Active Multitask Learning with Doubly Supervised Latent Dirichlet Allocation

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces two models – Doubly Supervised Latent Dirichlet Allocation (DSLDA) that makes use of both shared latent and *supervised* topics to accomplish multitask learning (MTL) and Active Doubly Supervised Latent Dirichlet Allocation (Act-DSLDA) that integrates MTL and active learning in the same framework. Experimental results on both document and image classification show that integrating MTL and active learning along with shared latent and supervised topics is superior to other methods which do not use all of these components.

1 Introduction

Research in computer vision for designing an automated object detector has primarily been focused on *either* gathering large datasets of web images [1, 2] *or* by formulating new algorithms that can reduce the degree of human intervention in the learning process. In one of the learning methodologies, *shared* attributes, abstract descriptors of object properties [3, 4, 5] are used to serve as an intermediate layer in a classifier cascade. If the *shared* attributes transcend object class boundaries, such a classifier cascade is beneficial for *transfer learning* [6]. Another group of researchers have formulated methods based on active learning for reducing the expense of human annotations where the system can request labels for the most informative examples [7, 4]. In this paper, our objective is to combine these two orthogonal approaches in order to leverage the benefits of both – learning from a shared feature space and making active queries.

2 Related Work

Unsupervised LDA has been extended to account for supervision. In *Labeled LDA* (LLDA [8]), the primary objective is to build a model of the words that indicate the presence of certain topic labels. Some other researchers [9, 10, 11] assume that supervision is provided for a single *response variable* to be predicted for a given document. In *Maximum Entropy Discriminative LDA* (MedLDA) [10], the objective is to infer some low-dimensional (topic-based) representation of documents which is predictive of the response variable. Transfer learning allows the learning of some tasks to benefit the learning of others through either simultaneous [12] or sequential [13] training. In multitask learning (MTL [12]), a single model is simultaneously trained to perform multiple related tasks (*e.g.*, see [6, 14]). Finally, there has been some effort to integrate active and transfer learning in the same framework [15, 16, 17] . However, none of these approaches deal with topic models or query over both supervised topic labels and class labels

3 Doubly Supervised Latent Dirichlet Allocation (DSLDA)

Assume we are given a training corpus consisting of N documents belonging to Y different classes (where each document belongs to exactly one class and each class corresponds to a different task). Further assume that each of these training documents is also annotated with a set of K_2 different topic "tags" (henceforth referred to as "supervised topics"). The objective is to train a model using the words in a data, as well as the associated supervised topic tags and class labels, and then use this model to classify completely unlabeled test data for which no topic tags nor class labels are provided. The DSLDA model is now described below.

• For the n^{th} document, sample a topic selection probability vector $\boldsymbol{\theta}_n \sim \text{Dir}(\boldsymbol{\alpha}_n)$, where $\boldsymbol{\alpha}_n = \boldsymbol{\Lambda}_n \boldsymbol{\alpha}$ and $\boldsymbol{\alpha}$ is the parameter of a Dirichlet distribution of dimension K, which is the total number of topics. The topics are assumed to be of two types – latent and supervised, and there are K_1 latent topics and K_2 supervised topics. Therefore, $K = K_1 + K_2$. Latent topics are never observed, while supervised topics are observed in training but not in test data. Henceforth, in each vector or matrix with K components, it is assumed that the first K_1 components correspond to the latent topics and the next K_2 components to the supervised topics. $\boldsymbol{\Lambda}_n$ is a diagonal binary matrix of dimension $K \times K$. The k^{th} diagonal entry is unity if *either* $1 \leq k \leq K_1$ *or* $K_1 < k \leq K$ and the n^{th} document is tagged with the k^{th} topic. Also, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ where $\boldsymbol{\alpha}_1$ is a parameter of a Dirichlet distribution of dimension K_1 and $\boldsymbol{\alpha}_2$ is a parameter of a Dirichlet distribution of dimension K_2.

