DRASO: Declaratively Regularized Alternating Structural Optimization

Partha Pratim Talukdar PARTHA@CIS.UPENN.EDU
Ted Sandler TSANDLER@CIS.UPENN.EDU
Mark Dredze MDREDZE@CIS.UPENN.EDU
Koby Crammer CRAMMER@CIS.UPENN.EDU
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 USA

John Blitzer JOHN@BLITZER.COM
Microsoft Research Asia, Hai Dian District, Beijing, China 100080

Fernando Pereira PEREIRA@GOOGLE.COM
Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 USA

Abstract

Recent work has shown that Alternating Structural Optimization (ASO) can improve supervised learners by learning feature representations from unlabeled data. However, there is no natural way to include prior knowledge about features into this framework. In this paper, we present Declaratively Regularized Alternating Structural Optimization (DRASO), a principled way for injecting prior knowledge into the ASO framework. We also provide some analysis of the representations learned by our method.

1. Introduction

While supervised learning algorithms achieve impressive results on a variety of NLP tasks, they rely on the availability of labeled data. The application of other available resources to improve over existing supervised methods has been explored in semi-supervised learning. There are two primary sources of information for semi-supervised algorithms: unlabeled data and prior knowledge. Alternating Structual Optimization (ASO) (Ando & Zhang, 2005) is a semi-supervised learning technique based on unlabeled data, which has achieved considerable success in many important problems (Blitzer et al., 2006; Blitzer et al., 2007). ASO learns a new data representation by constructing and solving a multitask learning problem using unlabeled data. While ASO makes excellent use of unlabeled data, there is currently no way to encode prior information in learning the representations. For example, in the sentiment classification task, a short list of positive and negative words can be used to bootstrap learning (Turney, 2002).

In this work we seek to combine ASO with this type of prior knowledge. We present Declaratively Regularized ASO (DRASO), which favors learning representations that are consistent with some side information. DRASO combines both unlabeled data and prior knowledge to find a single representation of the data. This paper describes DRASO and shows that the representations learned for sentiment classification using side information can improve over a standard ASO representation.

2. DRASO

Given a number of related supervised learning problems, ASO learns a shared low dimensional representation of the data in order to minimize the empirical risk across the various tasks. Specifically, let the training set for task ℓ be $\{(\mathbf{x}_i^\ell, y_i^\ell)\}_{i=1}^{n_\ell}$. Given m such training sets, ASO learns a shared representation $\hat{\mathbf{\Phi}}$ and associated weight vectors $\hat{\mathbf{w}}_\ell$, $\hat{\mathbf{v}}_\ell$, $\ell = 1, ..., m$ by minimizing

Appearing in *Appearing in the Workshop on Prior Knowledge at the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

the loss over the training sets:

$$\left[\{\hat{\mathbf{w}}_\ell, \hat{\mathbf{v}}_\ell\}, \hat{\mathbf{\Phi}}\right] = \underset{\mathbf{w}_\ell, \mathbf{v}_\ell, \mathbf{\Phi}}{\operatorname{argmin}} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L\left((\mathbf{w}_\ell + \mathbf{\Phi}'\mathbf{v}_\ell)'\mathbf{x}_i^\ell, y_i^\ell\right) + \lambda ||\mathbf{w}_\ell||^2 \right) \quad \text{s.t. } \mathbf{\Phi}\mathbf{\Phi}' = I_{k\times k}.$$

The matrix $\mathbf{\Phi}$ is a shared transformation which maps a feature vector $\mathbf{x} \in \mathbb{R}^D$ to a low-dimensional vector in $\mathbb{R}^k$. Given $\mathbf{\Phi}$, $\mathbf{w}_\ell$, and $\mathbf{v}_\ell$, the prediction for an instance $\mathbf{x}^\ell$ is the linear function $(\mathbf{w}_\ell + \mathbf{\Phi}'\mathbf{v}_\ell)'\mathbf{x}^\ell$ where $\mathbf{w}_\ell$ is the weight vector applied to the original instance and $\mathbf{v}_\ell$ is the weight vector applied to the shared, low-dimensional represention, $\mathbf{\Phi}\mathbf{x}^\ell$.

