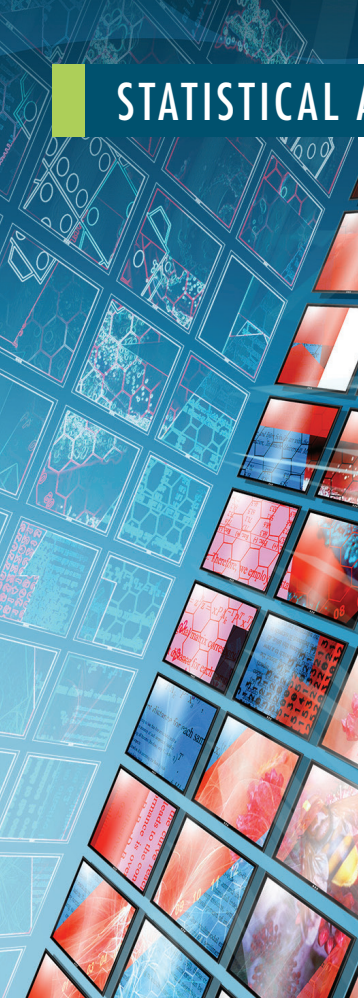

Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification

Rui Xia, *Nanjing University of Science and Technology*

Chengqing Zong, *Chinese Academy of Sciences*

Xuelei Hu, *Nanjing University of Science and Technology*

Erik Cambria, *National University of Singapore*

Domain adaptation problems arise often in sentiment classification. Here, the feature ensemble plus sample selection (SS-FE) approach offers better results by taking into account labeling adaptation and instance adaptation together.

The problem of domain adaptation has attracted more and more attention in the fields of both machine learning and natural language processing (NLP). Domain adaptation arises often in sentiment classification. For example, we might want a book review classifier, but the book-labeled reviews are scarce and the labeled reviews in the movie domain are abundant. Therefore, we need to adapt a classifier trained by labeled reviews in the movie domain to the book domain. According to the analysis carried out by Jing Jiang and ChengXiang Zhai,[1] there are two distinct needs in domain adaptation: *labeling adaptation* and *instance adaptation*. We interpret these two needs as follows:

- *labeling adaptation* models the changes of labeling function because one feature that's positive in the source domain might express the opposite meaning in the target domain; and
- *instance adaptation* models the change of instance probability, such as the change of vocabulary or word frequency from one domain to another.

To take full account of these two attributes, we propose a comprehensive method, called *feature ensemble plus sample selection* (SS-FE), for domain adaptation in sentiment classification. This approach could yield significant improvements compared to individual feature ensemble (FE) or sample selection (SS) methods, because it comprehensively considers both labeling adaptation and instance adaptation. We discuss others'

1541-1672/13/$31.00 © 2013 IEEE
Published by the IEEE Computer Society

efforts in domain adaptation in the related sidebar.

Method Overview

In formulating our SS-FE method, we first propose a labeling adaptation method based on feature ensemble (FE). This idea is based on the observation that, features with different type of part-of-speech (POS) tags have a distinct change in distribution in domain adaptation. We term the heavily changing features *domain-specific*, and the slightly changing features *domain-independent*. The domain-independent features generally perform consistently when the domain changes. For example, some adjectives and adverbs, such as "great" and "like," always have a strong correlation with the positive class label, regardless of domain. However, the domain-specific features indicate different sentiment in different domains.[2] The adjective "small," for example, is usually appraised as negative for a hotel review ("small room") but positive for a post office ("small queue"), or the concept "go read the book" most likely indicates positive sentiment for a book review but negative for a movie review.

In FE, we first train individual classifiers with different feature sets divided by their POS tags. The final model is a weighted ensemble of individual classifiers, in which the weights are turned based on a small amount of labeled data in the target domain, with the goal of increasing the weight of domain-free features and reducing the weight of domain-specific features in domain adaptation.

Second, we developed a sample selection method based on principal component analysis (PCA-SS) as an aid to FE, because FE models only labeling adaptation and neglects instance adaptation. To address this problem, we combine PCA-SS and FE, and refer to the joint approach as SS-FE. In SS-FE, we first employ PCA-SS to select a subset of source domain samples that are close to the target domain, and then use the selected samples for labeling adaptation. A singular value decomposition (SVD) algorithm is conducted on the target domain dataset to extract the latent concepts representing the target domain. Samples in the source domain are then projected onto the latent space. Those samples that are far away from the normal area are discarded, and the remaining samples with a close distance are used as training data in domain adaptation.

We empirically show that both FE and PCA-SS are effective for cross-domain sentiment classification, and that SS-FE performs better than either approach because it comprehensively considers both labeling and instance adaptation.

Feature Ensemble

POS tags are supposed to be significant indicators of sentiment. Previous work revealed a high correlation between the presence of adjectives and document sentiment; certain verbs and nouns are also strong indicators of sentiment.[3] For cross-domain sentiment classification, we observe that features with different types of POS tags might have different levels of distributional change in domain adaptation. For example, nouns change the most because domains are mostly denoted by nouns, while adjectives and adverbs are fairly consistent across domains. The cross-domain Kullback-Leibler (K-L) distance regarding different POS tags (reported in the "Discussion" section) confirms our observation.