• For the m^{th} word in the n^{th} document, sample a topic $z_{nm} \sim \text{multinomial}(\boldsymbol{\theta}'_n)$, where $\boldsymbol{\theta}'_n = (1-\epsilon)\{\boldsymbol{\theta}_{nk}\}_{k=1}^{k_1} \epsilon\{\Lambda_{n,kk}\boldsymbol{\theta}_{nk}\}_{k=1+k_1}^{K}$. This implies that the supervised topics are weighted by ϵ and the latent topics are weighted by $(1-\epsilon)$. Sample the word $w_{nm} \sim \text{multinomial}(\boldsymbol{\beta}_{z_{nm}})$, where $\boldsymbol{\beta}_k$ is a multinomial distribution over the vocabulary of words corresponding to the k^{th} topic.

• For the n^{th} document, generate $Y_n = \arg\max_y \boldsymbol{r}_y^T \mathbb{E}(\bar{\mathbf{z}}_n)$ where Y_n is the class label associated with the n^{th} document, $\bar{\mathbf{z}}_n = \sum_{m=1}^{M_n} \mathbf{z}_{nm}/M_n$. Here, $\mathbf{z}_{nm}$ is an indicator vector of dimension K. $\boldsymbol{r}_y$ is a K-dimensional real vector corresponding to the y^{th} class, and it is assumed to have a prior distribution $\mathcal{N}(0, 1/C)$. M_n is the number of words in the n^{th} document. The maximization problem to generate Y_n (or the classification problem) is carried out using a max-margin principle.

Note that predicting each class is effectively treated as a separate task, and that the shared topics are useful for generalizing the performance of the model across classes. In particular, when all classes have few training examples, knowledge transfer between classes can occur through the shared topics.

Let us denote the hidden variables by $\boldsymbol{Z} = \{\{z_{nm}\}, \{\boldsymbol{\theta}_n\}\}$, the observed variables by $\boldsymbol{X} = \{w_{nm}\}$ and the model parameters by $\boldsymbol{\kappa}_0$. To avoid computational intractability, inference and estimation are performed using Variational **EM** using a completely factorzied approximation $q(\mathbf{Z})$. With the use of the lower bound obtained by the factorized approximation, followed by Jensen's inequality, DSLDA reduces to solving the following optimization problem[1]:

$$\min_{q,\boldsymbol{\kappa}_0,\{\xi_n\}} \frac{1}{2}||\boldsymbol{r}||^2 - \mathcal{L}(q(\boldsymbol{Z})) + C\sum_{n=1}^{N}\xi_n, \text{ s.t. } \forall n, y \neq Y_n : \mathbb{E}[\boldsymbol{r}^T \Delta f_n(y)] \geq 1 - \xi_n; \xi_n \geq 0. \quad (1)$$

Here, $\Delta f_n(y) = f(Y_n, \bar{\boldsymbol{z}}_n) - f(y, \bar{\boldsymbol{z}}_n)$ and $\{\xi_n\}_{n=1}^N$ are the slack variables, and $f(y, \bar{\boldsymbol{z}}_n)$ is a feature vector whose components from $(y-1)K+1$ to yK are those of the vector $\bar{\boldsymbol{z}}_n$ and all the others are 0. $\mathbb{E}[\boldsymbol{r}^T \Delta f_n(y)]$ is the "expected margin" over which the true label Y_n is preferred over a prediction y. From this viewpoint, DSLDA projects the documents onto a combined topic space and then uses a max-margin approach to predict the class label. The parameter C penalizes the margin violation of the training data.We skip the update equations here and refer the reader to [5] instead.