Unfortunately, as written, the ASO criterion does not allow one to inject prior knowledge into the learned shared transformation $\mathbf{\Phi}$. For example, in the sentiment classification task, we may wish to represent the fact that presence of *excellent* or *superb* in a document express similar sentiment and hence a classifier should assign similar weights to the two features corresponding to the presence of these two words. To incorporate such declarative information, we suppose the existence of a prior knowledge graph which encodes knowledge about which features should be similarly correlated with the class labels in a "good" model. The nodes of the graph represent features and the edges represent feature similarities. The edges are weighted by the strength of similarity. These weights are encoded as a matrix $P \in R^{D\times D}$ with each entry $P_{ij} \geq 0$, $P_{ii} = 0$ and $\sum_j P_{ij} = 1$ for all i.

To enforce the similarity requirements, we replace the ridge regularization term with a penalty on the induced norm: $\mathbf{w}^{'} M \mathbf{w}$, where $M = (I - P)^{'}(I - P)$. This encourages features to be weighted similarly to the average of their neighbors' weights and is closely related to the LLE objective (Roweis & Saul, 2000; Sandler et al., 2008). The new optimization problem is then given as:

$$\left[\{\hat{\mathbf{w}}_\ell, \hat{\mathbf{v}}_\ell\}, \hat{\mathbf{\Phi}}\right] = \underset{\mathbf{w}_\ell, \mathbf{v}_\ell, \mathbf{\Phi}}{\operatorname{argmin}} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L\left((\mathbf{w}_\ell + \mathbf{\Phi}'\mathbf{v}_\ell)'\mathbf{x}_i^\ell, y_i^\ell\right) + \lambda \mathbf{w}^{'} M \mathbf{w} \right) \quad \text{s.t. } \mathbf{\Phi}M\mathbf{\Phi}' = I_{k\times k}.$$

We call this new objective DRASO, since the ASO objective is declaratively regularized. Solving for Φ yields a new eigenvalue problem, which can be solved efficiently (section 2.1).

2.1. Solving for Φ

Our main goal is to find the transformation Φ which we will use to create a new representation for the supervised problem. As in (Ando & Zhang, 2005), we can simplify the problem by making the change of variables $u_\ell = w_\ell + \Phi^{'} v_\ell$. This yields the optimization problem

$$\left[\{\hat{\mathbf{u}}_\ell, \hat{\mathbf{v}}_\ell\}, \hat{\mathbf{\Phi}}\right] = \underset{\mathbf{u}_\ell, \mathbf{v}_\ell, \mathbf{\Phi}}{\operatorname{argmin}} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L\left(\mathbf{u}_\ell^{'}\mathbf{x}_i^\ell, y_i^\ell\right) + \lambda(\mathbf{u}_\ell - \Phi^{'}\mathbf{v}_\ell)^{'} M (\mathbf{u}_\ell - \Phi^{'}\mathbf{v}_\ell) \right) \quad \text{s.t. } \mathbf{\Phi}M\mathbf{\Phi}' = I_{k\times k}.$$

Again following (Ando & Zhang, 2005), we can solve this problem using an alternating minimization technique. In the first step of the alternation, we fix Φ and $\mathbf{v}_\ell$ and solve for $\mathbf{u}$. As before, this step amounts to solving the decoupled linear predictions for each of the problems. In the second step, we fix $\mathbf{u}_\ell$ and solve for $\mathbf{v}_\ell$ and Φ. First we note that $\mathbf{v}_\ell$ has a closed form solution in terms of Φ and $\mathbf{u}_\ell$.

