Transfer Learning for Enhancing Information Flow in Organizations and Social Networks

Chris Pal, Xuerui Wang and Andrew McCallum
Computer Science Dept.
University of Massachusetts
Amherst, MA 01003

Abstract

The task of suggesting recipients for an email has recently received attention as it has potential to enhance the flow of knowledge and information within an organization or social network. We investigate two transfer learning techniques to improve recipient prediction performance through considering predictions for multiple users. We present a novel continuous hidden variable conditional random field for the recipient prediction problem. We characterize this construction as a type of discriminative author recipient topic or DART model. First we show transfer based performance increases achieved through shared hidden variables for prediction across different users. Second, we show how transfer from an organization wide model to a user specific model through parameter prior structure also confers substantial advantage, especially when models are constructed for new users.

1 Introduction

The problem of CC prediction was introduced in [10] along with a number of generative probabilistic models for solving the problem. Here we use discriminative methods for the general problem of recipient prediction and focus upon an exploration of two transfer learning techniques. We show that it is possible to leverage the information contained within the related prediction tasks for different users and increase overall prediction performance. These methods also show great potential for increasing the performance of models for new users.

There has been growing interest in the exploration of transfer learning methods within the Machine Learning community [3, 6, 1, 5, 11, 9]. Early work in [3] describes multitask learning as an approach to exploit information used for the training of other tasks to improve a given task. Other recent work explores multitask learning based on the minimization of regularization functionals in the context of Support Vector Machine (SVM) based approaches [5]. In contrast, the first aspect of our exploration here uses a discriminative low dimensional latent variable representation to make predictions for different senders across an organization. Our approach is thus related to [6] which sought to obtain a low dimensional latent space suitable for multiple image classification tasks. The second component of our exploration uses a graphical model over parameter priors to transfer information from an organization wide model for predictions to user specific models. Recently, [11] also explored transfer learning using more sophisticated parameter priors but for simpler discriminative models.

2 Recipient Prediction with DARTs

Email recipient prediction in the context of a recommending system for users in a social network or organization is a challenging problem for a number of reasons, including: (1) there are possibly hundreds or even thousands of possible recipients; (2) the true number of reasonable potential recipients is typically unknown; furthermore, (3) while some suggestions may indeed be reasonable, unless extensive analysis and hand labeling is used for augmenting labels, these suggestions may be flagged as incorrect. While these issues are important to consider, augmentations to recipient lists can be subjective and therefore we evaluate suggestions based on a test subset of emails and their observed recipients.

Topic models have also received substantial recent attention in the Machine Learning community. A variety of methods have emerged as alternatives to the original approach to Latent Semantic Analysis (LSA) [4] or direct Principal Component Analysis (PCA) [7] of a term document matrix. Latent Dirichlet Allocation (LDA) [2] is widely regarded as a state of the art topic modeling method when one wishes to use only the words of a document to obtain topics. Recently, McCallum *et al.* [8] extended the basic LDA approach to include explicit Author, Recipient and Topic variables, we shall refer to this approach as ART models. These models are very effective at extracting meaningful topics and have been shown to reveal user roles in social networks. Despite these attractive attributes, our experiments with these types of ART models for the task of recipient prediction have produced performance far below the baseline method discussed in Section 3. However, as discussed in Section 1, hidden variables and sophisticated prior structures can be used to implement a variety of transfer learning approaches. These factors motivate our development of the following Discriminative Author Recipient Topic (DART) models and two approaches to transfer learning with DARTs.

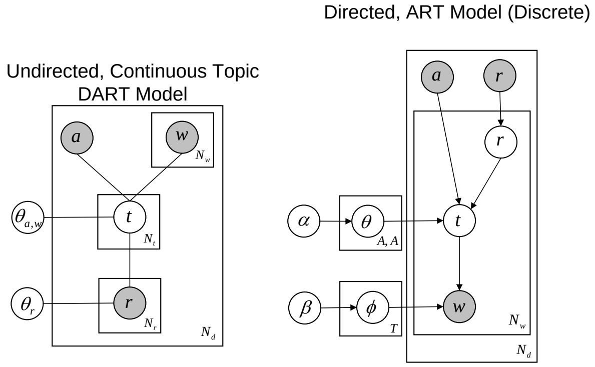

Figure 1: Our DART model vs. the ART model of [8]

