Hetero-Labeled LDA: A Partially Supervised Topic Model with Heterogeneous Labels

Dongyeop Kang[1,⋆], Youngja Park[2], and Suresh N. Chari[2]

[1] IT Convergence Laboratory, KAIST Institute, Daejeon, South Korea
[2] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
dykang@itc.kaist.ac.kr, {young_park,schari}@us.ibm.com

Abstract. We propose Hetero-Labeled LDA (*h*LLDA), a novel semi-supervised topic model, which can learn from multiple types of labels such as document labels and feature labels (i.e., heterogeneous labels), and also accommodate labels for only a subset of classes (i.e., partial labels). This addresses two major limitations in existing semi-supervised learning methods: they can incorporate only one type of domain knowledge (e.g. document labels or feature labels), and they assume that provided labels cover all the classes in the problem space. This limits their applicability in real-life situations where domain knowledge for labeling comes in different forms from different groups of domain experts and some classes may not have labels. *h*LLDA resolves both the label heterogeneity and label partialness problems in a unified generative process.

*h*LLDA can leverage different forms of supervision and discover semantically coherent topics by exploiting domain knowledge mutually reinforced by different types of labels. Experiments with three document collections–*Reuters*, *20 Newsgroup* and *Delicious*– validate that our model generates a better set of topics and efficiently discover additional latent topics not covered by the labels resulting in better classification and clustering accuracy than existing supervised or semi-supervised topic models. The empirical results demonstrate that learning from multiple forms of domain knowledge in a unified process creates an enhanced combined effect that is greater than a sum of multiple models learned separately with one type of supervision.

1 Introduction

Motivated by a diverse set of requirements such as information management and data security, there is an increasing need for large scale topic classification in large distributed document repositories. In these environments, documents are generated and managed independently by many different divisions and domain experts in the company. Often, it is prohibitively expensive to perform supervised topic classification at an enterprise scale, because it is very challenging to catalog what topics exist in the company let alone provide labeled samples for all the topics.

In recent years, probabilistic topic modeling, most notably Latent Dirichlet Allocation (LDA) has been widely used for many text mining applications as an alternative to expensive supervised learning approaches. Probabilistic topic modeling approaches can

⋆ This work was conducted while the author was an intern at IBM Research.

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 640–655, 2014.

discover underlying topics in a collection of data without training a model with labeled samples. However, unsupervised topic modeling relies primarily on feature (word) occurrence statistics in the corpus, and the discovered topics are often determined by dominant collocations and do not match with the true topics in the data.

A more realistic approach would be to use a semi-supervised learning in which the topic discovery process is guided by some form of domain knowledge. In recent years, many extensions to LDA, in both supervised and semi-supervised ways, have been proposed to generate more meaningful topics incorporating various side information such as correlation of words [16], word constraints [2, 12], document labels [20], and document network structure [7, 11]. Typically, these models extend LDA by constraining the model variables with newly observed variables derived from side information.

These methods have shown some success but are constrained by two major limitations: Firstly, they assume labels are present for all latent topics. This assumption can be satisfied in situations where all topics are known in advance and obtaining side information is relatively easy, such as a collection of user generated content and tags as in [20]. However, in a large distributed complex environment, this is not a realistic assumption. Secondly, they support only one type of supervision, e.g., the domain knowledge should be provided as either document labels or feature labels. In a large distributed environment, labeling is typically done by a diverse set of domain experts, and labels can be provided in different forms. For instance, some experts may be willing to label a small set of sample documents; while others can provide some topic-indicative features (i.e. features which are known *a priori* to be good indicators of the topics).

In this paper, we propose a new semi-supervised topic model to address these limitations in a unified generative process. It provides a unified framework that discovers topics from data that is *partially labeled* with *heterogenous labels*:

Heterogeneous Supervision: We assume that multiple types of supervision can exist in the training data. For instance, some training data are provided with document labels, and some others are associated with topic-indicative features. Further, we assume that a topic can receive multiple types of labels, e.g., feature and document labels. A simplistic approach to support multiple label types is to sequentially build topic models, i.e, build a model with one label type and use this model's output to bootstrap the next iteration with another label type. This naive approach is inefficient due to multiple learning steps and fail to capture new information reinforced by different label types. Instead, we develop a unified model to simultaneously learn from different types of domain knowledge.