Based on this observation, we divide features into four groups: adjectives and adverbs (J), verbs (V), nouns (N), and the others (O). Base classifiers will be trained on all four feature subsets. Thus, a feature vector $\mathbf{x}$ is made up of four parts: $\mathbf{x} = [\mathbf{x}_J, \mathbf{x}_V, \mathbf{x}_N, \mathbf{x}_O]$, and we use $g_k(\mathbf{x}_k)$ to denote the labeling function of each base-classifier. In the case of linear classifier, $g_k(\mathbf{x}_k) = w_k^T \mathbf{x}_k$, $k \in \{J, V, N, O\}$.

After base classification, we use the Stacking algorithm for meta-learning, where we construct the meta-learning feature vector $\tilde{\mathbf{x}} = [g_J, g_V, g_N, g_O]$ on a small amount of labeled data from the target domain (the *validation set*), and the weights of each base classifier θ_k are optimized by minimizing the perceptron loss function so that each component's final weights are tuned to adapt to the target domain. We represent the weighted ensemble as

$$f(\mathbf{x}) = \sum_k \theta_k g_k(\mathbf{x}_k). \quad (1)$$

In the meta-learning process, we assign larger weights to the domain-independent parts, such as adjectives and adverbs, and lower weights to the domain-specific parts, such as nouns. Figure 1 shows the FE algorithm's pseudocode.

Sample Selection

The FE approach adapts the labeling function $p_s(y|\mathbf{x}) \rightarrow p_t(y|\mathbf{x})$ (in our approach, $g \rightarrow f$). However, the labeling adaptation is based on the condition that the instance probability $p_s(\mathbf{x})$ and $p_t(\mathbf{x})$ are the same. If there's a big gap between them, the effect of labeling adaptation will be reduced. To address this issue, we propose PCA-SS as an aid to FE. PCA-SS first selects a subset of the source domain labeled data whose instance distribution is close to the target domain, and then uses these selected samples as training data in labeling adaptation.

For PCA-based sample selection, let X_t denote the document-by-feature matrix of the target domain dataset;

Related Work in Domain Adaptation

Researchers have taken a variety of approaches to improve domain adaptation. For example, Jiong Jiang and ChengXiang Zhai[1] proposed to address domain adaptation from two perspectives: labeling adaptation and instance adaptation. Sinno Jialin Pan and Qiang Yang[2] presented a comprehensive survey of transfer learning, which categorizes transfer learning approaches in a similar manner: feature-based transfer, instance-based transfer, and model-parameter-based transfer.

Existing works for domain adaptation in sentiment classification mostly use labeling adaptation. These methods aim to learn a new feature representation (or a new labeling function) for the target domain, using source domain-labeled data, with the help of some labeled (or unlabeled) target domain data.[3–11] Among them, the structural correspondence learning (SCL) algorithm[3] is the most representative. Hal Daumé introduced a simple method for domain adaptation based on feature space augmentation.[4] Sinno Jialin Pan and his colleagues[5] proposed two domain adaptation approaches via transfer component analysis (TCA) and spectral feature alignment (SFA),[6] respectively. Lixin Duan and his colleagues[7] proposed augmented feature representations for heterogeneous domain adaptation. Simultaneously, two of the authors of this article (Rui Xia and Chengqing Zong)[8] and Rajhans Samdani and Wen-Tau Yih[9] proposed feature reweighting based on an ensemble of feature sets, although they divide the feature sets in different ways. Xavier Glorot and his colleagues[10] proposed a deep learning approach, while Shou-Shan Li and Zong[11] proposed to use the ensemble technique to address multiple domain adaptation for sentiment classification.

Instance adaptation learns the importance of labeled data in the source domain by instance reweighting. These reweighed instances are then used for learning in the target domain. This problem is also known as *sample selection bias* in machine learning scenarios. Masashi Sugiyama and his colleagues[12] proposed the Kullback-Leibler importance estimation procedure (KLIEP) algorithm to address the density ratio estimation problem. However, it's only appropriate for the case of continuous distribution. To the best of our knowledge, research on instance adaptation in natural language processing (NLP) is pretty scarce. Recently, Amittai Axelrod and his colleagues[13] proposed a method called *pseudo in-domain data selection* for selecting source domain training samples based on language model. Erik Cambria his colleagues, instead, employed PCA[14] and linear discriminant analysis (LDA)[15] to build SenticNet (http://sentic.net), an affective commonsense knowledge base for open-domain sentiment analysis.

In contrast to these methods, which focus mainly on one type of adaptation, we propose a comprehensive model that simultaneously takes labeling adaptation and instance adaptation into consideration. We first employ principal component analysis sample selection (PCA-SS) to select a subset of source domain samples, and then use the selected samples as training data in feature ensemble (FE)-based labeling adaptation. It's an extended version of our previous work,[8] in which we conducted only FE-based transfer. Indeed, PCA has been widely used for dimension reduction and concept representation in feature-based domain adaptation. For example, SCL, TCA, and SFA all employed singular value decomposition (SVD) for building new feature representation. To our knowledge, this is the first time that PCA has been used for sample selection in domain adaptation.

