Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 5, N. 1, 2017.

Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Exponential Family Transfer Learning with Application to Text Document Modeling

Ricardo N. Rodrigues[1]
Centro de Ciências Computacionais, FURG, Rio Grande, RS, Brazil
Venu Govindaraju[2]
Center for Unified Biometrics, SUNY at Buffalo, Buffalo, NY, USA

Abstract. Transfer learning, as applied in machine learning, transfers knowledge between similar learning tasks with the objective of improving performance. We propose a method based on the prior selection principle that explores the transfer learning paradigm for estimating probability density functions belonging to the exponential family. Experiments on distribution of words over text documents, modeled as multinomial distributions, have shown better results when compared to maximum-likelihood estimation.

Keywords. Exponential family, transfer learning, document modeling.

1 Introduction

It is known that people learn more easily if they have previously learned similar or related information. Transfer learning, as studied in machine learning field, borrows ideas from psychology to develop computational methods that can explore the similarity between related tasks with the purpose of improving learning performance in computational tasks [6].

In this paper, we propose a transfer learning method for estimating exponential family probability distributions. A probability distribution function (pdf) belongs to the exponential family if it can be written in the following form:

$$p(\mathbf{x}|\boldsymbol{\eta}) = \exp\left\{\boldsymbol{\eta}^T\mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\eta}) + h(\mathbf{x})\right\}, \tag{1}$$

where $\mathbf{T}(\mathbf{x})$ is the vector of sufficient statistics, $\boldsymbol{\eta}$ is the vector of natural parameters, $h : \mathbf{x} \to \mathcal{R}$ is a base measure over $\mathbf{x}$ and $\psi(\boldsymbol{\eta})$ is the log partition function that assures that $\int p(\mathbf{x})d\mathbf{x} = 1$.

In this context, transfer learning refers to the parameter estimation of one or more target pdfs while leveraging the similarities from one or more source pdfs. Our approach is based on the prior selection principle, which leads to a prior distribution that incorporates the relation between pdfs. We present our method for both inductive transfer, where one

[1]ricardonagel@furg.br
[2]venu@cse.ub.edu

DOI: 10.5540/03.2017.005.01.0277

target task is learnt using information from other source tasks which solutions are known a priori, and for multi-task transfer, where several tasks are learnt simultaneously.

In Section 2 we present some related work in transfer learning. In Section 3 we describe the proposed method in detail. Experiments and their results are presented in Section 4. Finally, in Section 5, we present the conclusions and future works.

2 Related Work

Transfer learning algorithms have been proposed for several applications in classification [2], regression [11] and reinforcement learning [10]. For example, in [2] the AdaBoost classifier has been modified to make use of data coming from related classification tasks following a inductive transfer learning. So for example, if the classifier is being trained to classify a text document into classes 'automobile' and 'computer hardware', a related (source) classifier with classes 'motorcycle' and 'computer software' can be used to improve the classification performance.

Transfer learning for pdfs, as focused in this paper, has not been explicitly investigated, but it can be argued that some well know statistical machine learning methods implicitly apply transfer learning concepts when estimating pdf. Most of these methods are based on a Bayesian framework that has shown to be a natural option for transfer learning. In this framework, the prior information is encoded in terms of a common prior distribution over the model parameters of a group of pdfs or even a prior distribution over models. An example of such methods is Latent Dirichlet Allocation (LDA) [1], where the distribution over words for several documents (corpus) are learned together, allowing for transfer of knowledge about topic (i.e. category) distributions in the corpus. Folowing a similar line, [10] uses hierarchical Bayes models for multitask learning of Gaussian Processes.

Pan *et al.* [7] consider transfer learning via dimensionality reduction. They propose to learn a low-dimensional latent feature space where the distributions between the source domain data and the target domain data are the same or close to each other. However, their approach is quadratic on the number of training instances, which becomes problematic for large datasets. In [4] importance sampling is used for transferring samples from source to target domains. Recent surveys in the field can be found in [8] and [6].

3 Exponential Family Multi-task using the Prior Selection Principle

In this section, we first introduce the prior selection principle (section 3.1) and show how it can be used to fundamentally justify the usage of conjugate priors for exponential family inductive transfer learning (section 3.2). Then, we propose a new algorithm, called Iterative Multi-Task (section 3.3), that uses this same prior for implementing exponential family multi-task learning.