Partial Supervision: *h*LLDA also can handle the label partialness problem, where the training data are partially labeled. We allow for two types of partial labels:

- *Partially labeled document*: The labels for a document cover only a subset of all the topics the document belongs to. Our goal is to predict all the topics for the document.
- *Partially labeled corpus*: Only a small number of documents in a corpus are provided with labels. Our goal is to find the labels for all the documents.

We validate our algorithm using *Reuters*, *20 Newsgroup* and *Delicious*, which have been widely used in previous topic models and are adequate for testing the label partialness problem, since the documents contain multiple topics. The experiments for the

label heterogeneity shows that *h*LLDA achieves about 3 percentage points higher classification and clustering accuracy than LLDA by adding feature labels comprising only 10 words for each topic. The experiments for the label partialness shows that *h*LLDA produces 8.3 percentage points higher clustering accuracy and 34.4% improvement on Variational Information compared with LLDA. The results confirm that *h*LLDA significantly enhances the applicability of topic modeling for situations where partial, heterogenous labels are provided. Further we show that learning from multiple forms of domain knowledge in a unified process creates an enhanced combined effect that is greater than a sum of multiple models learned separately with one type of supervision.

In summary, the main contributions of the paper include:

- We propose a novel unified generative model that can simultaneously learn from different types of domain knowledge such as document labels and feature labels.
- *h*LLDA effectively solves the label partialness problem when the document label set is a subset of the topic set and/or the training data contain unlabeled documents.
- *h*LLDA is simple and practical, and it can be easily reduced to LDA, zLDA and LLDA depending on the availability of domain information.

The remainder of this paper is structured as follows. We first compare *h*LLDA with existing supervised and semi-supervised topic modeling algorithms in Section 2. Section 3 describes the generative process of *h*LLDA and the learning and inference algorithm in details. Experimental data and evaluation results are shown in Section 4 and Section 5. Section 6 provides final discussions and future work.

2 Related Work

*h*LLDA is broadly related to semi-supervised and supervised topic models. Existing (semi-)supervised topic models can be categorized into two groups based on the type of domain knowledge they utilize: *document supervision* and *feature supervision*.

Document Supervision

Existing approaches that utilize document labels fall in supervised learning assuming that all the documents in the training data have document labels. Supervised methods such as sLDA [5], discLDA [15], and medLDA [24] have shown a comparable performance on classification and regression tasks as general discriminative classifiers, but they support only one topic for a document. Labeled LDA (LLDA) [20] extends previous supervised models to allow multiple topics of documents, and Partially labeled LDA (PLDA) [21] further extends LLDA to have latent topics not present in the document labels. PLDA supports one-to-many mapping between labels and topics, but the number of latent topics is fixed constant for all documents. Recently, [14] propose a non-parametric topic model using Dirichlet Process with Mixed Random Measures (DP-MRM) that allows one label to be mapped with multiple topics. [18] propose a Dirichlet-multinomial regression (DMR) topic model that can incorporate arbitrary types of observed document features, such as author and publication venue, by providing a log-linear prior on document-topic distributions. DMR can be viewed as a supervised topic model by treating document labels as document features.

Table 1. Comparison of *h*LLDA with supervised and semi-supervised topic models using document labels

	No. of Topics per Document	Label-Topic Mapping	Label Partialness
sLDA	single	one-to-one	no
LLDA	multiple	one-to-one	no
PLDA	multiple	one-to-many	yes
DP-MRM	multiple	one-to-many	no
***h*LLDA**	multiple	one-to-one	yes

Table 2. Comparison of *h*LLDA with supervised and semi-supervised topic models using word labels

	Label Type	Label-Topic Mapping	Label Partialness
zLDA	unlabeled groups of features	one-to-one	no
SeededLDA	unlabeled groups of features	one-to-one	no
***h*LLDA**	labeled or unlabeled groups of features	one-to-many	yes

