

# Active Learning for Imbalanced Sentiment Classification

Shoushan Li<sup>†</sup>, Shengfeng Ju<sup>†</sup>, Guodong Zhou<sup>†\*</sup>

<sup>†</sup>Natural Language Processing Lab  
School of Computer Science and Technology  
Soochow University, Suzhou, 215006, China  
{shoushan.li, shengfeng.ju}@gmail.com,  
gdzhou@suda.edu.cn

Xiaojun Li<sup>‡</sup>

<sup>‡</sup>College of Computer and  
Information Engineering  
Zhejiang Gongshang University  
Hangzhou, 310035, China  
lixj@mail.zjgsu.edu.cn

## Abstract

Active learning is a promising way for sentiment classification to reduce the annotation cost. In this paper, we focus on the imbalanced class distribution scenario for sentiment classification, wherein the number of positive samples is quite different from that of negative samples. This scenario posits new challenges to active learning. To address these challenges, we propose a novel active learning approach, named co-selecting, by taking both the imbalanced class distribution issue and uncertainty into account. Specifically, our co-selecting approach employs two feature subspace classifiers to collectively select most informative minority-class samples for manual annotation by leveraging a certainty measurement and an uncertainty measurement, and in the meanwhile, automatically label most informative majority-class samples, to reduce human-annotation efforts. Extensive experiments across four domains demonstrate great potential and effectiveness of our proposed co-selecting approach to active learning for imbalanced sentiment classification.

## 1 Introduction

Sentiment classification is the task of identifying the sentiment polarity (e.g., positive or negative) of

a natural language text towards a given topic (Pang et al., 2002; Turney, 2002) and has become the core component of many important applications in opinion analysis (Cui et al., 2006; Li et al., 2009; Lloret et al., 2009; Zhang and Ye, 2008).

Most of previous studies in sentiment classification focus on learning models from a large number of labeled data. However, in many real-world applications, manual annotation is expensive and time-consuming. In these situations, active learning approaches could be helpful by actively selecting most informative samples for manual annotation. Compared to traditional active learning for sentiment classification, active learning for imbalanced sentiment classification faces some unique challenges.

As a specific type of sentiment classification, imbalanced sentiment classification deals with the situation in which there are many more samples of one class (called *majority class*) than the other class (called *minority class*), and has attracted much attention due to its high realistic value in real-world applications (Li et al., 2011a). In imbalanced sentiment classification, since the minority-class samples (denoted as *MI* samples) are normally much sparse and thus more precious and informative for learning compared to the majority-class ones (denoted as *MA* samples), it is worthwhile to spend more on manually annotating *MI* samples to guarantee both the quality and quantity of *MI* samples. Traditionally, uncertainty has been popularly used as a basic measurement in active learning (Lewis and Gale, 2004). Therefore, how to select most informative *MI* samples for manual annotation without violating the basic

---

\* Corresponding author

uncertainty requirement in active learning is challenging in imbalanced sentiment classification.

In this paper, we address above challenges in active learning for imbalanced sentiment classification. The basic idea of our active learning approach is to use two complementary classifiers for collectively selecting most informative *MI* samples: one to adopt a certainty measurement for selecting most possible *MI* samples and the other to adopt an uncertainty measurement for selecting most uncertain *MI* samples from the most possible *MI* samples returned from the first classifier. Specifically, the two classifiers are trained with two disjoint feature subspaces to guarantee their complementariness. This also applies to selecting most informative *MA* samples. We call our novel active learning approach co-selecting due to its collectively selecting informative samples through two disjoint feature subspace classifiers. To further reduce the annotation efforts, we only manually annotate those most informative *MI* samples while those most informative *MA* samples are automatically labeled using the predicted labels provided by the first classifier.

In principle, our active learning approach differs from existing ones in two main aspects. First, a certainty measurement and an uncertainty measurement are employed in two complementary subspace classifiers respectively to collectively select most informative *MI* samples for manual annotation. Second, most informative *MA* samples are automatically labeled to further reduce the annotation cost. Evaluation across four domains shows that our active learning approach is effective for imbalanced sentiment classification and significantly outperforms the state-of-the-art active learning alternatives, such as uncertainty sampling (Lewis and Gale, 2004) and co-testing (Muslea et al., 2006).

