ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Sparse Nonparametric Topic Model for Transfer Learning

Ali Faisal[1], Jussi Gillberg[1], Jaakko Peltonen[1], Gayle Leen[2], and Samuel Kaski[1,3] *

1- Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University
2- deCODE genetics

3- Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki

Abstract. Count data arises for example in bioinformatics or analysis of text documents represented as word count vectors. With several data sets available from related sources, exploiting their similarities by *transfer learning* can improve models compared to modeling sources independently. We introduce a Bayesian generative transfer learning model which represents similarity across document collections by *sparse sharing of latent topics* controlled by an Indian Buffet Process. Unlike Hierarchical Dirichlet Process based multi-task learning, our model decouples topic sharing probability from topic strength, making sharing of low-strength topics easier, and outperforms the HDP approach in experiments.

1 Introduction

Classical machine learning methods learn models for data from a single data source. When few training samples are available for the learning task, methods may overfit or have too little information to infer complicated models. To gain more information for the learning task, *transfer learning* [1] methods transfer knowledge from earlier tasks to a new one, and *multi-task learning* [2] methods learn several tasks together from their respective data sets, exploiting their underlying relationships. When set in the probabilistic modeling framework, such approaches typically build a hierarchical model describing how model parameters vary among tasks; models for all tasks are then learned simultaneously.

We introduce a multi-task learning (transfer learning) method for an unsupervised learning problem: generative modeling of count data in multiple tasks, such as bag-of-words text documents from several collections. We model each data source with the topic model family [3]; with a nonparametric extension where both the number of topics and their strengths are learned from data. To model sharing of information among tasks, we allow topics to be shared among tasks. We use an Indian Buffet Process (IBP; [4]) to model how many topics are active overall and which topics each task uses to model its respective documents; we allow a further sparsity-inducing step to turn off some topics from each task.

*AF, JG and JP had equal contributions. Authors belong to AIRC, a CoE of the Academy of Finland (AoF). The work was supported by AoF decisions 123983 and 252845; Finnish Doctoral Programme in Computational Sciences and in part by PASCAL2 NoE, ICT 216886.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Finally we generate the strengths of active topics in each task from a Gamma prior. We use Bayesian inference (MCMC sampling) to infer the posterior over topics and make predictions about new documents as in any Bayesian model.

The most relevant earlier work is the Hierarchical Dirichlet Process model (HDP; [5]) which extends the single-task Latent Dirichlet Allocation model (LDA; [3]) and learns the number of topics from data by a Dirichlet Process (DP) prior; it is also extended to multi-task problems by modeling topic strengths in each task as draws from an upper-level Dirichlet process prior; we denote the multi-task version by MT-HDPLDA. MT-HDPLDA implicitly assumes that the topics most likely to be shared are also the strongest topics. This neglects the possibility of sharing weak topics. In contrast, our IBP-based sharing separates the choice of which topics to share from generation of topic strengths, allowing more flexible sharing between multiple tasks. In experiments our model outperforms MT-HDPLDA on several data domains. Another related model is the single-task model in [6], which uses an IBP prior to control which topics are active in each document and draws strengths of active topics from Gamma priors. The model in [6] is for single-task learning only. Our model can be seen as a multi-task counterpart, where the "IBP+Gamma" type generation of topic strengths is used across multiple tasks rather than across documents in one task.

Sections 2 and 3 describe our model and the experiments; Section 4 concludes.

2 Multi-task topic models

The basic single-task topic model Latent Dirichlet Allocation (LDA; [3]) generates a document through activity of latent topics; to generate a document, a topic distribution is drawn from a prior, and to generate each word in the document, a topic is drawn from the topic distribution and the word is drawn from a topic-wise word distribution. LDA assumes a fixed number of topics.

The Hierarchical Dirichlet Process (HDP; [5]) generalizes LDA to learn the number of topics from data and model multiple document collections (data sets), by Bayesian nonparametric inference. In an HDP, topics for a document are drawn from a Dirichlet process (DP), which in turn is drawn from a data set level DP, which can in turn be drawn from an overall DP across data sets. The topmost DP in the hierarchy determines which topics are active overall and their strengths; lower-level DPs choose among their parent-level active topics, varying their strengths by a stick-breaking construction to yield differing topic distributions at each branch of the hierarchy. When inferring topics from data, the topmost DP can activate new topics as well as change their strength; the HDP can thus infer the number of topics from data. See [5] for details. Since sharing is done by the topic strength hierarchy, with the stick-breaking construction the strongest topics (which generate many words overall) are the most likely to survive in several branches of the hierarchy and thus be shared across data sets; this can make the model a bad fit for multi-task problems with low-strength shared topics (topics discussed in many document collections but not at great length).

