classifier C_{Cer}. From Figure 6, we can see that a choice of the proportion θ 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.

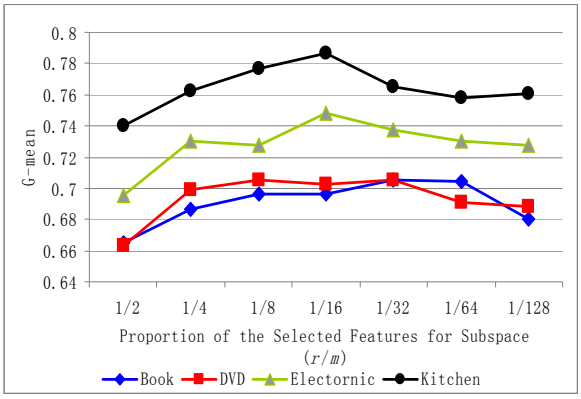

Figure 6: Performance of **co-selecting-plus** over varying sizes of feature subspaces (θ)

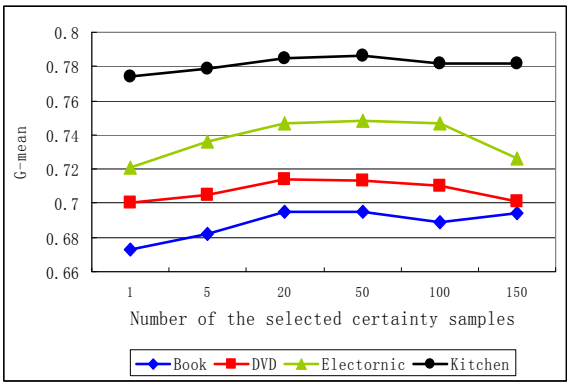

Figure 7: Performance of **co-selecting-plus** over varying numbers of the selected certain samples (*k*)

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