References

1. J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," *Proc. Ann. Meeting of the Assoc. Computational Linguistics*, ACL, 2007, pp. 264–271.
2. S.J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowledge and Data Engineering,* vol. 22, no. 10, 2010, pp. 1345–1359.
3. J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-Boxes, and Blenders: Domain Adaptation for Sentiment Classification," *Proc. Annual Meeting of the Assoc. for Computational Linguistics*, ACL, 2007, pp. 440–447.
4. H. Daume, III, "Frustratingly Easy Domain Adaptation," *Proc. Ann. Meeting of the Assoc. for Computational Linguistics*, ACL, 2007, pp. 256–263.
5. S.J. Pan et al., "Domain Adaptation via Transfer Component Analysis," *Proc. Int'l Joint Conf. Artificial Intelligence*, Morgan Kaufmann, 2009, pp. 1187–1192.
6. S.J. Pan et al., "Cross-Domain Sentiment Classification via Spectral Feature Alignment," *Proc. Int'l World Wide Web Conference*, ACM, 2010, pp. 751–760.
7. L. Duan, D. Xu, and I.W. Tsang, "Learning with Augmented Features for Heterogeneous Domain Adaptation," *Proc. Int'l Conf. Machine Learning*, 2012; http://icml.cc/2012/papers/373.pdf.
8 R. Xia and C. Zong, "A POS-Based Ensemble Model for Cross-Domain Sentiment Classification," *Proc. Int'l Joint Conf. Natural Language Processing*, Asian Federation of Natural Language Processing, 2011, pp. 614–622.
9. R. Samdani and W. Yih, "Domain Adaptation with Ensemble of Feature Groups," *Proc. Int'l Joint Conf. Artificial Intelligence* (IJCAI), IJCAI, 2011, pp. 1458–1464.
10. X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," *Proc. Int'l Conf. Machine Learning*, Int'l Machine Learning Soc., 2011, pp. 513–520.
11. S. Li and C. Zong, "Multi-Domain Adaptation for Sentiment Classification: Using Multiple Classifier Combining Methods," *Proc. IEEE Conf. Natural Language Processing and Knowledge Eng.*, IEEE, 2008, pp. 458–465.
12. M. Sugiyama et al., "Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation," *Annals of the Inst. of Statistical Mathematics*, vol. 60, no. 4, 2008, pp. 699–746.
13. A. Axelrod, X. He, and J. Gao, "Domain Adaptation via Pseudo In-Domain Data Selection," *Proc. Conf. Empirical Methods in Natural Language Processing*, ACL, 2011, pp. 355–362.
14. E. Cambria, D. Olsher, and K. Kwok, "Sentic Activation: A Two-Level Affective Common Sense Reasoning Framework," *Proc. Conf. Artificial Intelligence*, Assoc. for the Advancement of Artificial Intelligence, 2012, pp. 186–192.
15. E. Cambria, T. Mazzocco, and A. Hussain, "Application of Multi-Dimensional Scaling and Artificial Neural Networks for Biologically Inspired Opinion Mining," *Biologically Inspired Cognitive Architectures*, vol. 4, 2013, pp. 41–53.

we then get the target domain's latent concepts by solving the following SVD problem:

$$\mathbf{X}_t = U\Sigma V^T,$$

where U and V are unitary unit orthogonal matrices and Σ contains the positive real singular values of decreasing magnitude $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_M$. The latent concepts are the orthogonal column vectors in the matrix V, and the variance of the data projected along the ith column of V is equal to σ_i^2.

To optimally capture the data's variations, only those latent concepts corresponding to the $k < M$ largest singular values are typically retained. By selecting the columns of $P = V[:, 0 : k] \in R^{M \times k}$, which correspond to the latent concepts associated with the first k singular values as the projection matrix, we obtain the document-by-concept matrix $\hat{X}_t = X_t P$.

Once the principal component model is established, the projection of an arbitrary sample onto the concept space is given by $\hat{x}^T = Px^T$. Note that Hotelling's T^2 statistic reflects the trend and magnitude deviation degree of each sample in a PCA model. In the settings of domain adaptation, it measures the extent to which a sample deviates from the concept space. Therefore, as the criterion for sample selection, we use T^2 statistic, also known as the "concept distance" measure:

$$D(\mathbf{x}) = T^2(\mathbf{x}) = \mathbf{z}^T\mathbf{z} = \mathbf{x}P\Lambda_k^{-1}P^T\mathbf{x}^T,$$

where $\mathbf{z} = \Lambda_k^{-\frac{1}{2}}P^T\mathbf{x}^T$, and Λ_k is the diagonal matrix corresponding to the top k singular values.