Feature Supervision

A feature label is typically provided as a set of words that are likely to belong to the same topic. Feature labels are helpful for discovering non-dominant or secondary topics by enforcing the words be assigned to the labeled topics, while standard LDA usually ignore them in favor of more prominent topics. Andrzejewski *et al.* proposed three different approaches for incorporating feature labels. In zLDA, they constrain latent topic assignment of each word to a set of seed words [2]. [3] applies Dirichlet Forest which allows must-links and cannot-links on topics, and [4] uses First-Order-Logic to generate human friendly domain knowledge. [12] described Seeded LDA that restricts latent topics to specific interests of a user by providing sets of seed words. To maximize the usage of seed words in learning, they jointly constrain both document-topic and topic-word distributions with the seed word information.

To our knowledge, *h*LLDA is the only semi-supervised topic model that combine heterogeneous side information together in one generative process, and discover the topics of documents using partially labeled documents and/or corpus. Table 1 and Table 2 summarize the differences of *h*LLDA with other existing algorithms that support document supervision and word supervision respectively.

3 Hetero-Labeled LDA

In this section, we describe *h*LLDA in detail and discuss how it handles heterogeneous labels and partially labeled data. We propose a unified framework that can incorporate multiple types of side information in one simple generative process.

Preliminaries

We first introduce some notations that will be used in the paper as shown in Table 3.

Table 3. Notations

$\mathcal{D}$	a document collection, $\{d_1, d_2, \ldots, d_M\}$
M	the number of documents in $\mathcal{D}$
$\mathcal{V}$	the vocabulary of $\mathcal{D}$, $\{w_i, w_2, \ldots, w_N\}$
N	the size of $\mathcal{V}$, i.e., the number of unique words in $\mathcal{D}$
$\mathcal{T}$	the set of topics in $\mathcal{D}$, $\{T_1, T_2, \ldots, T_K\}$
K	the number of topics in $\mathcal{T}$
$\mathcal{L}_W$	the set of topics provided by word labels
K_W	the number of unique topics in $\mathcal{L}_W$
$\mathcal{L}_D$	the set of topics provided by document labels
K_D	the number of unique topics in $\mathcal{L}_D$
$\mathcal{L}$	the label space, i.e., $\mathcal{L} = \mathcal{L}_W \cup \mathcal{L}_D$
$\mathcal{D}_L$	labeled documents
$\mathcal{D}_U$	unlabeled documents, i.e., $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$

We also define three different levels of side information for both document supervision and feature supervision.

Definition 1 (Side Information) *Any domain knowledge that can constrain the topic distributions of documents or words.* hLLDA *supports the following three different levels of side information.*

- `Group Information`*: It only specifies that a group of documents or words that belong to a same set of topics (e.g.,* $L_d = \{d_1, d_2, \ldots, d_c\}$*) and* $L_w = \{w_1, w_2, \ldots, w_g\}$*).*
- `Label Information`*: This side information provides a group of labels with associated topic labels. For instance,* $L_d = \{d_1, d_2, \ldots, d_c; T_1, T_2, \ldots, T_k\}$ *specifies that the documents belong to topics* $T_1, \ldots, T_k$*, where* $1 \leq k \leq K$*.*
- `Topic Distribution`*: This information further provides topic distributions of the label information. For instance,* $L_d = \{d_1, \ldots, d_c; T_1, \ldots, T_k; p_1, \ldots, p_k\}$ *indicates that the documents belonging to the topic* T_i *with the likelihood of* p_i*. We note that* p_i *is a perceived likelihood value by domain experts, and* $\sum_i p_i < 1$ *in many cases.*

hLLDA Model

The main goals of *h*LLDA are to build a topic model that can incorporate different types of labels in a unified process and to discover all underlying topics when only a small subset of the topics are known in advance. We solve the problems by modifying both the document topic distribution (θ) and word topic assignment (z) with the side information. Figure 1 depicts the graphical representation of *h*LLDA. In *h*LLDA, the global topic distribution θ is generated by both a Dirichlet topic prior α and a label-specific topic mixture ψ obtained from the document labels Λ_d with a Dirichlet prior γ. Then, the word topic assignment z is generated from the global topic mixture θ constrained by word labels Λ_w.

