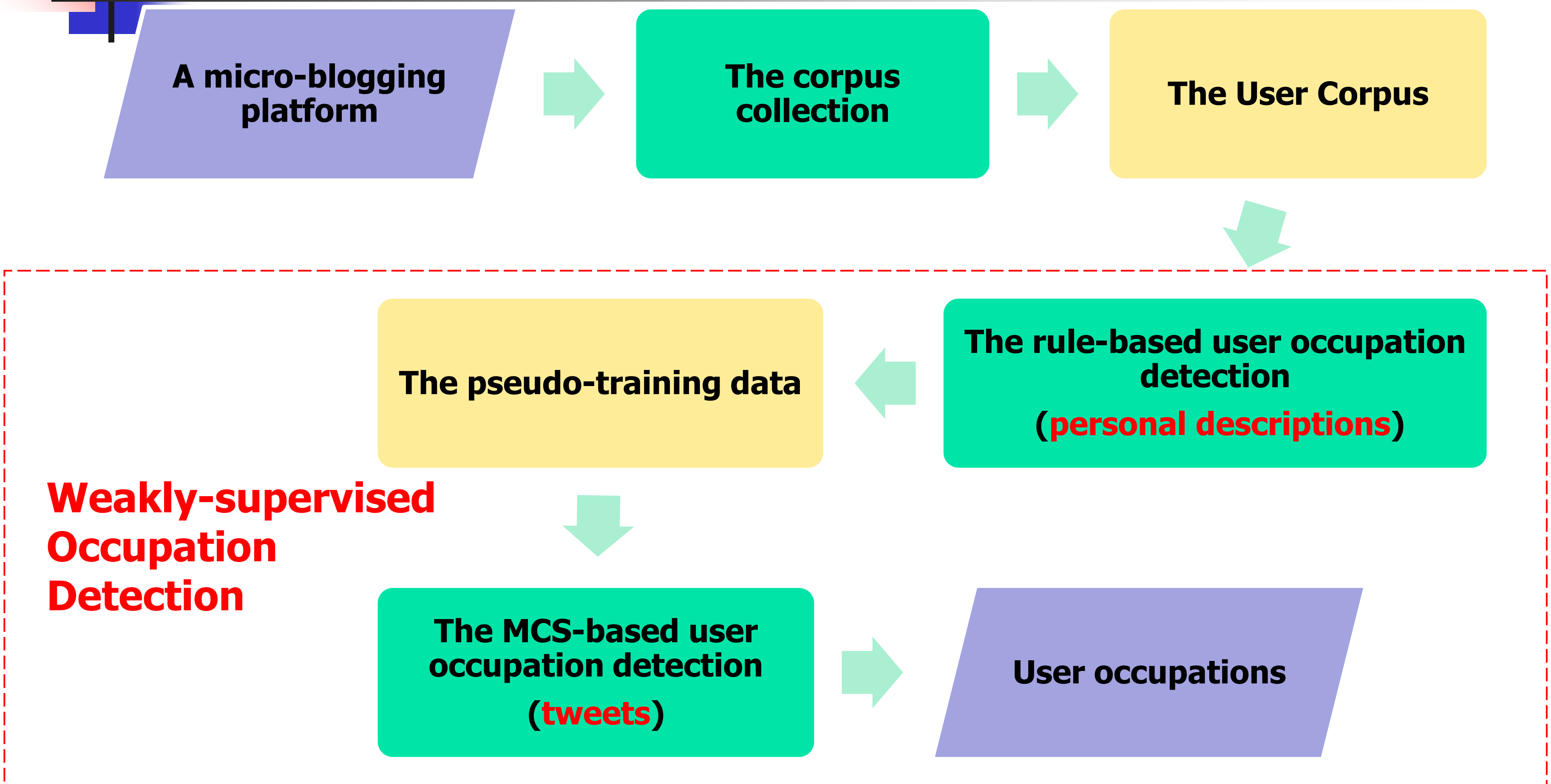

Figure 1: The architecture of our occupation detection

The Rule-based User Occupation Detection

- Detects the occupations of some users according to their personal descriptions.