The remainder of this paper is organized as follows. Section 2 overviews the related work on sentiment classification and active learning. Section 3 proposes our active learning approach for imbalanced sentiment classification. Section 4 reports the experimental results. Finally, Section 5 draws the conclusion and outlines the future work.

## 2 Related Work

In this section, we give a brief overview on sentiment classification and active learning.

### 2.1 Sentiment Classification

Sentiment classification has become a hot research topic in NLP community and various kinds of classification methods have been proposed, such as unsupervised learning methods (Turney, 2002), supervised learning methods (Pang et al., 2002), semi-supervised learning methods (Wan, 2009; Li et al., 2010), and cross-domain classification methods (Blitzer et al., 2007; Li and Zong, 2008; He et al., 2011). However, imbalanced sentiment classification is relatively new and there are only a few studies in the literature.

Li et al. (2011a) pioneer the research in imbalanced sentiment classification and propose a co-training algorithm to perform semi-supervised learning for imbalanced sentiment classification with the help of a great amount of unlabeled samples. However, their semi-supervised approach to imbalanced sentiment classification suffers from the problem that their balanced selection strategy in co-training would generate many errors in late iterations due to the imbalanced nature of the unbalanced data. In comparison, our proposed active learning approach can effectively avoid this problem. By the way, it is worth to note that the experiments therein show the superiority of under-sampling over other alternatives such as cost-sensitive and one-class classification for imbalanced sentiment classification.

Li et al. (2011b) focus on supervised learning for imbalanced sentiment classification and propose a clustering-based approach to improve traditional under-sampling approaches. However, the improvement of the proposed clustering-based approach over under-sampling is very limited.

Unlike all the studies mentioned above, our study pioneers active learning on imbalanced sentiment classification.

### 2.2 Active Learning

Active learning, as a standard machine learning problem, has been extensively studied in many research communities and several approaches have been proposed to address this problem (Settles, 2009). Based on different sample selection strategies, they can be grouped into two main categories: (1) uncertainty sampling (Lewis and Gale, 2004) where the active learner iteratively select most uncertain unlabeled samples for manual annotation; and (2) committee-based

sampling where the active learner selects those unlabeled samples which have the largest disagreement among several committee classifiers. Besides query by committee (QBC) as the first of such type (Freund et al., 1997), co-testing learns a committee of member classifiers from different views and selects those contention points (i.e., unlabeled examples on which the views predict different labels) for manual annotation (Muslea et al., 2006).

However, most previous studies focus on the scenario of balanced class distribution and only a few recent studies address the active learning issue on imbalanced classification problems including Yang and Ma (2010), Zhu and Hovy (2007), Ertekin et al. (2007a) and Ertekin et al. (2007b)<sup>2</sup>. Unfortunately, they straightly adopt the uncertainty sampling as the active selection strategy to address active learning in imbalanced classification, which completely ignores the class imbalance problem in the selected samples.

Attenberg and Provost (2010) highlights the importance of selecting samples by considering the proportion of the classes. Their simulation experiment on text categorization confirms that selecting class-balanced samples is more important than traditional active selection strategies like uncertainty. However, the proposed experiment is simulated and non real strategy is proposed to balance the class distribution of the selected samples.

Doyle et al. (2011) propose a real strategy to select balanced samples. They first select a set of uncertainty samples and then randomly select balanced samples from the uncertainty-sample set. However, the classifier used for selecting balanced samples is the same as the one for supervising uncertainty, which makes the balance control unreliable (the selected uncertainty samples take very low confidences which are unreliable to correctly predict the class label for controlling the balance). Different from their study, our approach possesses two merits: First, two feature subspace classifiers are trained to finely integrate the certainty and uncertainty measurements. Second, the *MA* samples are automatically annotated,

<sup>2</sup> Ertekin et al. (2007a) and Ertekin et al. (2007b) select samples closest to the hyperplane provided by the SVM classifier (within the margin). Their strategy can be seen as a special case of uncertainty sampling.

which reduces the annotation cost in a further effort.

### 3 Active Learning for Imbalanced Sentiment Classification

Generally, active learning can be either stream-based or pool-based (Sassano, 2002). The main difference between the two is that the former scans through the data sequentially and selects informative samples individually, whereas the latter evaluates and ranks the entire collection before selecting most informative samples at batch. As a large collection of samples can easily gathered once in sentiment classification, pool-based active learning is adopted in this study.