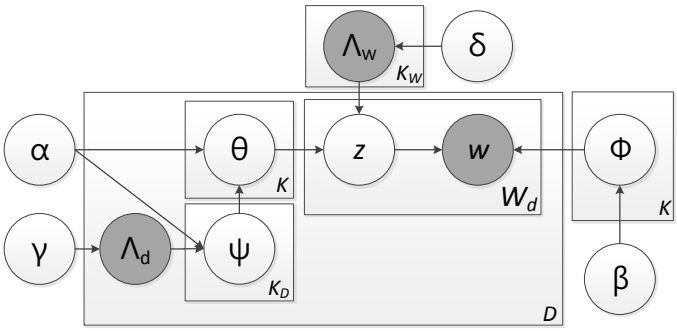


Fig. 1. Graphical representation of *h*LLDA. $|\Lambda_d| = K_D$ and $|\Lambda_w| = K_W$. Note that z is influenced by both the word side information (Λ_w) and the document side information (Λ_d) in *h*LLDA, producing synergistic effect of heterogeneous side information.

Table 4 describes the generative process of *h*LLDA in more detail. In *h*LLDA, the total number of topics (K) is set to the sum of the numbers of unique topics present in the document and word labels (i.e., $|\mathcal{L}_\mathcal{D} \cup \mathcal{L}_\mathcal{W}|$) and the number of additional latent topics (K_B) the user wants to discover from the corpus. Here, the number of latent topics (K_B) is an input parameter.

We first draw multinomial topic distributions over the words for each topic k, ϕ_k, from a Dirichlet prior β as in the LDA model [6] (line 1–2). However, unlike other LDA models, *h*LLDA has an additional initialization step for word topic assignment z, when word (feature) labels are provided as side information (line 3–5). For each topic appearing in the word labels, k_W, we draw multinomial topic distributions, Λ_{k_W}, over the vocabulary using a smoothed *Bernoulli* distribution, i.e., $\Lambda_{k_W}^{(w)} = (l_1, l_2, ..., l_V)$ where $l_v \in \{\delta, 1-\delta\}$. The $Bernoulli_{smooth}$ distribution generates smoothed values δ $(0 < \delta < 1)$ with success probability p or $1-\delta$ with failure probability $1-p$, rather than value 1 with probability p and value 0 with probability $1-p$ as in the *Bernoulli* distribution. We propose the $Bernoulli_{smooth}$ distribution to handle the label partialness. Note that the *Bernoulli* distribution does not allow words or documents to be assigned to the topics not provided in the document or feature labels. However, with $Bernoulli_{smooth}$, documents and words can be assigned to topics from other latent topics with a low probability $1-\gamma$ and $1-\delta$ respectively.

The $Bernoulli_{smooth}$ distribution drawn from word label information, Λ_{k_W}, contains a vector of topics for each word and is later used to constrain the global topic mixture θ as described in line 16. We multiply Λ_{k_W} with θ to generate the multinomial distribution z (line 16). The topic assignment z_i for each word i in a document d is chosen from a multinomial distribution $\{\lambda_1^{(d)}, \ldots, \lambda_K^{(d)}\}$, where $\lambda_i^{(d)}$ denotes the assigned topic for word i in document d and is generated by multiplying the global topic mixture θ and the word label constraint Λ_{k_W}. Applying soft constraints on word topic assignment z using word labels is similar to zLDA [2], but, zLDA puts constraints on word instances, while *h*LLDA puts constraints over the vocabulary elements. Further, by influencing z with the mixture of the word side information and the document side information (see

Table 4. Generative process for *h*LLDA. $Bernoulli_{smooth}$ distribution generates smoothed values (e.g., value $v, 0 < v < 1$ with success probability p or $1 - v$ with failure probability $1 - p$) rather than value 1 or value 0.