In the following exposition we present a continuous hidden variable random field based DART model. Our model can be characterized as a type of discriminative Boltzmann machine, or a rich discriminative multinomial generalization of probabilistic PCA [12]. We use random variables and notation shown in figure 1. Figure 1 (left) illustrates our model as a plated random field and contrasts our DART model with the ART model of [8] (right). Our DART model encodes the conditional probability of recipients and hidden topics given words and the email author as:

$$P(\mathbf{r_d}, \mathbf{t_d} | \mathbf{w_d}, a_d) = \mathcal{Z}(\mathbf{w_d}, a_d)^{-1} \exp\Bigg[\sum_{i=1}^{T}(-\log(t_{di})$$
$$-\frac{1}{2}\log^2(t_{di}) + (\sum_{j=1}^{N_d} M^w_{iw_{dj}} + \sum_{k=1}^{S_d} M^r_{ir_{dk}} + M^a_{ia_d})\log(t_{di}))\Bigg],$$

where we have coupled the random variables by the connection matrices M^w, M^r and M^a. $\mathcal{Z}$ is an intractable normalization constant. For notation convenience, we set $M^w_{iV} = 0$, for $i = 1, \cdots, T$, $M^r_{iR} = 0$, for $i = 1, \cdots, T$, and $M^a_{iA} = 0$, for $i = 1, \cdots, T$. Integrating out the uncertainty of hidden topics, we seek to optimize the marginal conditional likelihood of recipients given words and authors for the corpus,

$$\prod_{d=1}^{D} P(\mathbf{r_d}|\mathbf{w_d}, a_d) = \mathcal{Z}(\mathbf{w_d}, a_d)^{-1} \cdot$$
$$\exp\left[\frac{1}{2}\sum_{d=1}^{D}\sum_{i=1}^{T}(\sum_{j=1}^{V-1} M^w_{ij} m_{dj} + \sum_{k=1}^{S_d} M^r_{ir_{dk}} + M^a_{ia_d})^2\right] \quad (1)$$

SYMBOL	DESCRIPTION
T	number of topics
N_d	number of emails
V	number of unique words
R	number of recipients
A	number of authors
N_w	number of word tokens in email d
S_d	number of recipients on email d
M^w	$T \times (V-1)$ word connection matrix
M^r	$T \times (R-1)$ recipient connection matrix
M^a	$T \times (A-1)$ author connection matrix
t_{di}	the i^{th} topic of email d
w_{dj}	the j^{th} word of email d
r_{dk}	the k^{th} recipient of email d
a_d	the author of email d

Table 1: Notation used in this paper

To perform learning with our DART model we take the gradient of the conditional log likelihood and arrive at the following update rules:

$$\delta M^w_{ij} \quad \propto \quad \sum_{d=1}^{D}(m_{dj}\sum_{k=1}^{S_d} M^r_{ir_{dk}} - \sum_{k=1}^{S_d} M^r_{i\hat{r}_{dk}}) - \frac{M^w_{ij}}{\sigma^2} \quad (2)$$
$$\delta M^r_{ik} \quad \propto \quad \sum_{d=1}^{D}(I(k \in \mathbf{r_d})(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^r_{ir_{dk}} + M^a_{ia_d})$$
$$- \quad I(k \in \hat{\mathbf{r}}_\mathbf{d})(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^r_{i\hat{r}_{dk}} + M^a_{ia_d})) - \frac{M^r_{ik}}{\sigma^2}$$
$$\delta M^a_{ib} \quad \propto \quad \sum_{d=1}^{D} I(b = a_d)(\sum_{k=1}^{S_d} M^r_{ir_{dk}} - \sum_{k=1}^{S_d} M^r_{i\hat{r}_{dk}}) - \frac{M^a_{ib}}{\sigma^2},$$

where $\hat{\mathbf{w}}_\mathbf{d}$ or equivalently $\hat{m}_{dk}$, $d = 1, \ldots, D$, and $k = 1, \ldots, V-1$ denote draws from a Gibbs sampler and $I(q \in Q)$ and $I(a = b)$ are indicator functions. Under the DART model here, the conditional distributions required for the sampler are either log-normal for t or multinomial for the other variables. As with more tractable conditional random fields, these updates have an intuitive interpretation as consisting of differences between expectations involving the empirical or data distribution and expectations based on (approximated) model distributions. The final terms in these updates arise from the use of a zero mean Gaussian prior for parameters.

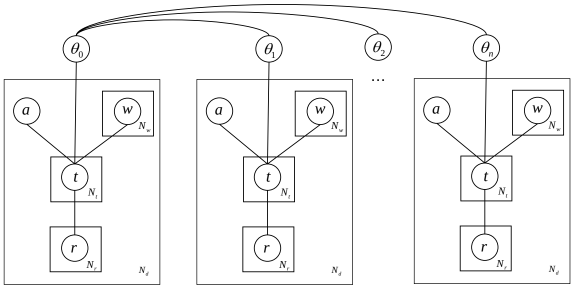

Figure 2: Transfer learning using information from an organization scale model (left) to user specific models.