1. If the job information is provided, the user is tagged as “employee”.
2. If the college information is provided, the user is tagged as “student”.
3. The user is tagged as “undermined”.

The Evaluation of the Rule-based User Occupation Detection (1)

- Datasets:
  - The test/dev datasets:
    - The rule-determined test/dev dataset: ~1000 instances with the tag “student” or “employee”.
    - The rule-undetermined test/dev dataset: ~1000 instances with the tag “undetermined”.
  - The pseudo-training data: ~27,000 instances.

The Evaluation of the Rule-based User Occupation Detection (2)

- Data noise
  - Noisy features
    - Tweets are intrinsically noisy -> the features based on tweets are noisy
  - Noisy pseudo tags

	rule-determined	*rule-undetermined*
Accuracy	~72%	~28%

Table 1: The accuracy of the rule-based user occupation detection on test datasets

The Evaluation of the Rule-based User Occupation Detection (3)

- Data imbalance
  - Eg. the imbalance ratio between "undetermined" and "employee" is ~5 in the pseudo-training data.
  - 3-class classification: student, employee and undetermined

	studen t	employee	un-employed	undetermined
rule-determined	50.8%	36.5%	**1.2%**	**11.5%**
rule-undetermined	40.2%	31.4%	**0.4%**	**28.0%**

Table 2: **The real occupation distribution over test datasets**

The MCS-based User Occupation Detection

data imbalance and data noise

- The training stage:
  - Some training instances selected by our class-based random sampling method.
  - A base classifier is trained with a supervised classification method as well as these training instances .
  - For each training instance, all of the tweets are catenated into a document on which feature extraction works.

data noise

- The test stage: our cascaded ensemble learning method is used.

The Class-based Random Sampling

Input: the initial training data, *K* (a value which controls the size of the outputted training dataset)

Output: a training dataset for a base classifiers

Procedure:

1. For each class c_i (i = 0 to M, M is the class number), K instances are randomly selected from the instances whose tags are c_i in the initial training data.
2. $M*K$ instances are combined to form a training dataset for a base classifier.