1	For each topic $k \in \{1, ..., K\}$
2	Generate $\phi = (\phi_{k,1}, \ldots, \phi_{k,V})^T \sim Dir(\cdot\|\beta)$
3	For each topic $k_W \in \{1, \ldots, K_w\}$
4	For each word $w \in \{1, \ldots, N\}$
5	Generate $\Lambda_{k_W}^{(w)} \sim Bernoulli_{smooth}(\cdot\|\delta)$
6	For each document d:
7	if $d \in \mathcal{D}_U$
8	Generate $\theta^{(d)} = (\theta_1, \ldots, \theta_K)^T \sim Dir(\cdot\|\alpha)$
9	if $d \in \mathcal{D}_L$
10	For each topic $k_D \in \{1, \ldots, K_D\}$
11	Generate $\Lambda_{k_D}^{(d)} \sim Bernoulli_{smooth}(\cdot\|\gamma)$
12	Generate $\Psi^{(d)} = (\psi_1, \ldots, \psi_{K_d})^T \sim Dir(\cdot\|\alpha \cdot \Lambda_{k_D}^{(d)})$
13	Generate $\theta^{(d)} = (\theta_{K_d+1}, \ldots, \theta_{(K)})^T \sim Dir(\cdot\|\alpha_{K_d+1:K})$
14	Generate $\theta^{(d)} = (\Psi^{(d)^T} \mid \theta^{(d)^T})^T$
15	For each i in $\{1, \ldots, N_d\}$
16	Generate $z_i \in \{\lambda_1^{(d)}, \ldots, \lambda_K^{(d)}\} \sim Mult(\cdot\|\Lambda_{k_W}^{(i)} \cdot \theta^{(d)})$
17	Generate $w_i \in \{1, \ldots, V\} \sim Mult(\cdot\|\phi_{z_i})$

Figure 1), *h*LLDA can benefit from the combined effect of multiple heterogeneous side information.

*h*LLDA generates the document topic distribution θ differently for documents with document side information and for documents without document labels (line 7–14). If the document is unlabeled (i.e., $d \in \mathcal{D}_U$), we generate topics using the Dirichlet prior α in the same way as in LDA (line 8). If the document is labeled (i.e., $d \in \mathcal{D}_L$), we first generate the document labels over topics $\Lambda_{K_D}^{(d)} = (l_1, l_2, ..., l_{K_D})$, where $l_k \in \{\gamma, 1 - \gamma\}$ is drawn from the smoothed *Bernoulli* distribution, $Bernoulli_{smooth}(\cdot|\gamma)$ (line 10-11). The soft constraints on document labels enable *h*LLDA to discover other latent topics for *partially labeled documents or corpus*, which do not exist in the document labels. We note that this is different from both Labeled LDA (LLDA) [20] and Partially Labeld LDA (PLDA) [21]. In LLDA, a document is strictly constrained to generate topics only from the provided document labels. PLDA relaxes this restriction and allows a document to be assigned a set of latent topics that are unseen in the document labels, but the number of the latent topics is arbitrarily fixed constant for all documents.

Note that, in $Bernoulli_{smooth}(\cdot|\delta)$ and $Bernoulli_{smooth}(\cdot|\gamma)$, the values for δ and γ are larger than $1-\delta$ and $1-\gamma$ respectively, ensuring that the topic distributions from the side information have more weights than the topics not covered by the side information. Further, when the document side information is provided in the form of Topic Distribution as described in Definition 1, the perceived likelihoods, p_i, are used as biased priors.

We generate a document label-topic mixture $\Psi^{(d)}$ of size K_D using the Dirichlet topic prior α and the document label constraints $\Lambda^{(d)}_{K_D}$ (line 12) and then generate a latent topic mixture $\theta^{(d)}$ of size K-K_D using the Dirichlet prior α (line 13). Finally, we concatenate the document label-topic mixture Ψ and the latent topic mixture θ to generate θ with size K (line 14). The concatenation together with the soft constraints on document topics allow the document to generate new topics that are not included in the document labels from *partially labeled documents or corpus*. Even though the concatenation of Dirichlet random variables does not produce a value that is an element of the simplex, our experiments show that it solves the label partialness very well.

The remaining steps (line 15–17) are similar to the processes in LDA. For each word i in document d, we generate topic assignment z_i from multinomial distribution $\theta^{(d)}$ and word label constraint $\Lambda^{(i)}_{k_W}$ and generate the word from multinomial distribution ϕ_{z_i}.