DOI: 10.5540/03.2017.005.01.0277

3.1 Prior Selection Principle

Suppose we have an set of k a priori guesses $\{\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_k\}$ for the estimation of an exponential family pdf. The prior selection principle [9] can be used to fundamentally incorporate this initial guesses into a prior distribution over the parameter space. The general idea is to find a prior distribution $\pi(\boldsymbol{\eta})$ that, on average, is similar to the a priori guesses but that is also not so different from a Jefrey's prior $g(\boldsymbol{\eta})$ representing our total ignorance about the optimal solution. The following functional formalizes these two objectives:

$$\mathcal{F}\left[\pi(\boldsymbol{\eta})\right] = \sum_{i=1}^{k} \lambda_i \int \pi(\boldsymbol{\eta}) KL(\boldsymbol{\eta}_i \| \boldsymbol{\eta}) d\boldsymbol{\eta} + \int \pi(\boldsymbol{\eta}) \log \frac{\pi(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} d\boldsymbol{\eta}, \tag{2}$$

where $KL(\boldsymbol{\eta}_i \| \boldsymbol{\eta})$ represents the Kullback-Liebler (KL) divergence between two pdf given by:

$$\begin{aligned} KL(\boldsymbol{\eta}_1 \| \boldsymbol{\eta}_2) &= \int_{x \in \mathcal{X}} p(x|\boldsymbol{\eta}_1) \log \frac{p(x|\boldsymbol{\eta}_1)}{p(x|\boldsymbol{\eta}_2)} dx, & (3) \\ &= \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) + (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)\boldsymbol{\mu}_1. & (4) \end{aligned}$$

Let $\boldsymbol{\mu}$ represent the expected parameter given by $\boldsymbol{\mu} = \nabla_{\boldsymbol{\eta}} \psi(\boldsymbol{\eta})$. By variational calculus, we find that the solution that minimizes 2 has the form:

$$\pi(\boldsymbol{\mu}) \propto \exp\left\{-\overline{\lambda} KL(\overline{\boldsymbol{\mu}} \| \boldsymbol{\mu})\right\} g(\boldsymbol{\mu}), \tag{5}$$

where:

$$\overline{\lambda} = \sum_{i=1}^{k} \lambda_i, \qquad \overline{\boldsymbol{\mu}} = \frac{1}{\overline{\lambda}} \sum_{i=1}^{k} \lambda_i \boldsymbol{\mu}_i. \tag{6}$$

We observe that the prior given by equation (5) is a conjugate prior [9]. The posterior distribution, after we observe n i.i.d samples $\{\mathbf{x}_j\}_{j=1}^{n}$ will have the same format but with parameters given by:

$$\overline{\lambda_p} = n + \overline{\lambda}, \qquad \overline{\boldsymbol{\mu}_p} = \frac{1}{\overline{\lambda_p}} \{\sum_{j=1}^{n} \mathbf{T}(\mathbf{x}_j) + \sum_{i=1}^{k} \lambda_i \boldsymbol{\mu}_i\}. \tag{7}$$

3.2 Inductive Pdf Transfer Learning

Our proposal for inductive pdf transfer learning consist of a direct usage of the prior given by equation (5), where the parameters $\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k\}$ are the source pdf parameters and $\{\lambda_1, ..., \lambda_k\}$ are their respective weights. Then the maximum a posteriori (MAP) solution is given by equation (7).

Our approach differs in two main ways from the standard hierarchical Bayes approach:

1. It is not assumed that the source parameters have been sampled from this prior. In the prior selection principle, the source pdfs are view as a priori "guesses" for the target pdf;

DOI: 10.5540/03.2017.005.01.0277

2. Each source pdf can have a different weight in the prior. In contrast, hierarchical Bayes models usually assume all tasks are i.i.d samples from the prior, and therefore, with same importance in the inference process.

3.3 Multi-task Pdf estimation

In the previous section, the source tasks parameters were assumed to be known a priori. In this section, we propose a new algorithm to solve the multitask problem, where the estimation for a set of target pdfs is done simultaneously. Our algorithm works by iteratively applying inductive transfer for one target pdf assuming the others are know. This algorithm shares some ideas with the Iterative Conditional Modes (ICM) algorithm [3], so we call it Iterative MultiTask (IMT).