By this definition, we can obtain a concept distance for each sample $x_s^{(n)}$ in the source domain:

$$D(\mathbf{x}_s^{(n)}) = \mathbf{x}_s^{(n)}P\Lambda_k^{-1}P^T\mathbf{x}_s^{(n)^T}.$$

Input:
Source domain training set $\mathcal{D}_S = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N_s)}, y^{(N_s)})\}$, where the feature vector $\mathbf{x}^{(n)} = [\mathbf{x}_J^{(n)}, \mathbf{x}_V^{(n)}, \mathbf{x}_N^{(n)}, \mathbf{x}_O^{(n)}]$;
Target domain validation set $\mathcal{D}_t = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N_t)}, y^{(N_t)})\}$, where N_t is (5–10)% the size of N_s;
Target domain test set $\mathcal{D}_s = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(M)}, y^{(M)})\}$.
Process:
for each component $k \in \{J, V, N, O\}$: % base-classification
train a base-classifier $\mathcal{H}_k$: $g_k = \omega_k^T\mathbf{x}_k$
predict each sample in the validation set using $\mathcal{H}_k$
predict each sample in the test set using $\mathcal{H}_k$
end;
build meta-learning training samples on the validation set $\tilde{\mathcal{D}} = \{\tilde{\mathbf{x}}^{(1)}, \tilde{\mathbf{x}}^{(2)}, \ldots, \tilde{\mathbf{x}}^{(N_t)}\}$, where $\tilde{\mathbf{x}}^{(i)} = [g_J^{(n)}, g_V^{(n)}, g_N^{(n)}, g_O^{(n)}]$;
train a meta-learning classifier $f = \theta^T\tilde{\mathbf{x}}$; % meta-learning
Output:
predict samples in the test set using $f(\mathbf{x}) = \sum_k \theta_k g_k(\mathbf{x}_k)$.

Figure 1. Pseudocode of the feature ensemble (FE) algorithm.

Correspondingly, we get a set of the concept distances for each sample $\mathbf{x}_t^{(n)}$ in the target domain:

$$D(\mathbf{x}_t^{(n)}) = \mathbf{x}_s^{(n)}P\Lambda_k^{-1}P^T\mathbf{x}_t^{(n)^T}.$$

Finally, we define the sample selection threshold as:

$$\bar{D} = \max_{\mathbf{x}_t^{(n)} \in \mathcal{D}_t} \{D(\mathbf{x}_t^{(n)})\}. \quad (2)$$

Samples in the source domain $\mathbf{x}_s^{(n)}$ with $D(\mathbf{x}_s^{(n)}) > \bar{D}$ are discarded, and those with lower concept distance than the threshold are selected as training samples in domain adaptation.

Experiments

Now that we've detailed our method's inner workings, we present our experiments and results

Dataset and Experimental Setup

We use the popular Multi-Domain Sentiment Dataset[4] for experiments. It consists of product reviews collected from four different domains of Amazon.com—book (B), DVD (D), electronics (E), and kitchen (K). Each domain contains 1,000 positive and 1,000 negative reviews. The term "source → target" is used to denote different cross-domain tasks. For example, "D → B" represents the task that's trained in the DVD domain but tested in the book domain. Each domain acts as the source domain and target domain, respectively, and this generates 12 total tasks.

In each of the tasks, labeled instances in the source domain are used for training base-classifiers. Labeled instances in the target domain are evenly split into 10 folds, where one fold is used as a validation set for meta-learning, and the other nine folds are used as a test set. All of the following experimental results are reported in terms of an average of the 10 folds' cross validation.

Experimental Results of PCA-SS

Here, we present the experimental results of sample selection. The PCA model is established by the feature-by-document data matrix of the target validation set. The number of principal components is chosen with the percentage of variance contribution larger than 99.5 percent. All samples in the source and target domain are projected onto the concept space. We sort all source domain samples in an ascending order,