Recently a single-task model with more flexible topic sharing was proposed

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[6] using an Indian Buffet Process Compound Dirichlet Process prior which can be seen as a spike-and-slab prior over topic strengths. An Indian Buffet Process prior is placed on binary flags of whether topics are present in documents rather than on their strengths which are generated separately from Gamma variables; the model then avoids the coupling of topic strength and topic sharing implicit in the HDP model. The model is for single-task learning only; when only few data are available from each data source, a multi-task solution is needed.

2.1 New sparse nonparametric topic model for transfer learning

We present a new hierarchical Bayesian multi-task (transfer learning) model which allows flexible sharing of low-strength and high-strength topics across multiple data sets, with a spike-and-slab prior. Learning the model for each data set is called a task; our model performs transfer learning by learning the tasks together. We draw the binary matrix of which topics are present in each task from an Indian Buffet Process (IBP); to draw a topic for a new task, the IBP chooses one of the existing topics according to how many tasks they are already present in, or activates a new topic, hence the number of active topics is inferred from data. Since we empirically found that IBP did not provide enough sparsity for our small data, we implemented an additional sparsity masking step turning off some components in each task. The strength of remaining active topics is drawn from Gamma distribution within each task; from this task-specific topic prior, the remaining generation proceeds as in LDA, drawing document-specific topic distributions and then the words for each document.

The full generative scheme (see Figure 1) is as follows: **Step 1.** Draw a binary matrix $\mathbf{B} \sim \text{IBP}(\alpha)$. The c^{th} row b_c of $\mathbf{B}$ tells which topics are active in task c. Draw topic strength prior $\gamma^{(k)} \sim \text{Gamma}(a_1, a_2)$ for each component $k = 1, 2, ...\infty$. **Step 2.** For each task c, sample an additional topic sparsity masking $\psi_c^{(k)} \sim \text{Bernoulli}(\epsilon)$ which turns off some topics k from b_c, then draw topic strength $\phi_c^{(k)} \sim \text{Gamma}(\gamma^{(k)}, 1)$. Draw the size of the task (total number of words) as a negative binomial $n_c^{(.)} \sim \text{NB}(\sum_k b_c^{(k)} \phi_c^{(k)} \psi_c^{(k)}, \frac{1}{2})$. **Step 3.** Draw the topic-to-word distributions $\boldsymbol{\beta}_k \sim \text{Dirichlet}(\eta)$. **Step 4.** For every document $d = 1, 2, ...D_c$ in task c, draw the distribution over topics $\boldsymbol{\theta}_{c,d} \sim \text{Dirichlet}(\mathbf{b}_c.\boldsymbol{\phi}_c.\boldsymbol{\psi}_c)$ where $\mathbf{b}_c$,$\boldsymbol{\phi}_c$ and $\boldsymbol{\psi}_c$ are multiplied elementwise; then for each word n, draw topic index $z_{c,d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_{c,d})$ and term $w_{c,d,n} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{c,d,n}})$. Here α, ϵ, η, a_1 and a_2 are the model hyperparameters. Inference by sampling, discussed next, is efficient and only processes a finite number of topics at each step as is usual in nonparametric models.

2.2 Bayesian inference for our model

To infer our model from the multi-task data sets, we use a combination of collapsed Gibbs sampling and the Metropolis-Hastings algorithm to sample from the posterior distribution of the model parameters. We integrate out the topic specific distribution over words, the topic mixture distribution and the binary IBP matrix; sampling is needed only over the remaining variables. In the Gibbs

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

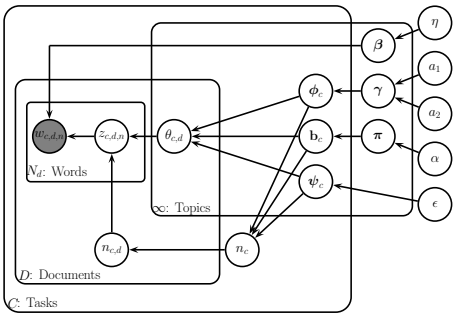

Fig. 1: Plate diagram for our sparse transfer learning topic model.