Figure 1 illustrates a standard pool-based active learning approach, where the most important issue is the sampling strategy, which evaluates the informativeness of one sample.

---

**Input:**

Labeled data  $L$ ;  
Unlabeled pool  $U$ ;

**Output:**

New Labeled data  $L$

**Procedure:**

Loop for  $N$  iterations:

- (1). Learn a classifier using current  $L$
  - (2). Use current classifier to label all unlabeled samples
  - (3). Use the sampling strategy to select  $n$  most informative samples for manual annotation
  - (4). Move newly-labeled samples from  $U$  to  $L$
- 

Figure 1: Pool-based active learning

#### 3.1 Sampling Strategy: Uncertainty vs. Certainty

As one of the most popular selection strategies in active learning, uncertainty sampling depends on an uncertainty measurement to select informative samples. Since sentiment classification is a binary classification problem, the uncertainty measurement of a document  $d$  can be simply defined as follows:

$$Uncer(d) = \min_{y \in \{pos, neg\}} P(y|d)$$

Where  $P(y|d)$  denotes the posterior probability of the document  $d$  belonging to the class  $y$  and  $\{pos,$

neg} denotes the class labels of positive and negative.

In imbalanced sentiment classification, *MI* samples are much sparse yet precious for learning and thus are believed to be more valuable for manual annotation. The key in active learning for imbalanced sentiment classification is to guarantee both the quality and quantity of newly-added *MI* samples. To guarantee the selection of *MI* samples, a certainty measurement is necessary. In this study, the certainty measurement is defined as follows:

$$Cer(d) = \max_{y \in \{pos, neg\}} P(y|d)$$

Meanwhile, in order to balance the samples in the two classes, once an informative *MI* sample is manually annotated, an informative *MA* sample is automatically labeled. In this way, the annotated data become more balanced than a random selection strategy.

However, the two sampling strategies discussed above are apparently contradicted: while the uncertainty measurement is prone to selecting the samples whose posterior probabilities are nearest to 0.5, the certainty measurement is prone to selecting the samples whose posterior probabilities are nearest to 1. Therefore, it is essential to find a solution to balance uncertainty sampling and certainty sampling in imbalanced sentiment classification,

### 3.2 Co-selecting with Feature Subspace Classifiers

In sentiment classification, a document is represented as a feature vector generated from the feature set  $F = \{f_1, \dots, f_m\}$ . When a feature subset, i.e.,  $F^S = \{f_1^S, \dots, f_r^S\}$  ( $r < m$ ), is used, the original  $m$ -dimensional feature space becomes an  $r$ -dimensional feature subspace. In this study, we call a classifier trained with a feature subspace a feature subspace classifier.

Our basic idea of balancing both the uncertainty measurement and the certainty measurement is to train two subspace classifiers to adopt them respectively. In our implementation, we randomly select two disjoint feature subspaces, each of which is used to train a subspace classifier. On one side, one subspace classifier is employed to select some certain samples; on the other side, the other classifier is employed to select the most uncertain sample from those certain samples for manual

annotation. In this way, the selected samples are certain in terms of one feature subspace for selecting more possible *MI* samples. Meanwhile, the selected sample remains uncertain in terms of the other feature subspace to introduce uncertain knowledge into current learning model. We name this approach as co-selecting because it collectively selects informative samples by two separate classifiers. Figure 2 illustrates the co-selecting algorithm. In our algorithm, we strictly constrain the balance of the samples between the two classes, i.e., positive and negative. Therefore, once two samples are annotated with the same class label, they will not be added to the labeled data, as shown in step (7) in Figure 2.

---

#### Input:

Labeled data  $L$  with balanced samples over the two classes

Unlabeled pool  $U$

#### Output:

New Labeled data  $L$

#### Procedure:

Loop for  $N$  iterations:

- (1). Randomly select a feature subset  $F^S$  with size  $r$  (with the proportion  $\theta = r/m$ ) from  $F$
  - (2). Generate a feature subspace from  $F^S$  and train a corresponding feature subspace classifier  $C_{Cer}$  with  $L$
  - (3). Generate another feature subspace from the complement set of  $F^S$ , i.e.,  $F - F^S$  and train a corresponding feature subspace classifier  $C_{Uncer}$  with  $L$
  - (4). Use  $C_{Cer}$  to select top certain  $k$  positive and  $k$  negative samples, denoted as a sample set  $CER_1$
  - (5). Use  $C_{Uncer}$  to select the most uncertain positive sample and negative sample from  $CER_1$
  - (6). Manually annotate the two selected samples
  - (7). If the annotated labels of the two selected samples are different from each other:  
Add the two newly-annotated samples into  $L$
- 

Figure 2: The co-selecting algorithm

There are two parameters in the algorithm: the size of the feature subspace for training the first subspace classifier, i.e.,  $\theta$  and the number of

selected certain samples, i.e.,  $k$ . Both of the two parameters will be empirically studied in our experiments.