Let $\mathcal{T} = \{p(x|\boldsymbol{\eta}_i)\}_{i=1}^{k}$ represent the set of target pdfs to be learned. Assume we have a training dataset $\mathcal{D}_i$ of i.i.d samples for each pdf $i = 1...k$. We assume the relation between each pdf is encoded in a weighted graph G, where the tasks are nodes and a weighted edge g_{ij} represent the weight of pdf i in the estimation of pdf j. If two pdfs are not related, then $g_{ij} = 0$. The IMT algorithm is described in details in the pseudo-algorithm 1 as follows:

Algorithm 1 Iterative MultiTask (IMT)

Require: A set of untrained pdfs $\mathcal{T} = \{p(x|\boldsymbol{\eta}_i)\}_{i=1}^{k}$; training datasets $\mathcal{D}_i$, $i = 1...k$; a weighted graph G.

Ensure: An estimation for the pdfs $\{\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_k\}$.

1: Initialize $\hat{\boldsymbol{\eta}}_i$, $i = 1...k$ via Maximum Likelihood.
2: **for** number of iterations **do**
3: **for all** target tasks in $\mathcal{T}$ **do**
4: Construct a prior according to equation (5) where $\hat{\boldsymbol{\eta}}_j$, $j \neq i$ are source references and g_{ij} are the weights.
5: Update $\hat{\boldsymbol{\eta}}_i$ by computing the MAP solution according to equation (7).
6: **end for**
7: **end for**

Similarly to the ICM algorithm, we do not provide any proof for the convergence (global or local) for the IMT algorithm, but the empirical results presented in the next section indicate local convergence.

4 Experiments on Text Documents

We apply our method to the learning of document topics. More specifically, our objective is to estimate a pdf over words $p(w|\boldsymbol{\eta}_t, t)$, $w \in \mathcal{W}$ for a given topic t, where $\mathcal{W}$ is the vocabulary of possible words. We use a multinomial distribution, which belongs to the exponential family, to model $p(w|\boldsymbol{\eta}_t, t)$.

We use the 20 Newsgroup dataset, that is a collection of 18,774 newsgroup text documents, organized into 20 different topics. Each document is represented as a "bag of

DOI: 10.5540/03.2017.005.01.0277

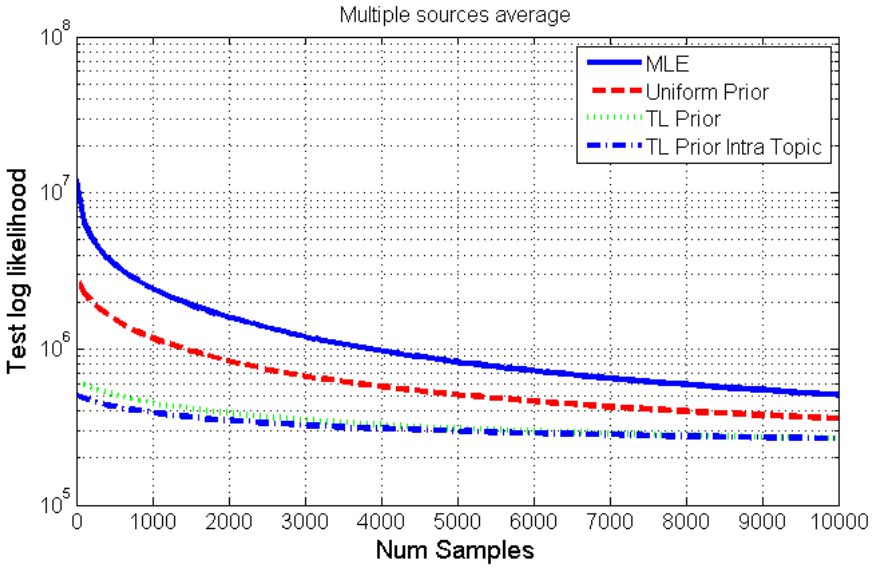

Figure 1: Multiple sources average performance. "Intra topic" transfer learning uses only source topics in the same category.

words" (i.e. the words position do not matter, only its frequency). 60% of the data is separated for training and 40% for testing. We filter the vocabulary by combining the 500 most popular words in each topic (based on training data). This gives an average of 106820 words per topic for training and 68331 for testing.