4 Active Doubly Supervised Latent Dirichlet Allocation (Act-DSLDA)

In the active learning setting, the model has to be changed slightly. We first state the notations used here. Suppose we are given an initial training corpus $\mathcal{L}$ with N documents belonging to Y different classes. When the learning starts, $\mathcal{L}$ is assumed to have fully labeled documents. However, as the learning progresses more documents are added to the pool $\mathcal{L}$ with class and/or a subset of supervised topics labeled. Therefore, at any intermediate point of the learning process, $\mathcal{L}$ can be assumed to contain several sets: $\mathcal{L} = \{\mathcal{T} \cup \mathcal{T}_C \cup \mathcal{T}_{A_1} \cup \mathcal{T}_{A_2} \cup \cdots \cup \mathcal{T}_{A_{K_2}}\}$, where $\mathcal{T}$ contains fully labeled documents (*i.e.* with both class and all of supervised topics labeled) and $\mathcal{T}_C$ represents the documents that have class labels. For $1 \leq k \leq K_2$, $\mathcal{T}_{A_k}$ represents the documents that have the k^{th} supervised topic labeled. Since, human provided annotations and class labels are expensive to obtain in general, we design an active learning framework where the model can query over an unlabeled pool $\mathcal{U}$ and request either class labels or a subset of the supervised topics. We use expected error reduction [18] as a measure in active selection. Such active selection mechanism is less immune to noise compared to uncertainty sampling [19] but requires the model parameters to be incrementally updated.

In the test data, the supervised topics are not observed and one has to infer them from either the parameters of the model or use some other auxiliary information. Since one of our objectives is to query over the supervised topics as well as the final category, we train a set of binary SVM classifiers that can predict the individual attributes from the features of the data. We denote the parameters of such classifiers by $\{\boldsymbol{r}_{2k}\}_{K_1<k\leq K}$. This is important to get an uncertainty measure over the supervised topics. To further clarify the issue, let us consider that only one supervised topic has to be labeled by the annotator for the n^{th} document from the set of supervised topics of size K_2. To select the most uncertain topic, one needs to compare the uncertainty of predicting the presence or absence of the individual topics. This uncertainty is different from that calculated from the conditional distribution which one might be tempted to calculate from the posterior over θ_n.

We change the notation slightly from DSLDA and denote by $\boldsymbol{r}_{1y}$ the K-dimensional real vector corresponding to the y^{th} class, and it is assumed to have a prior distribution $\mathcal{N}(0, 1/C)$. The maximization problem to generate Y_n (or the classification problem) is carried out using a max-margin principle and we use online support vector machines [20] for such update. Since the model has to be updated incrementally in the active selection step, a batch SVM solver is not applicable. Online SVM allows one to update the learnt weights incrementally given a new example.

Inference and parameter estimation in Act-DSLDA have two phases – one for the batch case when the model is trained with some labeled data and the other is for the active selection step where the model has to be incrementally updated to observe the effect of any labeled information that is queried from the oracle. In the batch mode, Act-DSLDA reduces to solving the

[1] Please see [10, 5] for further details.

following optimization problem:

$$\min_{q,\boldsymbol{\kappa}_0,\{\xi_n\}} \frac{1}{2}||\boldsymbol{r}_1||^2 - \mathcal{L}(q(\boldsymbol{Z})) + C\sum_{n=1}^{N}\xi_n \mathbb{I}_{\mathcal{T}_C,n},\ \text{s.t.}\ \forall n \in \mathcal{T}_C, y \neq Y_n : \mathbb{E}[\boldsymbol{r}_1^T \Delta f_n(y)] \geq 1 - \xi_n; \xi_n \geq 0. \quad (2)$$

The only difference in this objective from that of DSLDA is the presence of the indicator variable $\mathbb{I}_{\mathcal{T}_C,n}$ which is unity if the n^{th} document has a class label (*i.e.* $n \in \mathcal{T}_C$) and 0 otherwise. This implies that only the documents which have class labels are used for updating the parameters of online SVM. Rest of the updates are similar to DSLDA.