### 3.3 Co-selecting with Selected *MA* Samples Automatically Labeled

---

**Input:**

Labeled data  $L$  with balanced samples over the two classes

Unlabeled pool  $U$

*MA* and *MI* Label (positive or negative)

**Output:**

New Labeled data  $L$

**Procedure:**

Loop for  $N$  iterations:

- (1). Randomly select a proportion of features (with the proportion  $\theta$ ) from  $F$  to get a feature subset  $F^S$
  - (2). Generate a feature subspace from  $F^S$  and train a corresponding subspace classifier  $C_{Cer}$  with  $L$
  - (3). Generate another feature subspace from the complement set of  $F^S$ , i.e.,  $F - F^S$  and train a corresponding subspace classifier  $C_{Uncer}$  with  $L$
  - (4). Use  $C_{Cer}$  to select top certain  $k$  positive and  $k$  negative samples, denoted as a sample set  $CER_1$
  - (5). Use  $C_{Uncer}$  to select the most uncertain positive sample and negative sample from  $CER_1$
  - (6). Manually annotate the sample that is predicted as a *MI* sample by  $C_{Cer}$  and automatically annotate the sample that is predicted as *majority class*
  - (7). If the annotated labels of the two selected samples are different from each other:  
Add the two newly-annotated samples into  $L$
- 

Figure 3: The co-selecting algorithm with selected *MA* samples automatically labeled

To minimize manual annotation, it is a good choice to automatically label those selected *MA* samples. In our co-selecting approach, automatically labeling those selected *MA* samples is easy and

straightforward: the subspace classifier for monitoring the certainty measurement provides an ideal solution to annotate the samples that have been predicted as *majority class*. Figure 3 shows the co-selecting algorithm with those selected *MA* samples automatically labeled. The main difference from the original co-selecting is shown in Step (6) in Figure 3. Another difference is the input where a prior knowledge of which class is *majority class* or *minority class* should be known. In real applications, it is not difficult to know this. We first use a classifier trained with the initial labeled data to test all unlabeled data. If the predicted labels in the classification results are greatly imbalanced, we can assume that the unlabeled data is imbalanced, and consider the dominated class as *majority class*.

## 4 Experimentation

In this section, we will systematically evaluate our active learning approach for imbalanced sentiment classification and compare it with the state-of-the-art active learning alternatives.

### 4.1 Experimental Setting

**Dataset**

We use the same data as used by Li et al. (2011a). The data collection consists of four domains: Book, DVD, Electronic, and Kitchen (Blitzer et al., 2007). For each domain, we randomly select an initial balanced labeled data with 50 negative samples and 50 positive samples. For the unlabeled data, we randomly select 2000 negative samples, and 14580/12160/7140/7560 positive samples from the four domains respectively, keeping the same imbalanced ratio as the whole data. For the test data in each domain, we randomly extract 800 negative samples and 800 positive samples.

**Classification algorithm**

The Maximum Entropy (ME) classifier implemented with the Mallet<sup>3</sup> tool is mainly adopted, except that in the margin-based active learning approach (Ertekin et al., 2007a) where SVM is implemented with light-SVM<sup>4</sup>. The features for classification are unigram words with Boolean weights.