3 Results and Conclusions

To provide a straightforward, intuitive evaluation we use a mean reciprocal rank (MRR) metric. In information retrieval the reciprocal rank of a test document is the reciprocal of the rank at which the first relevant response was returned, or 0 if none of the responses contained a relevant answer. The score for a sequence of queries is the mean of the individual query's reciprocal ranks.

For our experiments we use two approaches to make recipient predictions. In our first experiment, we use the model itself to make recipient predictions by computing the mean of the hidden variable distribution given observed author and word features. We then compute the multinomial distribution obtained when conditioning upon this sample. The ordering of recipients produced by this distribution is then used to determine the MRR by finding the first predicted recipient also on the list of test email recipients. Our second experiment uses an approach based on computing the cosine similarity within the latent space for a given test email and all training set emails. Predictions are then obtained from recipients of the retrieved documents. Finally, our baseline method is a term frequency, inverse document frequency (TFIDF) based cosine similarity computation using the original document vectors and the same MRR computation as the latent space method.

We use the Enron email corpus with the processing described in [8]. The resulting corpus consists of 23,488 email messages sent among 147 users. Emails that were not received by at least one of the 147 users are not included. In order to capture only the new text entered by the author of a message, "quoted original messages" in replies were removed using some heuristic methods. Finally, to remove sensitivity to capitalization, all text was downcased. Finally, we randomly partition the data set into .9, .1 percent training and test sets for our experiments and use 200 topics.

For our first experiment we use the shared hidden variable structure of our DART model to learn a latent space model for predictions across all users in our training set. From figure 3 (top) we see that shared variable transfer learning increases MRR across the entire Enron corpus by 10% over the non-transfer TIDIF baseline. In our second experiment we investigate parameter prior based transfer learning as illustrated in figure 2. We first learn an organization wide model and then achieve transfer by using the parameters of this model as a non-zero mean Gaussian parameter prior for a model specifically trained for user 50. Figure 3 (bottom) compares a user specific model trained with transfer with a user specific model trained using a zero mean Gaussian prior. The TFIDF baseline is also given. From this analysis we see that the most dramatic benefits of transfer occur early in learning but benefits persist after many iterations of gradient based learning.

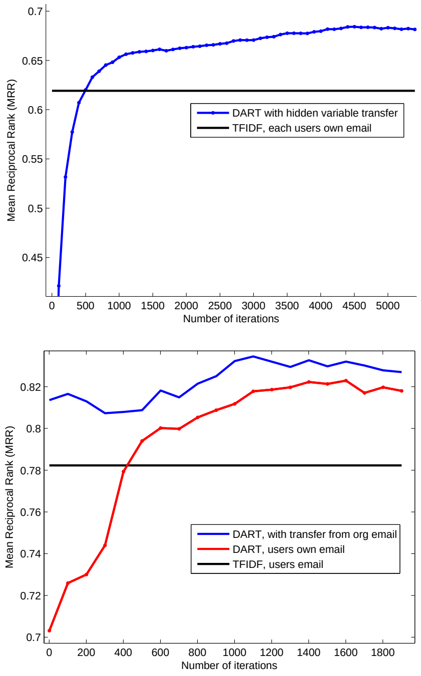

Figure 3: Transfer learning using: shared hidden variables (top), parameter priors (bottom).

In conclusion, we have shown how two types of transfer learning using DARTs confer significant advantages. Both shared hidden variables and prior based methods improve recipient prediction performance. In the latter case our experiments involved a user with over 2000 emails in their local training set. Early iterations in learning are analogous to situations where fewer training examples are available. We thus expect that transfer learning methods using these approaches could be most beneficial to models for new users.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, Microsoft Research under the Memex funding program and by a gift from Kodak Research. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, 2005.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[3] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[5] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, New York, NY, USA, 2004. ACM Press.

[6] N. Intrator and S. Edelman. Making a low-dimensional representation suitable for diverse tasks. In *In Learning to learn*. Kluwer, 1998.

[7] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2002.

[8] A. McCallum, A. Corrada-Emanuel, and X. Wang. Topic and role discovery in social networks. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2005.

[9] A. Niculescu-Mizil and R. Caruana. Learning the structure of related tasks. In *NIPS 2005 Workshop on Transfer Learning*, 2005.

[10] C. Pal and A. McCallum. CC prediction with grapical models. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, 2006.

[11] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 713–720, New York, NY, USA, 2006. ACM Press.

[12] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, 1990.