The Cascaded Ensemble Learning

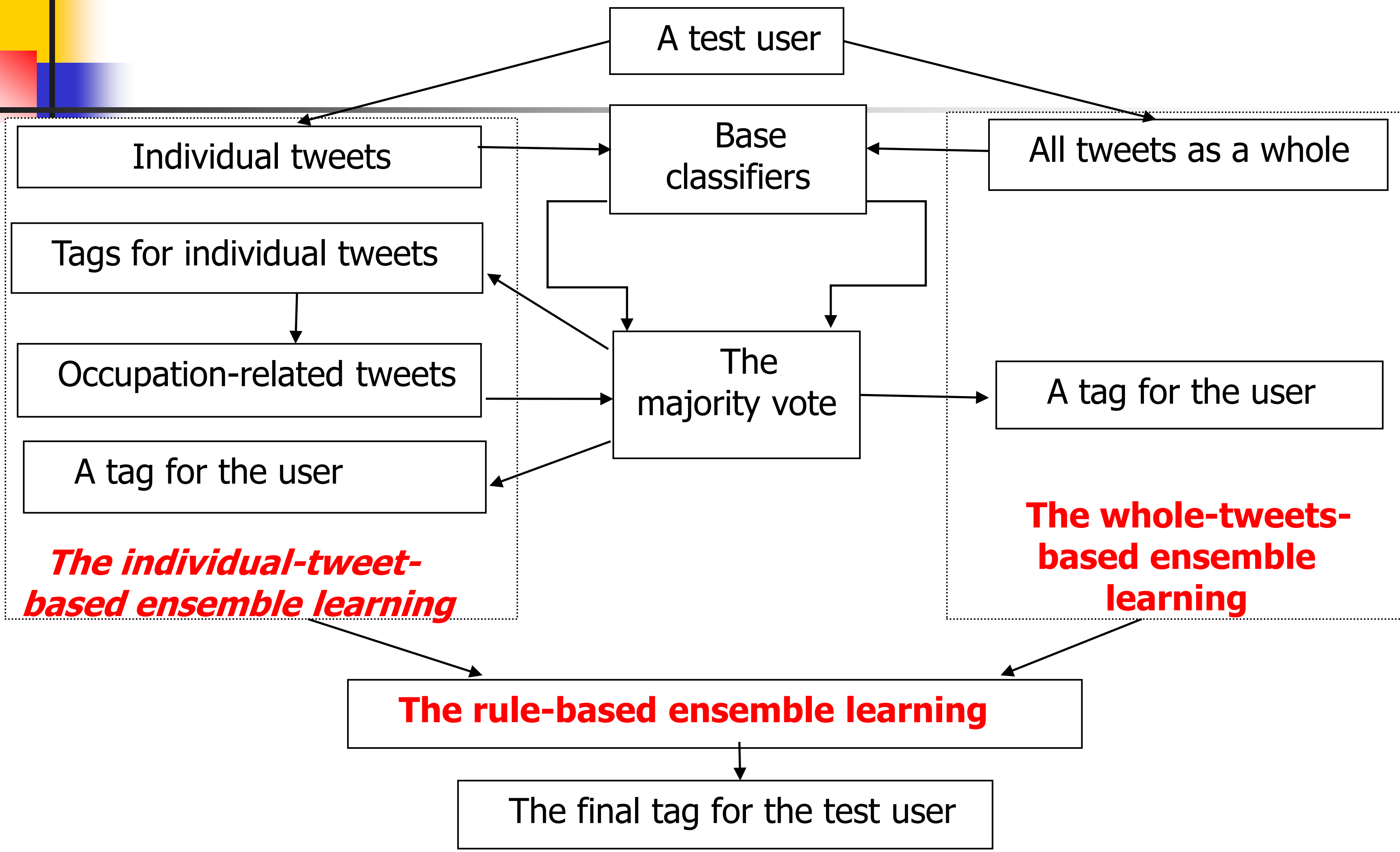

Figure 2: The cascaded ensemble learning method

The Evaluation of the MCS-based User Occupation Detection

- Basedlines :
  - **SC**: a common classification, which uses the following dataset to train one and only one classifier.
    - All of “student” instances, all of “employee” instances and some of “undetermined” instances .
  - **UndSamp+WTEnsem**: uses random under-sampling and the whole-tweets-based ensemble learning.

The Performances of Different Occupation Detection Models

	Prec	Rec	Fs	Acc
(1)SC	62.2	65.6	60.7	67.6
(2)UndSamp+WTEnsem	64.1	66.8	64.7	73.6
(3)RanSamp+WTEnsem	68.6	73.2	69.8	77.0
(4)RanSamp+CasEnsem	69.5	72.6	*70.6*	*77.7*

Table3: The performances for the rule-determined test dataset

	Prec	Rec	Fs	Acc
(1)SC	57.9	54.4	50.3	50.0
(2)UndSamp+WTEnsem	60.2	56.1	54.9	54.6
(3)RanSamp+WTEnsem	63.3	59.9	58.4	58.2
(4)RanSamp+CasEnsem	63.3	62.8	*61.7*	*61.9*

Table 4: The performances for the rule-undetermined test dataset

The Performances of Different Occupation Detection Models

- **SC-> UndSamp+WTEnsem**
  - A significant improvement is achieved.
  - A MCS-based framework with a sampling method can effectively overcome the data imbalance.
- **UndSamp+WTEnsem -> RanSamp+WTEnsem**
  - The performances are further improved
  - Our class-based random sampling can overcome both the data imbalance and the data noise.
- **RanSamp+WTEnsem -> RanSamp+CasEnsem**
  - A significant improvement is achieved for "rule-undet", and a slight improvement for "rule-det".
  - The improvement is from the tag "employee" and "undetermined".
  - Our individual-tweet-based ensemble learning can effectively solve the confusion of "employee vs. undetermined".

The Impact of the Class-based Random Sampling (1)

- Two important parameters:
  - K : the parameter of our class-based random sampling.
  - L : the number of the base classifiers in a MCS-based framework
- Relationships between the two parameters and the performance?