In our experiments, when a topic is a source task, we learn it via MLE with the full training data for that topic. Due to the large number training samples, we assume that this estimation has asymptotically approached the optimal solution. When a topic is a target, a relatively small subset of training samples is randomly selected from the full training dataset. Our main objective is to evaluate the target tasks performance according to the number of training samples (i.e. the target learning rate), in special, for cases when very few training samples are available. The performance of a given task is measured as the log likelihood on its respective test set.

4.1 Experiment 1: Inductive Transfer, Multiple Sources

In this experiments we evaluate the test likelihood on all 20 topics when multiple sources topics are present. We compare the case when only topics in the same category are used as sources (intra-topic) with the case when all topics are used as sources. The results are shown in figure 1. It can be seen that the intra-topic sources present slightly better results. This indicates that a selection in the source tasks, where only the relevant source tasks are picked, may improve transfer learning. Source selection has been proposed in [5] for classification multitask learning and is a option for minimizing the risk of negative transfer learning.

DOI: 10.5540/03.2017.005.01.0277

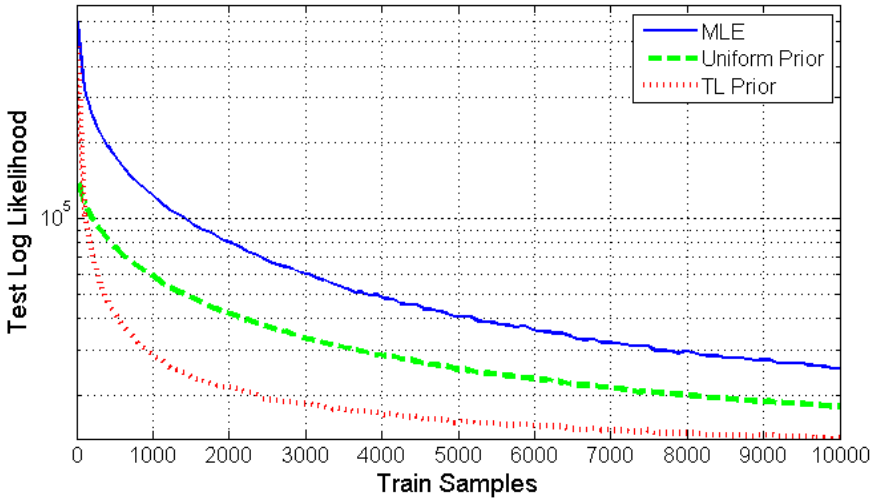

Figure 2: Multitask average performance.

4.2 Experiment 3: Multitask

In this experiment we apply the IMT algorithm to learn all topics simultaneously. Figure 2 shows the average performance over all topics. Note that when very few samples are available for each task (approximately less than 100), the IMT algorithm presents worse results than uniform prior. We conjecture that this happens because the estimation for such low numbers of training is so poor that transferring this information for other tasks is causing negative transfer. However, when training sample size increases, the IMT algorithm present significantly better results.

5 Conclusion

We presented new methods for inductive transfer learning and multi-task learning of pdfs belonging to the exponential family. The methods are based on the prior selection principle, and, in principle, can be applied to any exponential family pdf for which the maximum a posteriori solution (when using conjugate prior) can be computed. Experimental results on multinomial document topic pdf estimation have shown good results. Future work may include an automatic method to estimate weights and convergence analysis.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 193–200, New York, NY, USA, 2007. ACM.

[3] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society Series B Methodological*, 51(2):271–279, 1989.

[4] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association Computational Linguistics*, ACL '07, 2007.

[5] S. Kaski and J. Peltonen. Learning from relevant tasks only. In *Machine Learning: ECML 2007*. 2007.

[6] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[7] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1187–1192, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[8] L. Shao, F. Zhu, and X. Li. Transfer learning for visual categorization: A survey. *Neural Networks and Learning Systems, IEEE Transactions on*, 26(5):1019–1034, 2015.

[9] H. Snoussi and A. Mohammad-Djafari. Information geometry and prior selection. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 659, Mar. 2003.

[10] M. E. Taylor, P. Stone, and Y. Liu. Value functions for rl-based behavior transfer: a comparative study. In *AAAI'05: Proceedings of the 20th national conference on Artificial intelligence*, pages 880–885. AAAI Press, 2005.

[11] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 1012–1019, New York, NY, USA, 2005. ACM.

DOI: 10.5540/03.2017.005.01.0277