---

<sup>3</sup> <http://mallet.cs.umass.edu/>

<sup>4</sup> [http://www.cs.cornell.edu/people/tj/svm\\_light/](http://www.cs.cornell.edu/people/tj/svm_light/)

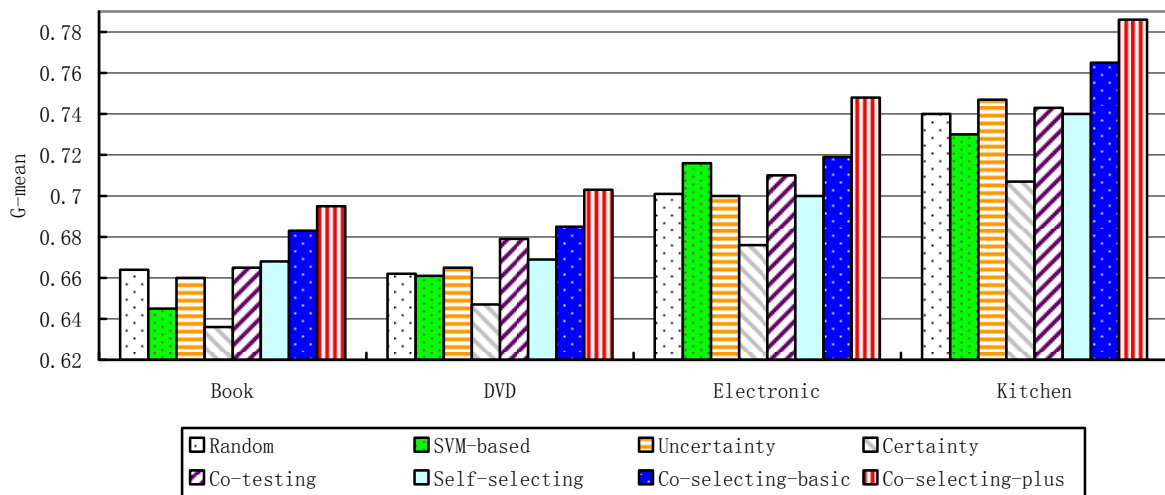


Figure 4: Performance comparison of different active learning approaches on imbalanced sentiment classification

### Evaluation metrics

The popular geometric mean  $G-mean = \sqrt{TP_{rate} \times TN_{rate}}$  is adopted, where  $TP_{rate}$  is the true positive rate (also called positive recall or sensitivity) and  $TN_{rate}$  is the true negative rate (also called negative recall or specificity) (Kubat and Matwin, 1997).

### 4.2 Experimental Results

For thorough comparison, various kinds of active learning approaches are implemented including:

- **Random:** randomly select the samples from the unlabeled data for manual annotation;
- **Margin-based:** iteratively select samples closest to the hyperplane provided by the SVM classifier, which is suggested by Ertekin et al. (2007a) and Ertekin et al. (2007b). One sample is selected in each iteration;
- **Uncertainty:** iteratively select samples using the uncertainty measurement according to the output of ME classifier. One sample is selected in each iteration;
- **Certainty:** iteratively select class-balanced samples using the certainty measurement according to the output of ME classifier. One positive and negative sample (the positive and negative label is provided by the ME classifier) are selected in each iteration;
- **Co-testing:** first get contention samples (i.e., unlabeled examples on which the member

classifiers predict different labels) and then select the least confidence one among the hypotheses of different member classifiers, i.e., the aggressive strategy as described Muslea et al. (2006). Specifically, the member classifiers are two subspace classifiers trained by splitting the whole feature space into two disjoint subspaces of same size;

- **Self-selecting:** first select  $k$  uncertainty samples and then randomly select a positive and negative sample from the uncertainty-sample set, which is suggested by Doyle et al. (2011). We call it self-selecting since only one classifier is involved to measure uncertainty and predict class labels.

For those approaches involving random selection of features, we run 5 times for them and report the average results. Note that the samples selected by these approaches are imbalanced. To address the problem of classification on imbalanced data, we adopt the under-sampling strategy which has been shown effective for supervised imbalanced sentiment classification (Li et al., 2011a). Our active learning approach includes two versions: the co-selecting algorithm as described in Section 3.2 and the co-selecting with selected *MA* samples automatically labeled as described in Section 3.3. For clarity, we refer the former as **co-selecting-basic** and the latter as **co-selecting-plus** in the following.

