Additional Experiments for "On the Use of Sparse Filtering for Covariate Shift Adaptation"

Fabio Massimo Zennaro[1], **Ke Chen**[2]
[1]Department of Informatics, University of Oslo, Oslo, Norway
[2]School of Computer Science, The University of Manchester, Manchester, UK

Keywords: covariate shift, feature distribution learning, sparse filtering, periodic sparse filtering, structure of the data, periodic structure

Real-World Data Experiments: Text Sentiment Analysis

We run further experiments on textual data for sentiment analysis (SA). We consider this type of data for two reasons: (i) SA data sets are well-known and widely studied in the literature on covariate shift adaptation (Blitzer, McDonald, & Pereira, 2006), and (ii) SA data may exhibit periodic behaviour within different categories.

In our simulations we worked with the popular Amazon Multi-Domain Sentiment Dataset (Blitzer et al., 2006). The data set contains 8000 online reviews evenly divided in four product categories (books, dvd, electronics and kitchen appliances) made available in the form of collections of unigrams and bigrams [1].

Samples are converted into a TF-IDF (term frequency-inverse document frequency) (Schütze, Manning, & Raghavan, 2008) bag-of-words representation using standard functions in the *scikit*[2] library. To reduce the dimensionality of the data vector we avoided further processing that could have drastically affected the distribution of the data (PCA, LDA); instead we followed the simple approach adopted by Glorot, Bordes, and Bengio (2011) and Gopalan, Li, and Chellappa (2011) of selecting the top-k most common features. We select a low number of features ($k = 50$), and to make the simulation more realistic and challenging, we select the top-k most common features over the training data only. Each review is provided with a binary label denoting whether the review is positive or negative. For each category, half of the reviews are labeled as positive and half of the reviews are labeled as negative. Reviews are balanced with respect to the category they belong to and with respect to the labels, so no upsampling is performed.

[1] `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`
[2] `http://scikit-learn.org/`

Table 1: Best hyper-parameter configurations chosen by model selection.

	noCSA+SVM	***SF+SVM***	***PSF+SVM***
book	SVM: $C = 1$	SF: $L = 120$ SVM: $C = 0.075$	PSF: $g = \sin()$ PSF: $L = 50$ PSF: $\lambda = 2.4$ SVM: $C = 0.75$
elec	SVM: $C = 0.01$	SF: $L = 120$ SVM: $C = 5 \cdot 10^{-5}$	PSF: $g = \sin()$ PSF: $L = 100$ PSF: $\lambda = 2.2$ SVM: $C = 1$
kitc	SVM: $C = 0.75$	SF: $L = 100$ SVM: $C = 0.05$	PSF: $g = \sin()$ PSF: $L = 120$ PSF: $\lambda = 2$ SVM: $C = 1$
	IW+LSPC	***SSA+SVC***	***DAE+SVM***
book	LSPC: $\sigma = 1$ LSPC: $\lambda = 1$	SSA: $P = 50$ SVM: $C = 1$	DAE: $f, g = \text{sigm}()$ DAE: $L = 100$ DAE: $\sigma = 0.01$ DAE: $\eta = 0.0005$ SVM: $C = 0.00075$
elec	LSPC: $\sigma = 1$ LSPC: $\lambda = 1$	SSA: $P = 35$ SVM: $C = 1$	DAE: $f, g = \text{sigm}()$ DAE: $L = 100$ DAE: $\sigma = 0.01$ DAE: $\eta = 0.001$ SVM: $C = 0.75$
kitc	LSPC: $\sigma = 3$ LSPC: $\lambda = 1$	SSA: $P = 35$ SVM: $C = 1$	DAE: $f, g = \text{sigm}()$ DAE: $L = 100$ DAE: $\sigma = 0.01$ DAE: $\eta = 0.0005$ SVM: $C = 0.75$

In each simulation we partition the data using the following protocol: we select a single category (dvd) as the training set $\mathbf{X}^{tr}$; we then select a second category and evenly split it between a target set $\mathbf{X}^{tgt}$ and a test set $\mathbf{X}^{tst}$. As before, adaptation is performed on the training and target data; classification uses training data for learning, target data for model selection and test data for evaluation

All the data are processed using the same classification systems and the same settings used in the previous experiment on ESR data. In order to perform model selection over a reasonable set of values for the hyper-parameters of each algorithm, the six CSA classification systems are configured as follows:

(i) *SVM*: a linear SVM is trained on $\{\mathbf{X^{tr}}, \mathbf{Y^{tr}}\}$. Model selection is performed on a single hyper-parameter: the soft cost C is chosen in the set $\{2.5 \cdot 10^{-5}, 5 \cdot 10^{-5}, 7.5 \cdot 10^{-5}, ..., 0.5, 0.75, 1.0\}$ as before.

(ii) *SF+SVM*: SF is trained on $\{\mathbf{X^{tr}}, \mathbf{X^{tar}}\}$ using an early stopping criterion. Model selection is performed on a single hyper-parameter: the learned features L is chosen in the set $\{20, 50, 80, 100, 120\}$ in order to explore a coarse-grained grid space around the original number of features $M = 50$. SVM is trained as for system (i).

(iii) *PSF+SVM*: PSF is trained on $\{\mathbf{X^{tr}}, \mathbf{X^{tar}}, \mathbf{Y^{tr}}\}$ using an early stopping criterion. Model selection is performed on three hyper-parameters: the non-linearity g chosen in the set $\{\sin(), \cos()\}$, the learned feature L in the set $\{20, 50, 80, 100, 120\}$ as before, the loss

parameter λ in the set $\{0.8, 0.9, 0.95, 1.0, 1.05, 1.1, 1.15, 1.2, 1.3, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8\}$ in order to explore a finer-grained grid space around the value $\lambda = 1$. The dimensionality of the learned space is given by the number of features L, evenly divided between the two classes, plus a fixed number of 10 features to account for unlabelled samples. SVM is trained as for system (i).