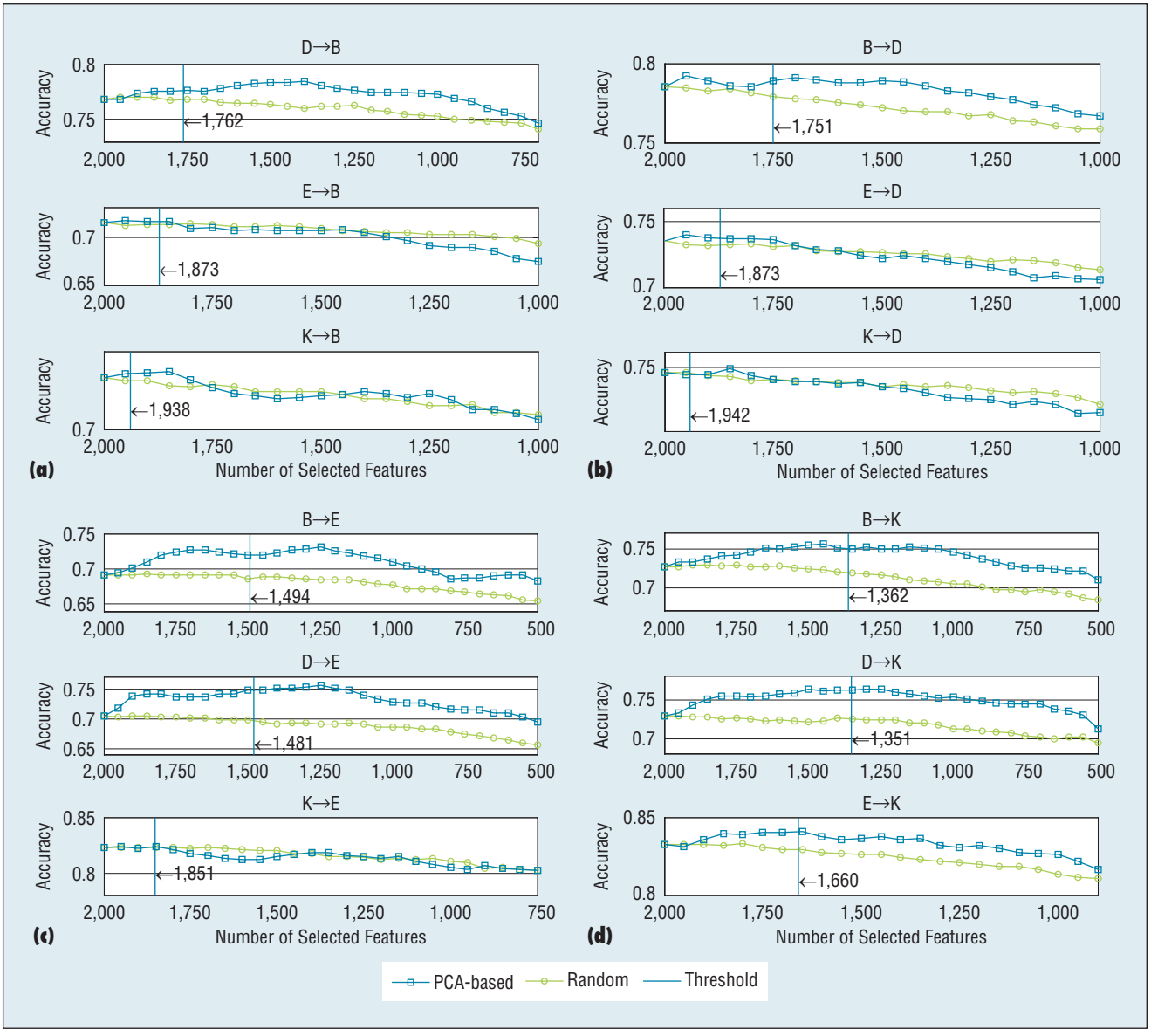

Figure 2. The performance of principal component analysis-based sample selection (PCA-SS).

according to their concept distance, and present the classification accuracy trained with a decreasing number of selected features in Figure 2. We chose Naïve Bayes (NB) as the base classification algorithm, as it reportedly performs the best among three classifiers (NB, MaxEnt, and SVMs) in the multi-domain dataset.[5] For comparison, we also report the performance of the same number of random selected samples. We observe the experimental results from the following three perspectives.

Random selection. We first observe the result of randomly selected samples. In all tasks, performance is gradually decreased by reducing the number of selected samples. This is in accordance with our standard understanding that decreasing training samples hurts machine-learning performance.

PCA-based sample selection. With respect to PCA-SS, the conclusion is different. In most of the tasks, the performance increases when the number of selected samples decreases at the first stage. As the selected number is reduced to a certain extent, the curve reaches the peak value. After that, the performance gradually decreases. This phenomenon leads to the conclusion that using a selected subset of samples as training data improves classification performance over training on all samples. This conclusion holds well in the tasks of D → B, B → D, B → E, D → E, B → K, D → K, and E → K. It's also worth noticing that sample selection isn't so effective in some tasks (such as E → B, K → D, and K → E). We'll discuss why later.

Table 1. Cross-domain classification accuracy.*

Tasks	Base-J	Base-V	Base-N	Base-O	Baseline	FE	PCA-SS	SS-FE
D → B	0.7589	0.6682	0.6307	0.6432	0.7685	0.7987	0.7759	0.8038
E → B	0.6907	0.6015	0.6038	0.5938	0.7156	0.7163	0.7162	0.7286
K → B	0.7034	0.5997	0.5943	0.5974	0.7265	0.7334	0.7285	0.7294
B → D	0.7665	0.6871	0.6597	0.6348	0.7854	0.7874	0.7889	0.7910
E → D	0.7228	0.5877	0.6142	0.5743	0.7350	0.7296	0.7373	0.7460
K → D	0.7179	0.5994	0.6142	0.5766	0.7469	0.7563	0.7458	0.7570
B → E	0.7319	0.6186	0.5649	0.5755	0.6915	0.7223	0.7199	0.7424
D → E	0.7431	0.6168	0.6005	0.5624	0.7040	0.7496	0.7477	0.7707
K → E	0.8051	0.7185	0.6961	0.6076	0.8235	0.8226	0.8243	0.8293
B → K	0.7568	0.6328	0.5936	0.5916	0.7265	0.7412	0.7507	0.7807
D → K	0.7334	0.6228	0.5987	0.6025	0.7299	0.7644	0.7630	0.7782
E → K	0.8083	0.7285	0.7037	0.6237	0.8330	0.8324	0.8412	0.8487

* Features are divided into four groups: adjectives and adverbs (J), verbs (V), nouns (N), and the others (O). FE = feature ensemble, and PCA-SS = principal component analysis sample selection.