For active selection, consider that a completely unlabeled or partially labeled document, indexed by n', is to be included in the labeled pool with one of the (K_2+1) labels (one for the class label and each different supervised topic), indexed by k'. In the E step, variational parameters corresponding to all other documents except for the n'th one is kept fixed and the variational parameters for only the n'th document are updated. In the M-step, we keep the priors $\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2\}$ over the topics and the SVM parameters $\boldsymbol{r}_2$ fixed as there is no easy way to update such parameters incrementally. From the empirical point of view, these parameters do not change much w.r.t. the variational parameters or features of a single document. However, the update of the parameters $\{\boldsymbol{\beta}, \boldsymbol{r}_1\}$ is easier. Updating $\boldsymbol{\beta}$ is accomplished by a simple update of the sufficient statistics. Updating $\boldsymbol{r}_1$ is done using the "ProcessNew" operation of online SVM followed by a few iterations of "ProcessOld".

5 Experimental Evaluation

Our evaluation used two datasets, a text corpus consisting of abstracts from ACM conferences and a multi-class image database named aYahoo [3]. Please see [5] for more details about these datasets. In order to demonstrate the contribution of each aspect of the overall model, DSLDA is compared against the following simplified models – 1. MedLDA with **o**ne-**v**s-**a**ll classification (MedLDA-OVA), 2. MedLDA with **m**ulti**t**ask **l**earning (MedLDA-MTL), 3. DSLDA with **o**nly **s**hared **s**upervised **t**opics (DSLDA-OSST), 4. DSLDA with **n**o **s**hared **l**atent **t**opics (DSLDA-NSLT), 5. **M**ajority **c**lass **m**ethod (MCM). We skip the rationales for using such baselines here to save space and details are available in [5].

Figs. 1 and 2 present representative learning curves for the image and text data respectively, showing how classification accuracy improves as the amount of class supervision (p_2) is increased. Results are shown for two different amounts of topic supervision ($p_1 = 0.5$ and $p_1 = 0.7$). The error bars in the curves show standard deviations across the 50 trials.

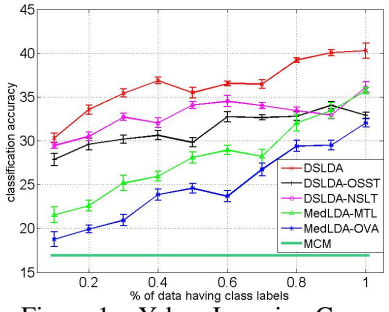

Figure 1: aYahoo Learning Curves

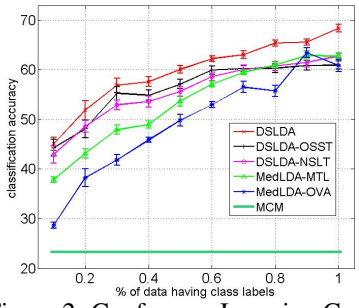

Figure 2: Conference Learning Curves

We also compare Act-DSLDA against the following models: 1. Active Learning in MedLDA with **o**ne-**v**s-**a**ll classification (Act-MedLDA-OVA) – a separate MedLDA model is trained for each class using a one-vs-all approach leaving no possibility of transfer across classes; 2. Active Learning in MedLDA with **m**ulti**t**ask **l**earning (Act-MedLDA-MTL)– a single MedLDA model is learned for all classes where the latent topics are shared across classes (this baseline is supposed to be stronger than baseline 1 where the latent topics are not shared); 3. Act-DSLDA with **o**nly **s**hared **s**upervised **t**opics (Act-DSLDA-OSST)– a model in which supervised topics are used and shared across classes but there are no latent topics (both the supervised topics and the class labels are queried using active selection strategy); 4. Act-DSLDA with **n**o **s**hared **l**atent **t**opics (Act-DSLDA-NSLT) – a model in which only supervised topics are shared across classes and a separate set of latent topics is maintained for each class (both the supervised topics and the class labels are queried using active selection strategy); 5. **R**andom **s**election of only **c**lass labels (RSC) – a MedLDA-MTL model where only the class labels are selected at random but the supervised topics are not used at all [2]; (this baseline shows the utility of active selection of classes in MedLDA-MTL framework); 6. **R**andom **s**election of **c**lass and **a**ttribute labels (RSCA) – a DSLDA model where both the class and the supervised topics are selected at random (this baseline is weaker than RSC since the supervised topics are less informative compared to the class labels).