Learning and Inference

We use the Gibbs sampling algorithm [9] to estimate the latent variables θ, ψ, and ϕ. We note that the word and document label priors δ and γ are independent from the rest of model parameters, and, since we simply concatenate ψ into θ (line 14), we can use the same inference as in LDA. Thus, our inference process follows the Gibbs sampling procedure that estimates only θ and ϕ.

At each iteration, the topic of ith document, z_i, is estimated by the conditional probability

$$\begin{aligned} &P(z_i = k|\boldsymbol{z_{-i}}, \boldsymbol{w}, \boldsymbol{\Lambda_W}, \boldsymbol{\Lambda_D}, \alpha, \eta, \gamma, \delta) \\ &\propto P(z_i = k|\boldsymbol{z_{-i}}, \boldsymbol{w}, \boldsymbol{\Lambda_W}, \alpha, \eta, \gamma) \\ &\propto \Lambda_k^{(w_i)} \times \left(\frac{n^{(w_i)}_{-i,k} + \eta}{\sum_{w'}^{W} \left(n^{(w')}_{-i,k} + \eta \right)} \right) \left(\frac{n^{(d)}_{-i,k} + \alpha}{\sum_{k'}^{T} \left(n^{(d)}_{-i,k'} + \alpha \right)} \right) \end{aligned} \tag{1}$$

where $\Lambda_k^{(w_i)}$ is a word label constraint that outputs $\gamma, 0 < \gamma < 1$ when $w_i \in \Lambda_W$, and 1-γ when $w_i \notin \Lambda_W$. The soft constraints on sampling procedure is similar to zLDA [2], except that the topic k can be a new topic not in the word labels. Then, we obtain the estimated probability ϕ_{kw} of word w in topic k and the estimated probability θ_{dk} of topic k in document d using Equation 2 and 3 respectively.

$$\phi_{kw} = \frac{n^{(w_i)}_{-i,k} + \eta}{\sum_{w'}^{W} \left(n^{(w')}_{-i,k} + \eta \right)} \tag{2}$$

$$\theta_{dk} = \frac{n^{(d)}_{-i,k} + \alpha}{\sum_{k'}^{T} \left(n^{(d)}_{-i,k'} + \alpha \right)} \tag{3}$$

When no side information is provided, hLLDA is reduced to LDA. Compared to LLDA, θ in hLLDA is limited by soft constraints drawn from the documents labels,

and, thus, becomes the same as LLDA, when only document side information is considered, and the document label prior γ is a binary vector representing the existence of topic labels for each document. Compared to zLDA, z in hLLDA is softly constrained by both the word labels and the document labels in assigning topics for each word in each document. hLLDA can be reduced to zLDA, when the side information contains only word labels, and K_W is equal to K. Based on these observations, hLLDA can be viewed as a generalized version of LDA, LLDA and zLDA. Further, we note that the existence of latent topic mixture θ enables hLLDA to find latent topics not covered by the document or word labels without harming the original distribution of topics from the labels.

4 Experiments

We conduct experiments to answer the following questions:

Q1 How effective is learning from mixture of heterogeneous labels for topic categorization?
Q2 How well does hLLDA discover latent topics from partially labeled documents and corpus?
Q3 How accurate are the generated topics?

Data

All experiments are conducted with three public data sets–*Reuters-21578* [22], *20 Newsgroup* [1], and *Delicious* [8]. The *Reuters-21578* data set contains a collection of news articles in 135 categories, and we chose the 20 most frequent topics for the experiments (hereafter called *Reuters*). For the *20 Newsgroup* dataset, we use all 20 categories in the data set (hereafter called *20News*). For the *Delicious* data set, we first selected the 50 most frequent tags in *Delicious.com*, and then manually chose 20 tags from the 50 tags and 5,000 documents for the selected 20 categories (hereafter called *Delicious*). Table 5 shows the topic categories in the the experiment data sets. We then conducted the following text processing on the documents: First, all stopwords were removed and words were stemmed using Porter's Stemmer [19]. Then, all words occurring in fewer than 5 documents were discarded. After the preprocessing, *Reuters* contains 11,305 documents and 19,271 unique words; *20News* has 19,997 documents with 57,237 unique words; and *Delicious* contains 5,000 with 141,787 unique words.