Concept distance threshold. Finally, we verify the validity of the concept distance measure for selecting samples. The size of selected samples provided by distance threshold is denoted by a red vertical line in each subfigure, and the golden size is denoted by a black vertical line. As we can see, although there's a bias between the number of samples given by the threshold and the golden size, the concept distance threshold still provides a reasonable value (at least a reasonable area) that's close to the golden one in classification performance.

Experimental Results of SS-FE

In this section, we present the results of SS-FE. Four base NB classifiers are trained on the four feature sets, where features with a term frequency of four or more are selected and the BOOL weight is adopted. Table 1 reports the classification accuracy of each component classifier (Base-J, Base-V, Base-N, and Base-O), and the baseline system using all features (Baseline), only FE, only PCA-SS, and SS-FE. We compared the following three aspects.

FE versus Baseline. We first compare base classifiers and Baseline. It's interesting that Base-J yields a comparative performance to Baseline. In some tasks, J is even better. With an efficient ensemble of all base classifiers, FE performs consistently better than each base classifier and the Baseline system.

PCA-SS versus Baseline. When comparing PCA-SS and Baseline, we reconfirm the conclusion from Figure 2 that, with a selected subset of training samples, the PCA-SS could improve significantly.

SS-FE, FE, and PCA-SS. To make the SS-FE comparison clearer, we plot the results of Baseline, FE, PCA-SS, and SS-FE in Figure 3. As the figure shows, SS-FE is consistently better than Baseline, FE, and PCA-SS, except for one task—K → B—where SS-FE is slightly weaker than FE. We summarize the results as follows: first, in the tasks, such as D → B and K → D, where FE is more effective but the PCA-SS is less effective (denoted as FE $\succ$ SS), the SS-FE improvements are generally gained by labeling adaptation. Second, in the tasks where the effects of PCA-SS are more significant than FE (FE $\prec$ SS), such as E → D and E → K, the SS-FE improvements are mainly from instance adaptation. Third, in the tasks where FE and PCA-SS are both effective, such as D → E and B → K, the improvements finally gained by SS-FE are remarkable.

Table 2. Average Kullback-Leibler (K-L) distance of different part-of-speech (POS) tags.

POS	K-L distance
J	0.3058
V	0.3123
N	1.0483
O	0.0831

Discussion

Based on the experimental results, now let's conduct some in-depth discussion.

Why Does POS-Based Feature Ensemble Work?

The experimental results showed the effectiveness of POS-based FE for labeling adaptation. In this section, we further discuss why we obtained the observed results. Table 2 shows the average K-L distance of each domain pair regarding different POS tags. Generally, the K-L distances of different types of POS tags can be ranked as: N >> V > J > O. O changes the least, because words such as