For experiments with both image and text data in Act-DSLDA, we start with a completely labeled dataset $\mathcal{L}$ consisting of 300 documents. In every active iteration, we query for 50 labels (class labels or supervised topics). Figs. 3 and 4 present representative learning curves for the image and the text data respectively, showing how classification accuracy improves as the amount of supervision is increased. The error bars in the curves show standard deviations across the 20 trials.

[2] Note that designing a DSLDA based model where only class labels are selected at random is tricky as one needs to balance the number of supervised topics queried and the number of class labels selected at random.

Overall, the results support the hypothesis that DSLDA's ability to incorporate both supervised and latent topics allow it to achieve better predictive performance compared to baselines that exploit only one, the other, or neither. Similarly, Act-DSLDA quite consistently outperforms all of the baselines, clearly demonstrating the advantage of combining both types of topics and integrating active learning and transfer learning in the same framework.

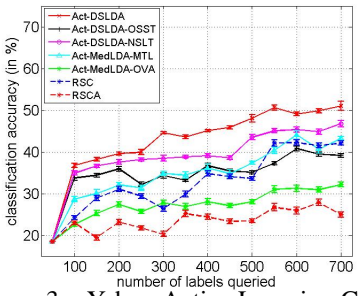

Figure 3: aYahoo Active Learning Curves

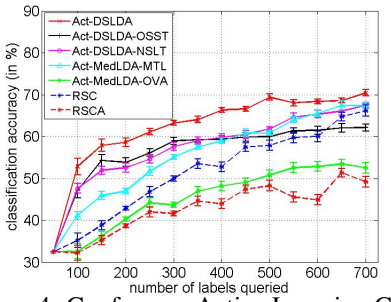

Figure 4: Conference Active Learning Curves

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*, pages 248–255, 2009.

[2] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation, 2008.

[3] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. of CVPR*, pages 1778–1785, 2009.

[4] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *Proc. of ICCV*, pages 1403–1410, 2011.

[5] A. Acharya, A. Rawal, R. J. Mooney, and E. R. Hruschka. Using both supervised and latent shared topics for multitask learning. In *ECML PKDD, Part II, LNAI 8189*, pages 369–384, 2013.

[6] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[7] A.J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *Proc. of CVPR*, pages 2372–2379, 2009.

[8] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of EMNLP*, pages 248–256, 2009.

[9] D. M. Blei and J. D. Mcauliffe. Supervised topic models. In *Proc. of NIPS*, 2007.

[10] J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: maximum margin supervised topic models for regression and classification. In *Proc. of ICML*, pages 1257–1264, 2009.

[11] J. Chang and D. Blei. Relational topic models for document networks. In *Proc. of AISTATS*, 2009.

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

[13] K. D. Bollacker and J. Ghosh. Knowledge transfer mechanisms for characterizing image datasets. In *Soft Computing and Image Processing*. Physica-Verlag, Heidelberg, 2000.

[14] A. Passos, P. Rai, J. Wainer, and H. Daumé III. Flexible modeling of latent task structures in multitask learning. In *Proc. of ICML*, pages 1103–1110, 2012.

[15] P. Rai, A. Saha, H. Daumé, III, and S. Venkatasubramanian. Domain adaptation meets active learning. In *Proc. of NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.

[16] A. Harpale and Y. Yang. Active learning for multi-task adaptive filtering. In *Proc. of ICML*, pages 431–438. Omnipress, 2010.

[17] A. Saha, P. Rai, H. Daum III, and S. Venkatasubramanian. Online learning of multiple tasks and their relationships. *JMLR - Proceedings Track*, 15:643–651, 2011.

[18] N. Roy and A. K. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of ICML*, pages 441–448, 2001.

[19] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[20] A. Bordes, L. Bottou, P. Gallinari, and J. Weston. Solving multiclass support vector machines with larank. In *Proc. of ICML*, pages 89–96, 2007.