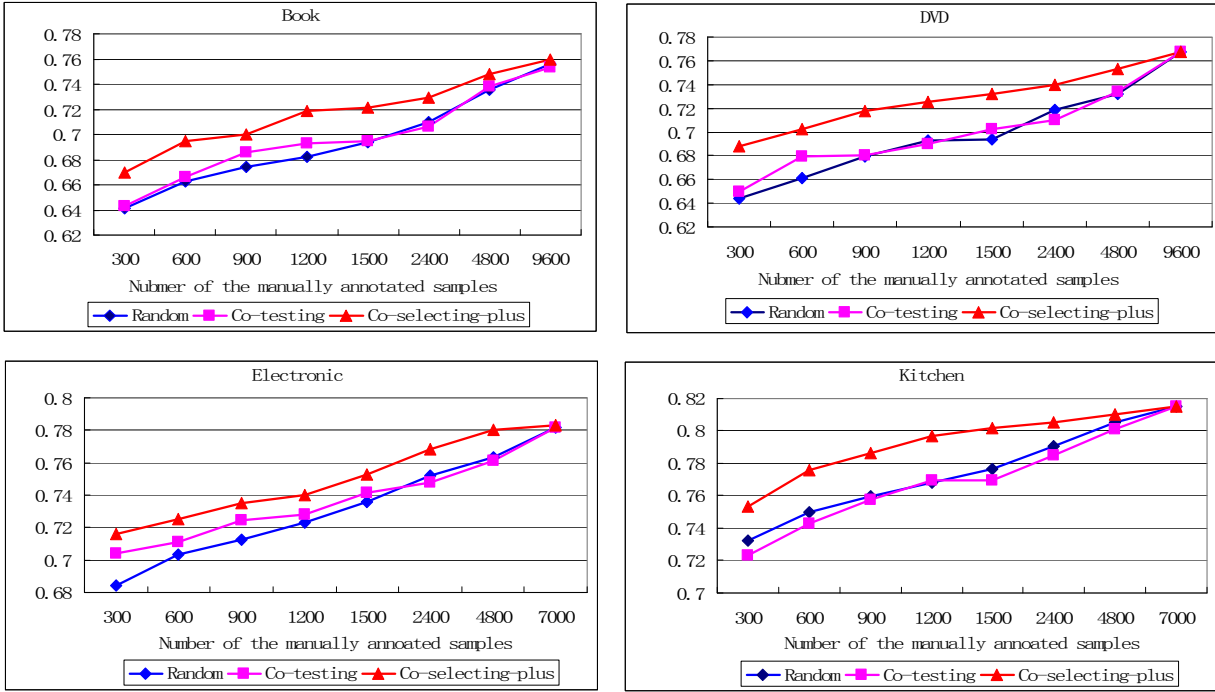


Figure 5: Performance comparison of three active learning approaches: random selection, co-testing and co-selecting-plus, by varying the number of the selected samples for manually annotation

### Comparison with other active learning approaches

Figure 4 compares different active learning approaches to imbalanced sentiment classification when 600 unlabeled samples are selected for annotation. Specifically, the parameters  $\theta$  and  $k$  is set to be 1/16 and 50 respectively. Figure 4 justifies that it is challenging to perform active learning in imbalanced sentiment classification: the approaches of **margin-based**, **uncertainty-based** and **self-selecting** perform no better than random selection while **co-testing** only outperforms random selection in two domains: DVD and Electronic with only a small improvement (about 1%). In comparison, our approaches, both **co-selecting-basic** and **co-selecting-plus** significantly outperform the random selection approach on all the four domains. It also shows that **co-selecting-plus** is preferable over **co-selecting-basic**. This verifies the effectiveness of automatically labeling those selected *MA* samples in imbalanced sentiment classification.

Specifically, we notice that only using the certainty measurement (i.e., **certainty**) performs worst, which reflects that only considering sample

balance factor in imbalanced sentiment classification is not helpful.

Figure 5 compares our approach to other active learning approaches by varying the number of the selected samples for manually annotation. For clarity, we only include random selection and **co-testing** in comparison and do not show the performances of the other active learning approaches due to their similar behavior to random selection. From this figure, we can see that **co-testing** is effective on Book and Electronic when less than 1500 samples are selected for manual annotation but it fails to outperform random selection in the other two domains. In contrast, our **co-selecting-plus** approach is apparently more advantageous and significantly outperforms random selection across all domains ( $p$ -value<0.05) when less than 4800 samples are selected for manual annotation.

### Sensitiveness of the parameters $\theta$

The size of the feature subspace is an important parameter in our approach. Figure 6 shows the performance of **co-selecting-plus** with varying sizes of the feature subspaces for the first subspace

classifier  $C_{Cer}$ . From Figure 6, we can see that a choice of the proportion  $\theta$  between 1/8 and 1/32 is recommended. This result also shows that the size of the feature subspace for selecting certain samples should be much less than that for selecting uncertain samples, which indicates the more important role of the uncertainty measurement in active learning.