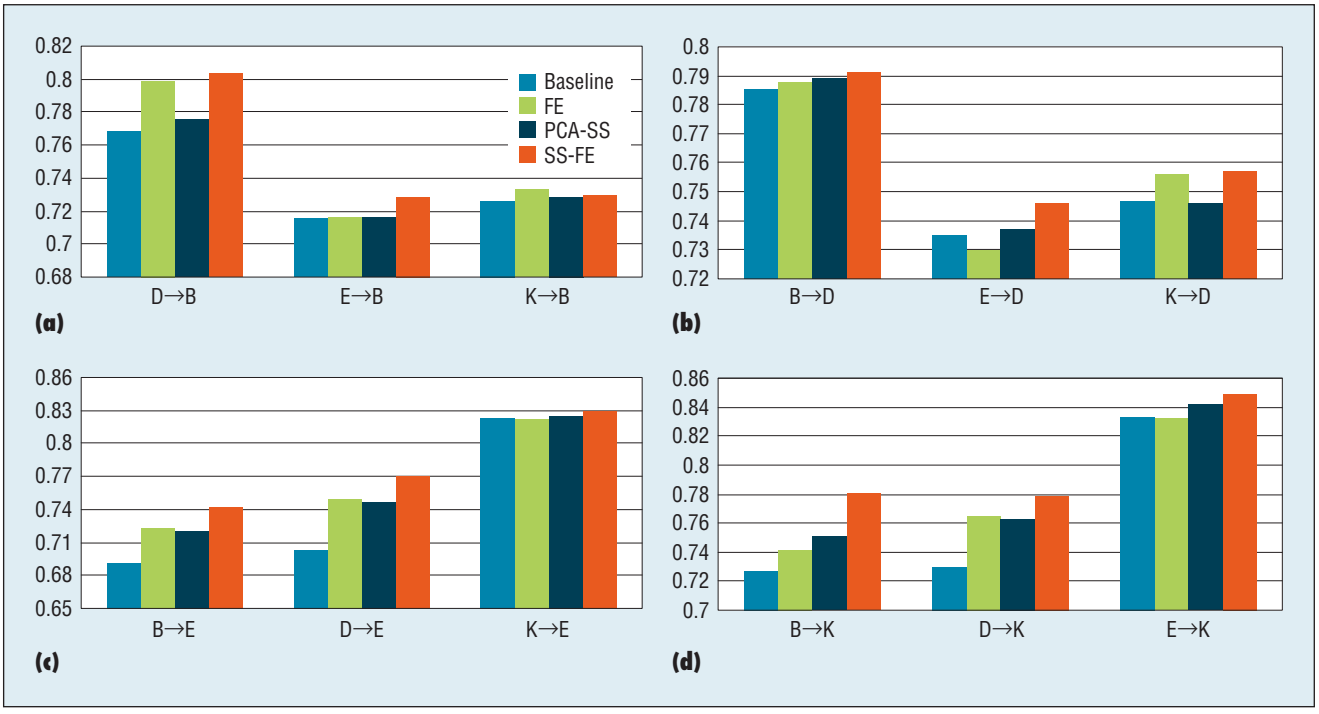

Figure 3. Cross-domain classification accuracy of baseline, FE, PCA-SS, and SS-FE.

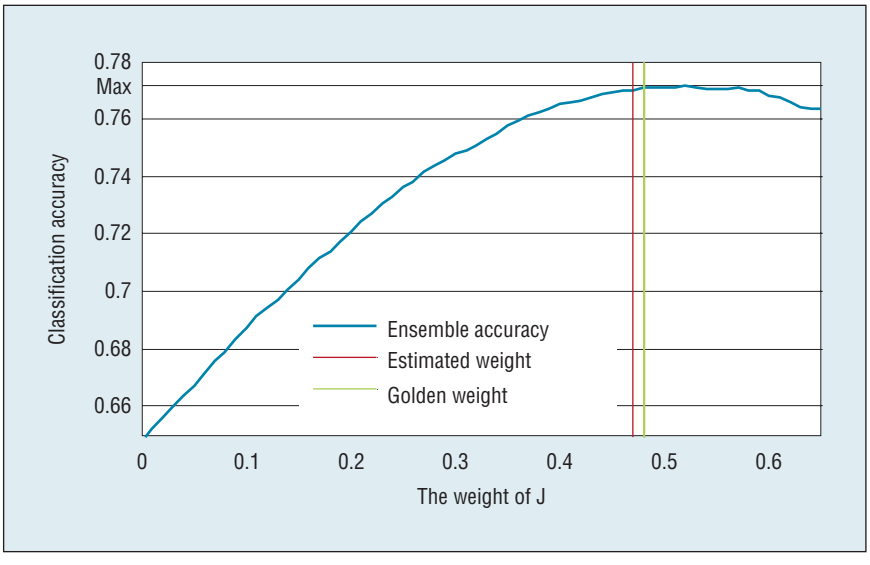

Figure 4. Parameter sensitivity test of FE.

prepositions and pronouns are mostly domain-free. The K-L distance of N is the largest and is significantly larger than the other POS tags, indicating that the change of N is the biggest across domains. The K-L distance of J is significantly smaller than that of N. V gives the comparable K-L distance. This suggests that features in J and V are partially domain-free. It also coincides with our intuition that words such as "great" and "like" always express positive meaning, even when domain changes.

We then observe the weights of different parts of POS tags trained by FE. The ensemble weights trained by FE with respect to POS tags J, V, N, and O are $\omega_J = 0.47$, $\omega_V = 0.19$, $\omega_N = 0.16$, and $\omega_O = 0.18$. The weight of J is the largest, while the weight of N is comparatively small.

We further test the sensitivity of parameter tuning. For the sake of simplicity, we fix the weights of V and O to be 0.19 and 0.16, respectively. We use w_J to denote the weight of J; the weight of N is thus $0.65 - w_J$. We tune the value of w_J from 0 to 0.65, and report the accuracy of FE in Figure 4. We can conclude that the performance is quite sensitive to the weights assigned to J and N. When w_J is close to 0, the weight of N is comparatively larger, and the ensemble performance drops sharply. The best result was obtained when w_J locates at the area close to 0.5, which is close to the results trained by FE. This proves that FE is effective at tuning parameters.

Table 3. K-L distance of unigram distribution regarding different tasks.

Task	K-L distance
B-D	0.1874
B-E	0.1874
B-K	0.4721
D-E	0.4429
D-K	0.4753
E-K	0.2843

Why Feature Ensemble Plus Sample Selection?