Domain Knowledge

We use the topic labels in the data sets as document side information. To evaluate the label heterogeneity (**Q1**) and partialness problems (**Q2**), we conduct experiments with varying amount of document side information comprising the first 5, 10, 15 and 20 labels from the topics in Table 5. We treat the documents belonging to the selected categories as labeled and the remaining documents as unlabeled.

For word side information, we extracted top 20 words for each class based on TF-IDF (term frequency-inverse document frequency), manually filtered irrelevant words

Table 5. The 20 topics in *Reuters*, *20News*, and *Delicious* data sets

Reuters	earn, acq, money-fx, crude, grain, trade, interest, wheat, ship, corn, dlr, oilseed, sugar, money-supply, gnp, coffee, veg-oil, gold, nat-gas, soybean
20News	alt.atheism, sci.space, comp.os.ms-windows, rec.sport.baseball, misc.forsale, soc.religion.christian, rec.autos, sci.crypt, talk.religion.misc, sci.med, comp.sys.ibm.pc.hardware, rec.sport.hockey, talk.politics.guns, sci.electronics, comp.graphics, rec.motorcycles, talk.politics.misc, comp.sys.mac.hardware, talk.politics.mideast, comp.windows.x
Delicious	design, web, software, reference, programming, art, education, resources, photography, music, business, technology, research, science, internet, shopping, games, marketing, typography, graphics

out and chose top 10 words as final word labels. When a word appears in multiple classes, we remove the word from all the classes except the class for which the word has the highest TF-IDF value. In real world, word labels are given by domain experts so they have more meaningful information than our artificially generated word labels. Even though we have conducted an experiment with real business data that contains document and word labels with successful experimental results, they are not included in this paper due to confidential information.

Evaluation Methods

We implement two variations of *h*LLDA and compare them with three existing topic modeling algorithms–LDA [6], LLDA [20] and zLDA [2]. (For multi-label classification task such as *Reuters* and *Delicious*, sLDA is not appropriate to compare with [20] so we does not include sLDA in our experiment.) The first version, *h*LLDA (L=T), assumes that all the topics are present in the labels to directly compare it with LLDA. The second version. *h*LLDA (L<T), is for cases where the label set is a subset of the topic set and validate the label partialness problem. For all the models, we use a Collapsed Gibbs sampler [10] for inference with standard hyper-parameter values $\alpha = 1.0$ and $\beta = 0.01$ and run the sampler for 1,000 iterations.

All comparisons are done using 5-fold cross validation over 10 random runs. For question **Q1** and **Q2**, we measure the following three evaluation metrics. For **Q3**, we compare the discovered topics qualitatively by visualizing the topics.

Prediction Accuracy: We predict a label of a new document by choosing the topic with the highest probability from the posterior document-topic distribution θ and check whether the label exists in the topic set of the document.

Clustering F-measure: We simulate clustering by assigning each document to the topic (i.e., cluster) that has the highest probability in θ. If two documents belong to the same topic by both the ground truth and by the simulated clustering, then it is regarded as correct. The F-measure is then calculated for all the pairs of documents. Even though clustering may not be a general metric to evaluate topic modeling algorithms, it can be

a good indicator of how topics are coherently grouped together especially when label information is incomplete (i.e., label partialness). Section explains the details.

Variational Information(VI): VI measures the amount of information lost and gained in changing clustering C_1 to clustering C_2 [17]. The VI of two clusters X and Y is calculated as $VI(X, Y) = H(X) + H(Y) - 2 * I(X, Y)$ where $H(X)$ (or $H(Y)$) denotes the entropy of the clustering X (or Y), and $I(X, Y)$ is the mutual information between X and Y. Lower VI values indicate better clustering results.

5 Experimental Results

We measure the performance of *h*LLDA and the baseline systems for the label heterogeneity the label partialness problems and also visually compare the discovered topics by *h*LLDA and LLDA.