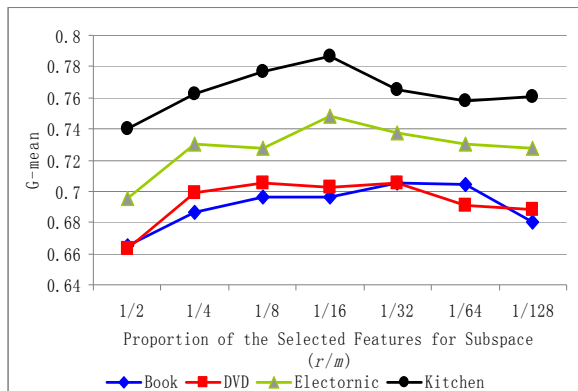


Figure 6: Performance of **co-selecting-plus** over varying sizes of feature subspaces ( $\theta$ )

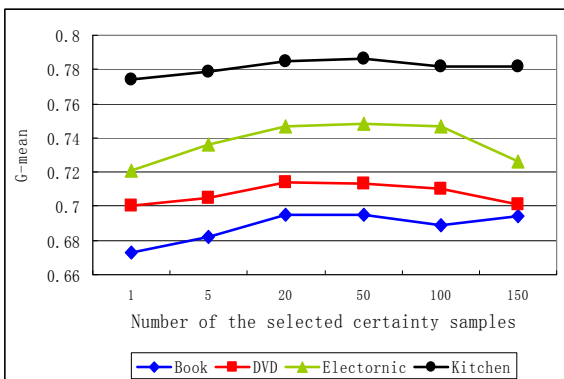


Figure 7: Performance of **co-selecting-plus** over varying numbers of the selected certain samples ( $k$ )

### Sensitiveness of parameter $k$

Figure 7 presents the performance of **co-selecting-plus** with different numbers of the selected certain samples in each iteration, i.e., parameter  $k$ . Empirical studies suggest that setting  $k$  between 20 and 100 could get a stable performance. Also, this figure demonstrates that using certainty as the only query strategy is much less effective (see the result when  $k=1$ ). This once again verifies the importance of the uncertainty strategy in active learning.

### Number of $MI$ samples selected for manual annotation

In Table 1, we investigate the number of the  $MI$  samples selected for manual annotation using different active learning approaches when a total of 600 unlabeled samples are selected for annotation. From this table, we can see that almost all the existing active learning approaches can only select a small amount of  $MI$  samples, taking similar imbalanced ratios as the whole unlabeled data. Although the **certainty** approach could select many  $MI$  samples for annotation, this approach performs worst due to its totally ignoring the uncertainty factor. When our approach is applied, especially **co-selecting-plus**, more  $MI$  samples are selected for manual annotation and finally included to learn the models. This greatly improves the effectiveness of our active learning approach.

Table 1: The number of  $MI$  samples selected for manual annotation when 600 samples are annotated on the whole.

	Book	DVD	Electronic	Kitchen
Random	71	82	131	123
SVM-based	65	72	135	106
Uncertainty	78	93	137	136
Certainty	160	200	236	227
Co-testing	89	84	136	109
Self-selecting	87	95	141	126
Co-selecting-basic	101	112	179	174
Co-selecting-plus	161	156	250	272

### Precision of automatically labeled $MA$ samples

In **co-selecting-plus**, all the added  $MA$  samples are automatically labeled by the first subspace classifier. It is encouraging to observe that 92.5%, 91.25%, 92%, and 93.5% of automatically labeled  $MA$  samples are correctly annotated in Book, DVD, Electronic, and Kitchen respectively. This suggests that the subspace classifiers are able to predict the  $MA$  samples with a high precision. This indicates the rationality of automatically annotating  $MA$  samples.

## 5 Conclusion

In this paper, we propose a novel active learning approach, named **co-selecting**, to reduce the annotation cost for imbalanced sentiment classification. It first trains two complementary



classifiers with two disjoint feature subspaces and then uses them to collectively select most informative *MI* samples for manual annotation, leaving most informative *MA* samples for automatic annotation. Empirical studies show that our co-selecting approach is capable of greatly reducing the annotation cost and in the meanwhile, significantly outperforms several active learning alternatives

For the future work, we are interested in applying our co-selecting approach to active learning for other imbalanced classification tasks, especially those with much higher imbalanced ratio.