Here, we use the maximum likelihood estimation framework to explain why SS-FE could gain a consistent improvement over the single use of PCA-SS or FE. Let $p(\mathbf{x}, y|\theta)$ denote the joint probability of instance $\mathbf{x}$ and class label y. Our aim in domain adaptation is to find the best parameter θ^* that maximizes the expected likelihood in the target domain. Because only the source domain-labeled data are available, we instead maximize the empirical likelihood of the data in the source domain:

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \int_{x\in\chi}\sum_{y\in\gamma} p_t(\mathbf{x},y)\log p(\mathbf{x},y|\theta)d\mathbf{x} \\ &= \arg\max_{\theta} \int_{x\in\chi^{p_t(\mathbf{x})}}\sum_{y\in\gamma} p_t(y|\mathbf{x})\log p(\mathbf{x},y|\theta)d\mathbf{x} \\ &\approx \arg\max_{\theta} \int_{x\in\chi^{\tilde{p}_s(\mathbf{x})}}\sum_{y\in\gamma} \tilde{p}_s(y|\mathbf{x})\log p(\mathbf{x},y|\theta)d\mathbf{x},\end{aligned} \quad (3)$$

where $\tilde{p}_s(\mathbf{x})$ and $\tilde{p}_s(y|\mathbf{x})$ denote the empirical distribution of the source domain.

In Equation 3, both labeling adaptation and instance adaptation are needed. In this work, labeling adaptation is conducted in a feature reweighting manner by the FE. Instance adaptation is embodied in the manner of sample selection, where the empirical likelihood is obtained based on the selected subset of samples. To validate this analysis, we present the K-L distance of different domain pairs in Table 3. We find that the K-L distances of B-D and E-K are relatively low, which suggests that these domains are similar in distribution. This corresponds well with our experimental results: first, in tasks such as D → B and K → D, the improvement of SS-FE is generally gained by labeling adaptation. This is quite reasonable, because their K-L distance is relatively low and the need for instance adaptation isn't urgent. Second, for the domain pairs whose distributional change is larger, such as E → D and E → K, the improvement of SS-FE is mainly due to instance adaptation rather than labeling adaptation.

We developed SS-FE to meet two distinct needs in domain adaptation: labeling adaptation and instance adaptation. Experimental results showed the effectiveness of SS-FE in both labeling adaptation and instance adaptation. One shortcoming of our approach is that training samples are selected in a "hard" way, which is sometimes too arbitrary. A "soft" manner, which assigns a sampling weight to each of the training samples, seems to be more promising.[6]

In the future, we plan to consider a "soft" manner in instance adaptation. We also plan to integrate instance adaptation with some unsupervised labeling adaptation methods, such as structural correspondence learning (SCL) and spectral feature alignment (SFA), to test our model's effectiveness over a broad range. ■

THE AUTHORS

Rui Xia is an assistant professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include natural language processing, machine learning, and text mining. Xia has a PhD in computer science from the Institute of Automation, Chinese Academy of Sciences. Contact him at rxia@njust.edu.cn.

Chengqing Zong is a professor at the National Laboratory of Pattern Recognition, Chinese Academy of Sciences' Institute of Automation. His research interests include machine translation, natural language processing, and sentiment classification. Zong has a PhD in computer science from the Chinese Academy of Sciences' Institute of Computing Technology. He's a director of the Chinese Association of Artificial Intelligence and the Society of Chinese Information Processing, and is a member of International Committee on Computational Linguistics (ICCL). He's an associate editor of *ACM Transactions on Asian Language Information Processing* and an editorial board member of *IEEE Intelligent Systems* and *Machine Translation*. Contact him at cqzong@nlpr.ia.ac.cn.

Xuelei Hu is an associate professor in the School of Computer Science and Engineering at Nanjing University of Science and Technology, China. Her research interests include machine learning, pattern recognition, and computer vision. Hu has a PhD in computer science from the Chinese University of Hong Kong. Contact her at xlhu@njust.edu.cn.

Erik Cambria is a research scientist in the Cognitive Science Programme, Temasek Laboratories, National University of Singapore. His research interests include AI, the Semantic Web, natural language processing, and big social data analysis. Cambria has a PhD in computing science and mathematics from the University of Stirling. He is on the editorial board of Springer's *Cognitive Computation* and is the chair of many international conferences, including Brain-Inspired Cognitive Systems (BICS) and Extreme Learning Machines (ELM). Contact him at cambria@nus.edu.sg.

Acknowledgments

This research work is supported by the Jiangsu Provincial Natural Science Foundation of China (BK2012396), the Research Fund for the Doctoral Program of Higher Education of China (20123219120025), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR). This work is also partly supported by the Hi-Tech Research and Development Program of China (2012AA011102 and 2012AA011101), and the External Cooperation Program of the Chinese Academy of Sciences.

References

1. J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," *Proc. Ann. Meeting of the Assoc. Computational Linguistics*, ACL, 2007, pp. 264–271.
2. E. Cambria et al., "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, 2013, pp. 15–21.
3. B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012.
4. H. Daume, III, "Frustratingly Easy Domain Adaptation," *Proc. Ann. Meeting of the Assoc. for Computational Linguistics*, ACL, 2007, pp. 256–263.
5. R. Xia, C. Zong, and S. Li, "Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification," *Information Sciences*, vol. 181, no. 6, 2011, pp. 1138–1152.
6. R. Xia et al., "Instance Selection and Instance Weighting for Cross-Domain Sentiment Classification via PU Learning," *Proc. Int'l Joint Conf. on Artificial Intelligence* (IJCAI), IJCAI, 2013.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.