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 θ) 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 [3] tool is mainly adopted, except that in the margin-based active learning approach (Ertekin et al., 2007a) where SVM is implemented with light-SVM [4]. The features for classification are unigram words with Boolean weights.

[3] http://mallet.cs.umass.edu/

[4] http://www.cs.cornell.edu/people/tj/svm_light/