we cyclically sample the topic assignment z, the topic strength ϕ and the IBP prior π (stick-breaking parameters) for topic activation. To sample topic assignmentswe integrate out the topic distribution. The posterior that the n^{th} word comes from topic k is $p(z_{c,d,n} = k|\mathbf{z}_{\setminus c,d,n}, w_{c,d,n}, \Delta) \propto (n^{(k)}_{w_{c,d,n},\setminus c,d,n} + \eta) \int d\boldsymbol{\theta}_{c,d}\, p(z_{c,d,n} = k|\boldsymbol{\theta}_{c,d}) p(\boldsymbol{\theta}_{c,d}|\mathbf{z}_{\setminus c,d,n}, \Delta)$ where$\Delta = \{\boldsymbol{\phi}^{\bullet}_c, \boldsymbol{\pi}^{\bullet}, \boldsymbol{\gamma}, \alpha, \epsilon\}$ and superscript $^{\bullet}$ denotes active topics. The likelihood involves a combinatorial integration over values of the sparse IBP matrix, but since we only need the posterior for taking topic k, the integral simplifies to $E[\boldsymbol{\theta}^{(k)}_{c,d}|\mathbf{z}_{\setminus c,d,n}, \Delta] \propto E[((n^{(k)}_{c,d,\setminus c,d,n} + \phi^{(k)}_c) b^{(k)}_c \psi^{(k)}_c)/(n^{(\cdot)}_{c,d,\setminus c,d,n} + \sum_j b^{(j)}_c, \psi^{(j)}_c, \phi^{(j)}_c)]$. While not combinatorial, this expectation is inefficient to evaluate in closed form as we would need to do so for every word in Gibbs sampling. We use an approximation similar to [6], using 1st order Taylor expansion for the three possible cases: topic k is active in the current task (data set); topic k does not appear in the current task but is active in the corpus (all data sets); or topic k is inactive in the whole corpus. We must process inactive topics in case the sampling activates one: a topic is inactive (denoted by a superscript $^{\circ}$) if it is never used in the whole corpus, i.e. $n^{k}_{(\cdot)(\cdot)} = 0$ even if $\sum_c b^{(k)}_c > 0$, and active otherwise. Inactive topics have an ordering of decreasing stick lengths: $P(\pi^{\circ}_k|\pi^{\circ}_{k-1}, \mathbf{z}_{k:k>K^{\dagger}} = 0) \propto \exp(\sum_{i=1}^{N} \frac{1}{i}(1-\pi^{\circ}_k)^i)\pi^{\circ}_k(1-\pi^{\circ}_k).\mathbb{I}(0 \leq \pi^{\circ}_k \leq \pi^{\circ}_{k-1})$, where $K^{\dagger}$ is an index such that all active topics have index $k < K^{\dagger}$. Stick parameters for the inactive topics are sampled using the above equation by adaptive rejection sampling[1] [7]. For active topics we sample the stick parameters $\pi^{\bullet}_k$ underlying the IBP matrix by semi ordered stick breaking [8]: $p(\pi^{\bullet}_k|\mathbf{B}) \sim \text{Beta}(\sum_{c=1}^{C} b^{(k)}_c, 1 + C - \sum_{c=1}^{C} b^{(k)}_c)$. Even though topic assignments can be sampled while integrating over the binary IBP matrix, the IBP matrix is still required here for sampling the stick parameters for the active topics. The current value of the IBP matrix is reinstantiated according to

[1]Multiple samples were generated and an average was used to get a better approximation.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

$p(b_c^{(k)} = 1|\pi_{(.)}^{(k)}, \boldsymbol{\phi}_k, \psi_c^{(k)}, \mathbf{z}_c)$ which has value 1 if $n_{c,(.)}^{(k)} > 0$, π_k if $n_{c,(.)}^{(k)} = 0$ and $\psi_c^{(k)} = 0$, and $\pi^{(k)}/(\pi^{(k)}+2^{\phi_c^{(k)}}(1-\pi^{(k)}))$ if $n_{c,(.)}^{(k)} = 0$ and $\psi_c^{(k)} = 1$. The masking vector $\boldsymbol{\psi}_c$ is initialized by a similar equation as the IBP matrix by interchanging $b_c^{(k)}$ with $\boldsymbol{\psi}_c^{(k)}$ and π_k with ϵ.

Lastly, for the topic strength parameters, the joint probability of $\phi_c^{(k)}$ and total number of counts assigned to topic k is $p(\phi_c^{(k)}, n_c^{(k)}|\gamma^{(k)}, b_c^{(k)}, \psi_c^{(k)}) =$
$((\phi_c^{(k)})^{\gamma^{(k)}-1} e^{-\phi_c^{(k)}}/\Gamma(\gamma^{(k)})) \prod_{c:b_c^{(k)}.\psi_c^{(k)}=1}^{C} \Gamma(n_c^{(k)}+\phi_c^{(k)}/(\Gamma(\phi_c^{(k)})\, n_c^{(k)}!\, 2^{(\phi_c^{(k)}+n_c^{(k)})}))$.
We use Metropolis-Hastings to compute the posterior, and sample $\gamma^{(k)}$ in a similar manner from the joint posterior for $\gamma^{(k)}$ and $\phi_{(.)}^{(k)}$.