(iv) *IW+LSPC*: IW is trained on $\{\mathbf{X^{tr}}, \mathbf{X^{tar}}\}$ setting the number of basis to the cardinality of the target set $\mathbf{X^{tar}}$. Model selection is performed on two hyper-parameters: the candidate σ chosen in the set $\{0.1, 0.2, 0.5, 1, 2, 3\}$ and the candidate λ in the set $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ following Hachiya, Sugiyama, and Ueda (2012). LSPC is trained on $\{\mathbf{X^{tr}}, \mathbf{Y^{tr}}\}$. Model selection is performed on two hyper-parameter: the candidate σ chosen in the set $\{0.1, 0.2, 0.5, 1, 2, 3\}$ and the candidate λ in the set $\{0.1, 0.17, 0.32, 0.56, 1\}$ following Hachiya et al. (2012).

(v) *SSA+SVM*: SSA is trained on $\{\mathbf{X^{tr}}, \mathbf{X^{tar}}\}$. Model selection is performed on a single hyper-parameter: the number of PCA components P chosen in the set $\{15, 35, 50\}$ in order to explore a coarse-grained grid below the original number of features $M = 50$. SVM is trained as for system (i).

(vi) *DAE+SVM*: DAE is trained on $\{\mathbf{X^{tr}}, \mathbf{Y^{tr}}\}$ for 10000 epochs and the noise set to a Gaussian with zero mean. Model selection is performed on four hyper-parameters: the non-linearity (for both encoding and decoding) is chosen in the set $\{\text{sigmoid}(), \tanh()\}$ following the standard in the neural network literature, the learned features L in the set $\{50, 70, 100\}$ in order to explore a coarse-grained grid equal or above the original number of features $M = 50$, the variance of the noise σ in the set $\{0.01, 0.05, 0.1\}$ and the learning rate η in the set $\{0.0005, 0.001, 0.005\}$ following the standard in the neural network literature.

Hyper-parameters for the classification systems are selected by cross-validation using a standard grid search method. Table 1 reports the hyper-parameter configurations selected by model selection.

Performance is evaluated again in terms of unweighed average recall (UAR); for each configuration, we report the mean and the standard error achieved over 10 independent simulations. Statistical validation of the results is evaluated by using a paired Wilcoxon test with the null hypothesis that the classification performance with and without CSA has the same mean (under the assumption that the results are symmetrically distributed around the true mean performance); statistics for the hypothesis test are collected from 100 independent trials.

Figure 1 shows the UAR of different classification systems on the book, electronics and kitchen categories using the protocol described above. In general, the Amazon sentiment data set proved challenging for all CSA algorithms. When adapting from the dvd category to the book category, only the PSF and the DAE algorithms provide an improvement in the final classification. This may suggest that the structure of the data has some form of weak periodicity, while, at the same time, it seems to strongly exclude any other sort of structure discovered by SF (radial structure), SSA (projectability of the PCA components of different categories on each other) or IW (continuity between the sub-domains of different categories); in general, though, on this specific adaptation problem, the structure-agnostic

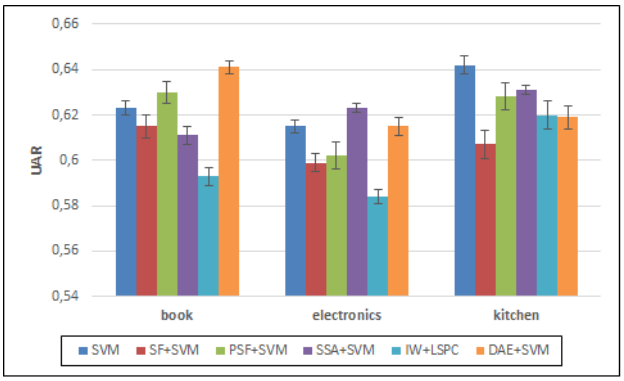

Figure 1: UAR Accuracies of different CSA methods along with the baselines.

DAE performs the best. When adapting from the dvd category to the electronics category, the only algorithm able to provide a consistent improvement is SSA. This may suggest that the PCA components of the dvd category and the electronics category carry useful discriminative information and may be projected on each other. All the remaining CSA algorithms either perform at the level of the starting representation or negatively affect the final performance. Finally, when adapting from the dvd category to the kitchen category, no algorithm is able to learn a good representation. This may be due to strongly different data structures underlying the data in the dvd category and the data in the kitchen category. Such a difference may be indeed justified by the arguably large semantic gap between the domain of reviews of dvd movies and the domain of reviews of kitchen appliances. Despite the limited success of CSA, processing the data using algorithms grounded on different assumptions offered us potential insights on the structure of the data at hand. Our assumption of periodicity proved once again more reliable than the assumption of radial structure. Indeed, the PSF algorithm was able to provide an improvement in the adaptation task on the book data, thus guaranteeing good results and suggesting that the book data and the dvd data may share some similarity over the conditional distribution of the labels; on the other hand, the SF algorithm could not succeed in any adaptation task, hinting once again to the fact that a radial structure underlying the data is a less common scenario. Also, the failure of specific CSA algorithms provide us with a reasonable ground to formulate hypothesis about the structure of the conditional distribution of the labels of the data at hand.

References

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 120–128).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (icml-11)* (pp. 513–520).

Gopalan, R., Li, R., & Chellappa, R. (2011). Domain adaptation for object recognition: An unsupervised approach. In *Computer vision (iccv), 2011 ieee international conference on* (pp. 999–1006).

Hachiya, H., Sugiyama, M., & Ueda, N. (2012). Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, *80*, 93–101.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.