The Impact of the Class-based Random Sampling (2)

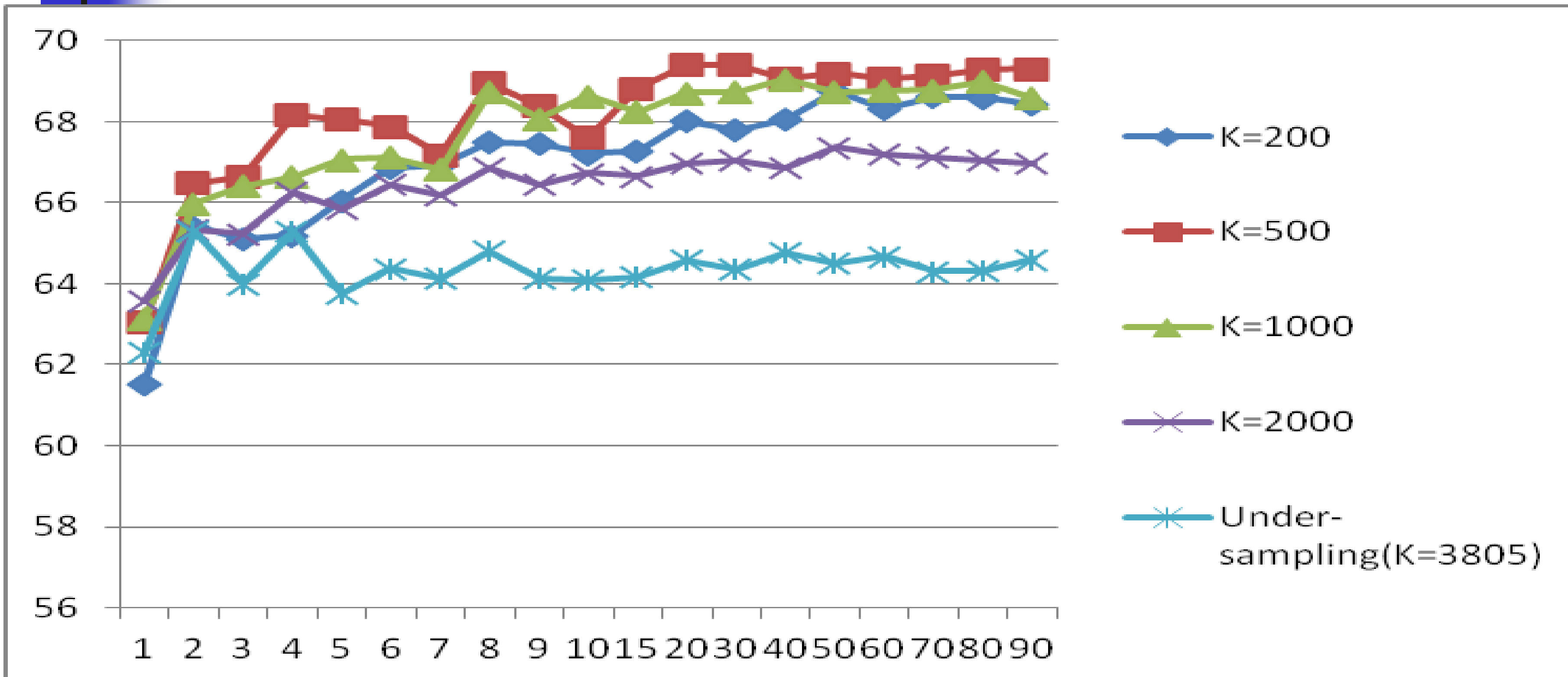

Figure 3. The performances of the RanSamp+CasEnsem model for the rule-determined test (x axis: the value of *L*; y axis: F-score)

The Impact of the Class-based Random Sampling (3)