3 Empirical results

We compare our model to the nearest method Hierarchical Dirichlet Process based multi-task learning (MT-HDPLDA).

Experiment 1: Continuum of problem domains. We expect our model to perform well in the case of multi-task problems where both overall strong and non-strong topics are shared; we build a continuum of multi-task problem domains where this situation occurs. At either end of the continuum, data is generated from a model where shared topics are strong (generate many words overall); the left end is a simpler case where both models can work well, and the right end is a complicated case especially suitable for MT-HDPLDA. Interesting domains lie between the two ends: in these intermediate domains, the topic generation mechanisms from either end are mixed together linearly, yielding small shared topics from both generators. We create nine domains across the continuum, identified by the mixing coefficient (0 to 1) between the generators.

Each problem domain is a multi-task scenario where each learning problem has 10 tasks (data sets). We use the setting where one task is more interesting than others; the interesting task has 24 documents with 8 words each, other tasks have 8 documents with 8 words each, all generated from 10 topics with a vocabulary of 150 words. We generate 10 such learning problems in each domain and run our method and MT-HDPLDA on each problem (for sampling, we use 1500 iterations initial burnin, then draw 100 samples 15 iterations apart); results are evaluated by predictive likelihood on held-out documents from the interesting task.Figure 2(left) shows that in the intermediate domains where weak topics are shared in the interesting task, we outperform MT-HDPLDA.

Experiment 2: NIPS data. We take the five most frequent sections of NIPS articles from 1987 to 1999 (http://www.gatsby.ucl.ac.uk/∼ywteh); in total they contain 1147 documents with vocabulary size 1321 and average document length ∼ 950 words. The most frequent group is "Algorithms and Architecture", which we choose as the interesting task. We run our model, MT-HDPLDA and single-task HDP as a baseline; we follow [5] for MT-HDPLDA, set $\eta = 0.5$ in both models, and $\alpha = 5$ and $\gamma \sim Gamma(5, 0.1)$ for ours; for sampling we initialize the Gibbs samplers randomly, take 1000 burn-in iterations, and then draw a total

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

of 10 samples 50 iterations apart. We learn models for different sizes of training data in the interesting task (5-40 documents) with 50 documents in each other task, and use 5-fold cross-validation in each case. Results are again evaluated by average predictive log-likelihood of held-out documents from the interesting task. Figure 2(right) shows the results: single-task learning naturally works poorly, and our model outperforms MT-HDPLDA in the challenging scenarios where training data is small and hence multi-task learning is most needed.

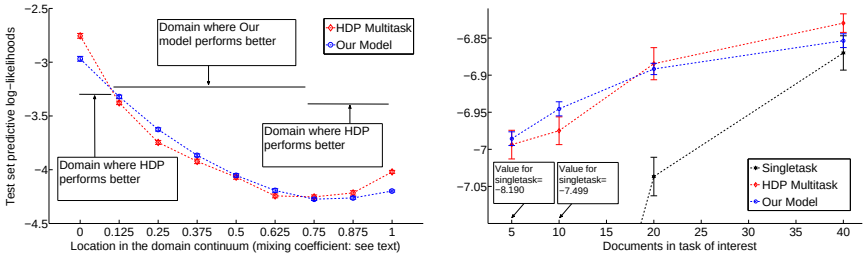

Fig. 2: Experiment results. Left: test set predictive likelihoods for simulated data continuum, error bars are over 10 random datasets. Right: test set predictive likelihoods for NIPS articles, error bars are over 5 folds.

4 Conclusions

We have introduced a sparse multi-task topic model that is a robust and flexible method to model sharing of strong and weak topics in multiple tasks. The proposed non-parametric model outperforms a state of the art HDP based topic model on a simulated data continuum and on real data with small training sets.

References

[1] S. Thrun. Is learning the n-th thing any easier than learning the first. In *Advances in Neural Information Processing Systems*, pages 640–646. The MIT Press, 1996.

[2] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

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

[4] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, 2006.

[5] Y. W. Teh, M. I Jordan, M. J Beal, and D. M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[6] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1151–1158. Omnipress, 2010.

[7] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41:337–348, 1992.

[8] Y. W. Teh. Stick-breaking construction for the indian buffet process. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1–10, 2007.