## Acknowledgments

The research work described in this paper has been partially supported by three NSFC grants, No.61003155, No.60873150 and No.90920004, one National High-tech Research and Development Program of China No.2012AA011102, Open Projects Program of National Laboratory of Pattern Recognition, and the NSF grant of Zhejiang Province No.Z1110551. We also thank the three anonymous reviewers for their helpful comments.

## References

- Attenberg J. and F. Provost. 2010. Why Label when you can Search? Alternatives to Active Learning for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance. In *Proceeding of KDD-10*, 423-432.
- Blitzer J., M. Dredze and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of ACL-07*, 440-447.
- Cui H., V. Mittal, and M. Datar. 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proceedings of AAAI-06*, pp.1265-1270.
- Doyle S., J. Monaco, M. Feldman, J. Tomaszewski and A. Madabhushi. 2011. An Active Learning based Classification Strategy for the Minority Class Problem: Application to Histopathology Annotation. *BMC Bioinformatics*, 12: 424, 1471-2105.
- Ertekin S., J. Huang, L. Bottou and C. Giles. 2007a. Learning on the Border: Active Learning in Imbalanced Data Classification. In *Proceedings of CIKM-07*, 127-136.
- Ertekin S., J. Huang, L. Bottou and C. Giles. 2007b. Active Learning in Class Imbalanced Problem. In *Proceedings of SIGIR-07*, 823-824.
- Freund Y., H. Seung, E. Shamir and N. Tishby. 1997. Selective Sampling using the Query by Committee algorithm. *Machine Learning*, 28(2-3), 133-168.
- He Y., C. Lin and H. Alani. 2011. Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification. In *Proceeding of ACL-11*, 123-131.
- Lewis D. and W. Gale. 1994. Training Text Classifiers by Uncertainty Sampling. In *Proceedings of SIGIR-94*, 3-12.
- Li F., Y. Tang, M. Huang and X. Zhu. 2009. Answering Opinion Questions with Random Walks on Graphs. In *Proceedings of ACL-IJCNLP-09*, 737-745.
- Li S. and C. Zong. 2008. Multi-domain Sentiment Classification. In *Proceedings of ACL-08*, short paper, pp.257-260.
- Li S., C. Huang, G. Zhou and S. Lee. 2010. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In *Proceedings of ACL-10*, pp.414-423.
- Li S., Z. Wang, G. Zhou and S. Lee. 2011a. Semi-supervised Learning for Imbalanced Sentiment Classification. In *Proceeding of IJCAI-11*, 826-1831.
- Li S., G. Zhou, Z. Wang, S. Lee and R. Wang. 2011b. Imbalanced Sentiment Classification. In *Proceedings of CIKM-11*, poster paper, 2469-2472.
- Lloret E., A. Balahur, M. Palomar, and A. Montoyo. 2009. Towards Building a Competitive Opinion Summarization System. In *Proceedings of NAACL-09 Student Research Workshop and Doctoral Consortium*, 72-77.
- Kubat M. and S. Matwin. 1997. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of ICML-97*, 179-186.
- Muslea I., S. Minton and C. Knoblock. 2006. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research*, vol.27, 203-233.
- Pang B. and L. Lee. 2008. Opinion Mining and Sentiment Analysis: Foundations and Trends. *Information Retrieval*, vol.2(12), 1-135.
- Pang B., L. Lee and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP-02*, 79-86.

- Settles B. 2009. Active Learning Literature Survey. *Computer Sciences Technical Report 1648*, University of Wisconsin, Madison, 2009.
- Turney P. 2002. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In *Proceedings of ACL-02*, 417-424.
- Wan X. 2009. Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of ACL-IJCNLP-09*, 235-243.
- Yang Y. and G. Ma. 2010. Ensemble-based Active Learning for Class Imbalance Problem. *J. Biomedical Science and Engineering*, vol.3,1021-1028.
- Zhang M. and X. Ye. 2008. A Generation Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval. In *Proceedings of SIGIR-08*, 411-418.
- Zhu J. and E. Hovy. 2007. Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. In *Proceedings of ACL-07*, 783-793.