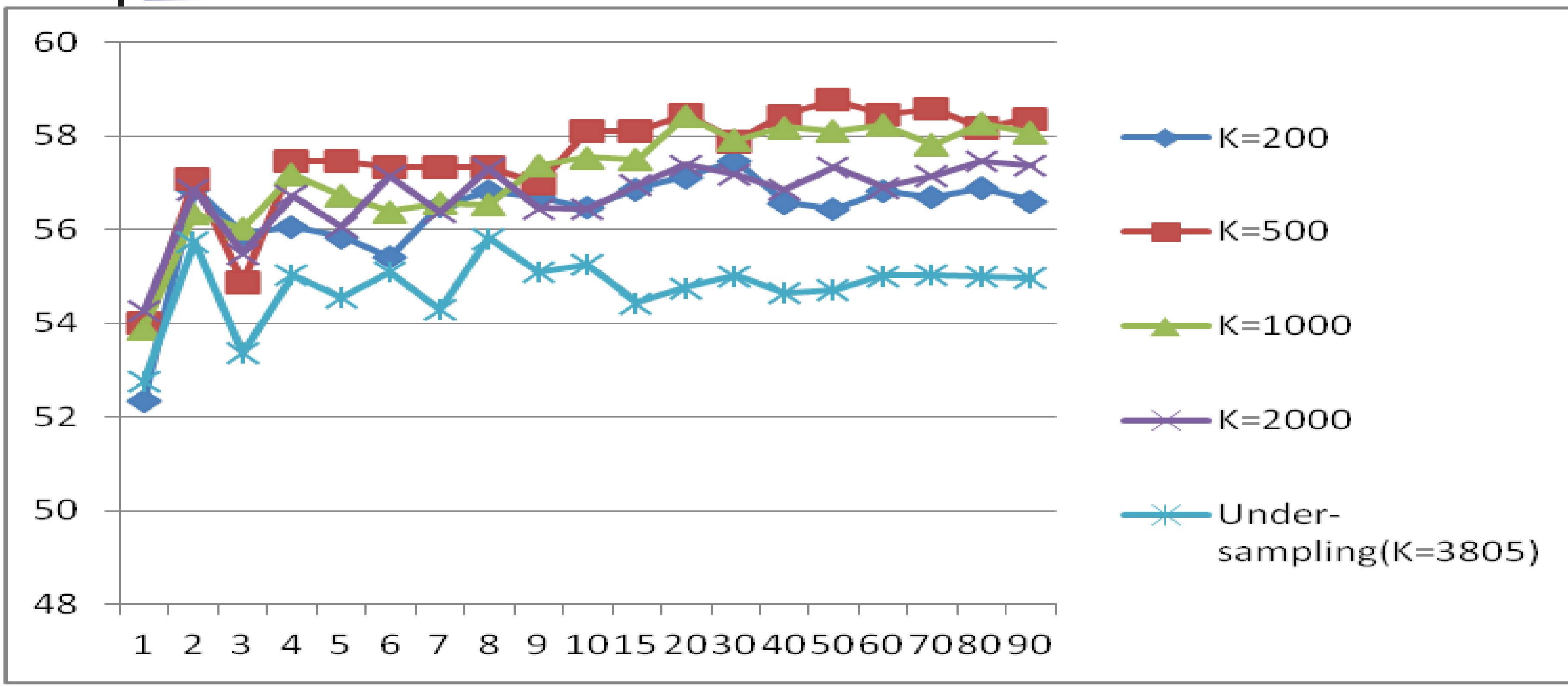

Figure 4. The performances of the RanSamp+CasEnsem model for the rule-undetermined test(x axis: the value of *L*; y axis: F-score)

The Impact of the Class-based Random Sampling (4)

- The impact of *L*
  - For a given *K*, the curve greatly varies when *L* is small, and becomes stable when *L* is large enough.
    - The performance of a supervised model is often determined by the size of the training data.
  - For a given *K*, the curve generally increases with the increasing *L*.
    - Even if the initial training data are noisy, more diverse training datasets -> an effective feature is likely to be selected.

- The impact of *K*
  - For a given *L*, the performance generally increases when *K* decreases from 3805 (the under-sampling) to 500.
    - The larger training dataset is -> the more conflicts -> the more confused a base classifier is
  - For a given *L*, the performance generally decreases when *K* decreases from 500 to 200.
    - Too small training dataset

Conclusions

- Proposed a weakly-supervised user occupation detection which achieves a significant improvement.
- Examine the contributions of different kind of user textual information to the occupation detection.
- Propose the class-based random sampling and the cascaded ensemble learning to overcome the data noise problem.

Questions?