$$\left[\{\hat{\mathbf{v}}_\ell\}, \hat{\mathbf{\Phi}}\right] = \underset{\mathbf{v}_\ell, \mathbf{\Phi}}{\operatorname{argmin}} \sum_{\ell=1}^{m} \left(\lambda(\mathbf{u}_\ell - \Phi^{'}\mathbf{v}_\ell)^{'} M (\mathbf{u}_\ell - \Phi^{'}\mathbf{v}_\ell) \right) \quad \text{s.t. } \mathbf{\Phi}M\mathbf{\Phi}' = I_{k\times k}.$$

Solving for $\mathbf{v}_\ell$ in this quadratic form gives us $\mathbf{v}_\ell = \Phi M \mathbf{u}_\ell$. Now we can solve for the following minimization problem for Φ:

$$\left[\hat{\mathbf{\Phi}}\right] = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \sum_{\ell=1}^{m} ||\Phi M \mathbf{u}_\ell||_2^2 \quad \text{s.t. } \mathbf{\Phi}M\mathbf{\Phi}' = I_{k\times k}.$$

Following (Ando & Zhang, 2005), we have that this problem is equivalent to the problem

$$\left[\hat{\mathbf{\Phi}}\right] = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \operatorname{tr}\left(\mathbf{\Phi} M U U^{'} \mathbf{\Phi}\right) \quad \text{s.t. } \mathbf{\Phi}M\mathbf{\Phi}' = I_{k\times k}.$$

By looking at the first order conditions for the Lagrangian, we can see that the solutions have the form

$$MUU^{'}M\mathbf{\Phi} = \alpha M \mathbf{\Phi}$$

We can transform this generalized eigenvalue problem into one that is smaller and easier to manage if we let $\theta = U^{'} M\mathbf{\Phi}$. Now, right multiplying by $U^{'}$, we get:

$$U^{'} MU\theta = \alpha\theta$$

That is we can solve for the eigenvectors of the modified gram matrix (transformed via M). Now, we can substitute back into the original problem (noting that M is symmetric).

$$\begin{aligned} MUU^{'} M\mathbf{\Phi} &= \alpha M\mathbf{\Phi} \\ MU\theta &= \alpha M\mathbf{\Phi} \end{aligned}$$

Thus we have, $\mathbf{\Phi} = \frac{1}{\alpha} U\theta$.

3. Experimental Results

ASO and DRASO representations were compared on the sentiment classification task using Amazon book reviews from Blitzer et al. (2007). Auxiliary problems were selected using mutual information. Prior knowledge was obtained from SentiWordNet (Esuli & Sebastiani, 2006) by manually selecting 31 positive and 42 negative words from the top ranked positive and negative words in SentiWordNet. Each selected word was connected in graph P to its 10 nearest neighbors according to SentiWordNet rank.

The learned Φs were used to project 32,502 words into a two dimensional space (Figure 1). Words on the prior knowledge lists are indicated by squares (negative) and triangles (positive). Points are color coded based on their behavior in a large sample of labeled training data (13,391 instances) as red (positive), blue (negative) and grey (neutral). The figures indicate that list words clumped by ASO are separated by DRASO. Additionally, while pulling apart high-sentiment words, neutral words are left together. Finally, observe that additional points not on the list have been pulled as well, showing the effect of prior knowledge on new features. These results indicate that DRASO can incorporate prior information into ASO in a principled and effective way.

4. Related Work

The feature graph based additional regularization term in the DRASO objective is close in spirit to Fused Lasso (Tibshirani et al., 2005). However, there are crucial differences. Firstly, fused lasso assumes an ordering over the features while no such restriction is necessary in case of DRASO. Secondly, fused lasso imposes an L_1 penalty over differences in weights of consecutive features (assuming the features are ordered as mentioned above). In contrast, DRASO uses an L_2 norm and the regularization is imposed over immediate neighborhood rather than pairwise constraints. The L_1 penalty in (Tibshirani et al., 2005) prefers weights of linked features to be exactly same. However, in many problem domains (including the ones considered in this paper), it is desirable to have similar rather than identical weights.

