denotes the value of calibration parameter α) is used. In Table 1, we explicitly compare the best results obtained by different methods, where “All” denotes the results using all training samples. The K-L divergence (KLD) between the training and test data is also reported for each task. The results of PUIW are in terms of that obtained by WNB, and the results of PUIS are in terms of that obtained by NB trained with selected samples.

It should be noticed that we have not compared our model with traditional labeling adaptation approaches (such as SCL) due to different task settings. The focus of this paper is instance adaptation for sentiment classification. Our method do not rely on any additional target domain data, while the labeling adaptation methods need either a small amount of labeled data or a large amount of unlabeled data in the target domain for help.

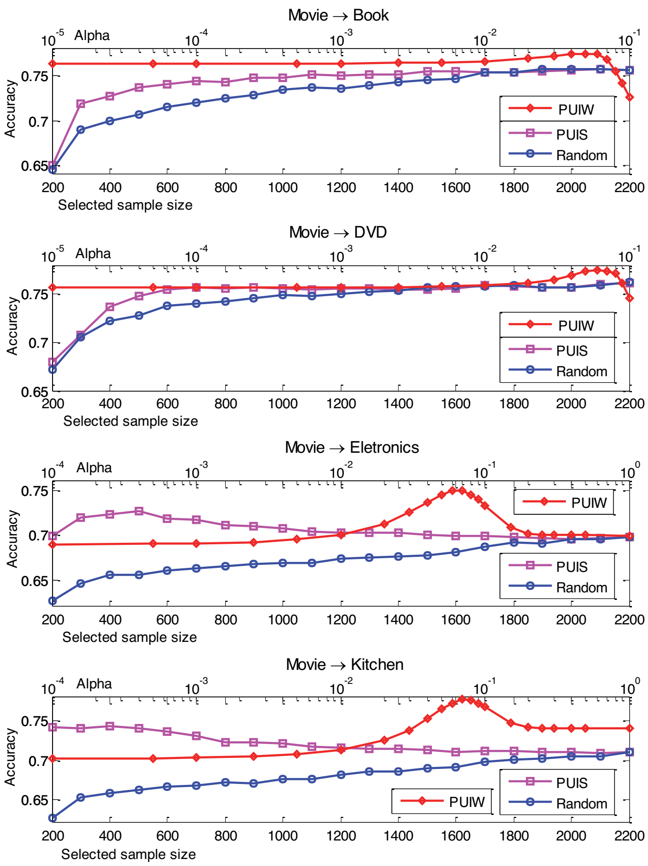

Figure 3: Accuracy curve using a normal-size training set. For Random and PUIS, the bottom x-axis (number of selected samples) is used. For PUIW, the top x-axis is used (the value of α).

Task	KLD	ALL	PUIS	PUIW
Movie → Book	43.81	0.7568	0.7572	**0.7747**
Movie → DVD	33.54	0.7622	0.7622	**0.7818**
Movie → Electronics	104.52	0.6975	0.7265	**0.7500**
Movie → Kitchen	119.70	0.7097	0.7435	**0.7775**
Average	—	0.7316	0.7474	**0.7710**

Table 1: Accuracy comparison using a normal-size training set.

We observe the results in the following aspects:

1) Random Selection. In all tasks, with the increase of training samples, the performance of Random Selection is gradually improved. This agrees with our general knowledge that more training data will improve the machine learning performance.

2) PUIS. Using the same number of training data, PUIS is significantly better than Random Selection. It indicates that PU learning is effective at selecting the most useful training samples. The accuracy increase is significant, especially in high-KLD tasks. For example, in Movie → {Electronics and Kitchen}, the improvements are 2.90 and 3.38 percentages, respectively. When the KLD is relatively low, however, the effect of PUIS are limited. The average increase is 1.58 percentages.

3) PUIW. Compared to PUIS, the effect of PUIW is more significant and robust. Across four tasks, the improvements are 1.79, 1.96, 5.25 and 6.78 (in average 3.94) percentages, respectively. It is reasonable since the calibrated in-target-domain probabilities used as sampling weight are more appropriate than the usage of 0/1 weight in instance selection.

4.3 Experiments with A Large-size Training Set

In this section, we report the experimental results using a large-size training set, which contains 10,000 labeled Video reviews. The reviews from each of the other 12 domains are used as test data respectively. In Figure 4, we draw the accuracy curve of Random Selection, PUIS and PUIW. Due to the space limitation, we only present four of them. In Table 2, we report the best results obtained by different methods as well as the KLD across all 12 tasks.

We observe the results the same way as in Section 4.2. But this time we lay the emphasis on the results that are special for large-size training data

1) The effect of adding more training samples. It can be observed in Random Selection that, when the size of training samples is relatively small (<2000), the accuracy increases significantly as the training size increases. But when the size becomes larger (>3000), the improvement by adding more training data is very slight. It indicates that when the size of training data is already large, adding more training samples will not cause significant improvements.

2) The necessity of PUIS/PUIW. Compared to the results in Section 4.2, the effects of PUIS and PUIW in this case are more significant. The average of increase of PUIS and PUIW are 3.32 and 4.48 percentages, respectively. It is worth noting that in many tasks, only about 10% of selected training samples could result in better performance than that trained with all samples. It indicates that instance adaptation is very necessary when the size of training data is large.

3) The stability of PUIW. In addition to its remarkable performance, it also can be observed that the accuracy curve of PUIW is unimodal and quite stable. While it is not easy to determine the best number of selected instances in PUIS, most of the best accuracy in PUIW are obtained when α locates at the area around 0.1. It suggests another advantage of PUIW: the stability in model selection.