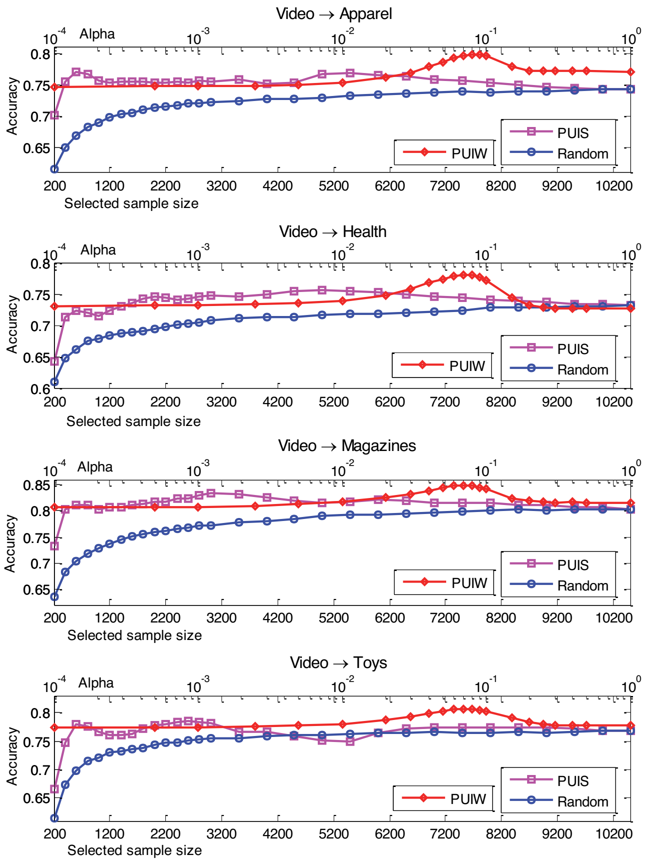

Figure 4: Accuracy curve using a large-size training set. For Random and PUIS, the bottom x-axis (number of selected samples) is used. For PUIW, the top x-axis is used (the value of α).

Task	KLD	ALL	PUIS	PUIW
Video → Apparel	166.69	0.7440	0.7704	**0.7995**
Video → Baby	160.50	0.7494	0.7759	**0.7932**
Video → Books	85.61	0.7328	**0.7952**	0.7893
Video → Camera	146.61	0.7747	0.8164	**0.8278**
Video → DVD	66.71	0.7877	0.8169	**0.8180**
Video → Electronics	143.87	0.7213	0.7603	**0.7712**
Video → Health	159.73	0.7331	0.7576	**0.7826**
Video → Kitchen	155.72	0.7424	0.7736	**0.7980**
Video → Magazines	122.53	0.8030	0.8344	**0.8484**
Video → Music	99.49	0.7562	0.7581	**0.7734**
Video → Software	136.48	0.7411	**0.8078**	0.7830
Video → Toys	134.84	0.7679	0.7858	**0.8066**
Average	—	0.7545	0.7877	**0.7993**

Table 2: Accuarcy comparison using a large-size traing set.

4.3 Further Discussion

We finally investigate the relation between K-L divergence (KLD) and the accuracy improvements of our approach. It is known that KLD measures the difference of two distributions. In our tasks, KLD represents the distributional change from the training set to test set. Hence, when KLD is small, the space of improvements in domain adaptation is limited; when KLD increases, the space of improvements also becomes larger.

In Figure 5, we draw the relation of KLD and the accuracy increase gained by PUIW. It can be observed that, KLD and accuracy increase are in a linear relation generally, except for few aberrant points. It indicates that our approach has a good property: the larger the KLD of the training and test data is, the more effective our approach will be.

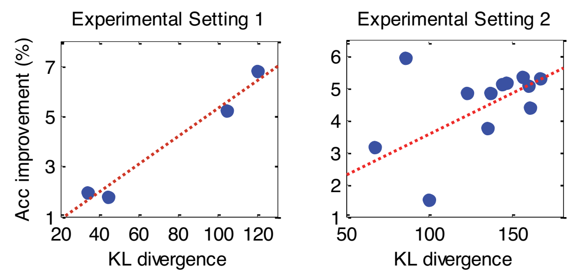

Figure 5: The relation between KLD and Accuracy Increase across all of the tasks (Each point represents one task).

5 Conclusions

In this paper, we propose a novel approach for cross-domain sentiment classification, based on instance selection and instance weighting via PU learning. PU learning is first used to learn an in-target-domain selector, and assign an in-target-domain probability to each sample in the training set. Based on the in-target-domain probabilities, two models namely PUIS and PUIW, are developed. The experimental results prove the necessity and effectiveness of the approach, especially when the size of training data is large. The results also indicate another good property of our approach: the larger the K-L divergence between the training and test data is, the more effective our approach will be.

Shortcomings of this work contain two aspects: 1) Explicit model selection, such as the determination of the number of selected samples and the value of the calibration parameter α, are not involved; 2) It lacks the consideration for labeling adaptation (we simply assume $p_s(y|\mathbf{x}) \approx p_t(y|\mathbf{x})$ in Section 3.3) in instance adaptation. Both of them are very important issues, and we will perform some related investigation in our future work.

Acknowledgments

The research work is supported by the Jiangsu Provincial Natural Science Foundation of China under Grant No. BK2012396, the Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20123219120025, the Open Project Program of the National Laboratory of Pattern Recognition (NLPR). The work is also partially supported by the Hi-Tech Research and Development Program (863 Program) of China under Grant No. 2012AA011102 and 2012AA011101, and the National Science Fund for Distinguished Young Scholars under Grant No. 61125305.