The additional regularization term in the DRASO objective is similar to the one in Penalized Discriminant Analysis (PDA) (Hastie et al., 1995). While PDA performs standard classification, DRASO is focused on learning a new and more effective representation in ASO's multitask learning setting. The learned representation could in turn be used as additional features in a standard classifier, which is currently being investigated (Section 5).

5. Conclusion

In this paper we have presented DRASO, which extends ASO by adding a regularization term. This additional term makes it possible to inject valuable prior knowledge into the ASO framework. We have shown that while solving for Φ, incorporation of the additional regularization term results in an eigenvalue problem (different from ASO) which can be solved efficiently. We have also presented experimental evidence demonstrating effectiveness of DRASO over ASO.

For future work, we are considering applications to learning tasks for which ASO has performed well. For many of these tasks, prior knowledge can be added through existing resources or through the use of unsupervised methods to infer relations between features. We are also investigating under what conditions prior knowledge can improve over labeled data alone.

Acknowledgment

We would like to thank the anonymous reviewers for helpful feedback. This work is supported in part by NSF IIS-0513778.

References

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research (JMLR)*, *6*, 1817–1853.

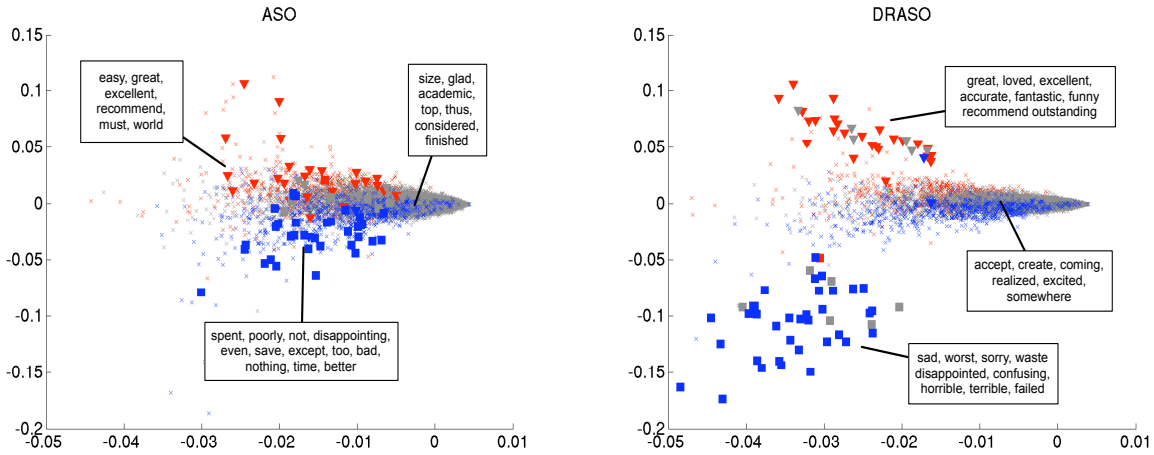

Figure 1. ASO and DRASO projections of 32,502 words into a two-dimensional space. Squares (negative) and triangles (positive) indicate prior knowledge words. Polarity of features as measured from labeled data is indicated by blue (negative), red (positive) and grey (neutral). Some of the features are annotated to demonstrate the effects of the projection.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Association for Computational Linguistics (ACL)*.

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. *Empirical Methods in Natural Language Processing (EMNLP)*.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC*, 417–422.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, *23*, 73–102.

Roweis, S., & Saul, L. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding.

Sandler, T., Blitzer, J., & Ungar, L. H. (2008). Learning with locally linear feature regularization. *Snowbird Learning Workshop*.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, *67*, 91–108.

Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Association for Computational Linguistics (ACL)*.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. *ACM International Conference Proceeding Series*.