Label Heterogeneity

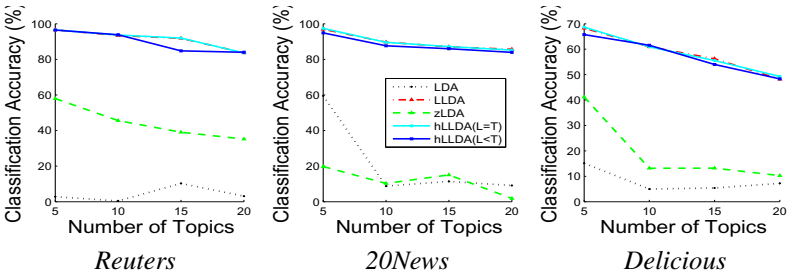
We first validate the effectiveness of *h*LLDA in dealing with heterogenous labels. In this experiment, we used document labels and feature labels as heterogeneous domain knowledge for *h*LLDA, but we can easily extend to other types of labels such as document structure labels. Further, we assume that all topics appear in the labels, and all training documents are labeled with document labels or feature labels.

Figure 2(a) shows the accuracy of multi-class prediction. As we can see, both versions of *h*LLDA perform well for all three data sets. The accuracy levels of *h*LLDA are significantly better than LDA and zLDA and slightly higher than LLDA. This indicates that mixture of two heterogeneous domain information improve the prediction accuracy. Figure 2(b) shows the F-measure of the multi-class clustering task. The F-measure of both *h*LLDA algorithms show similar performance as LLDA while significantly outperforming LDA and zLDA. We notice that, however, for *Delicious*, *h*LLDA is better than LLDA confirming that adding feature label information is beneficial. These results indicate that *h*LLDA can combine different types of supervision successfully, and the combination of heterogeneous label types is beneficial for both classification and clustering tasks.

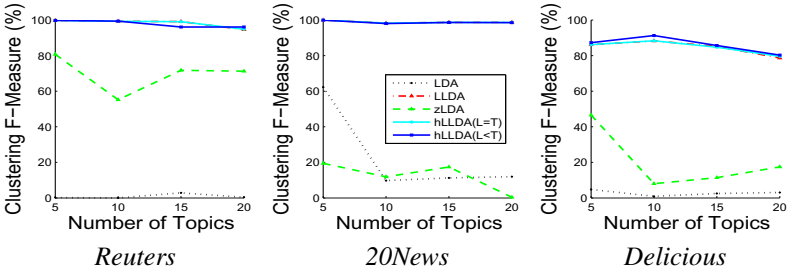
Label Partialness

For the label partialness problem, we consider two types of label partialness: *partially labeled document* and *partially labeled corpus*.

Partially Labeled Documents: The goal is to predict the full set of topics for a document when only a subset of topics is provided as labels for the document. We conduct experiments for different levels of partialness ranging from 10% to 100% with 10% interval. For p% partialness, we include a topic in the document's label set with probability p. In this experiment, *20News* and *Delicious* were used because most documents in the data sets have multiple topics. As we can see from the results shown in Figure 3, *h*LLDA, especially *h*LLDA ($L < T$), outperforms all other algorithms both in terms of clustering F1-measure and VI.

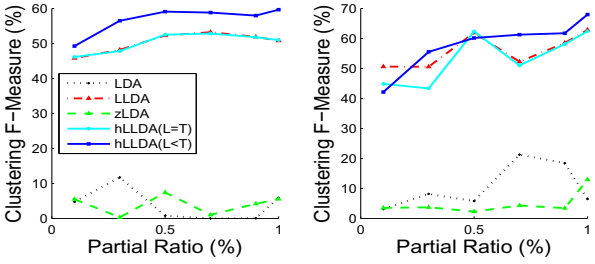


(a) Multi-Class Prediction Accuracy

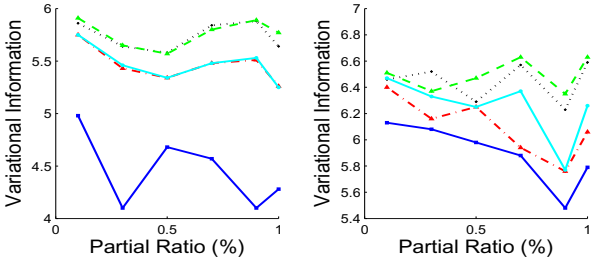


(b) Clustering F-measure

Fig. 2. Performance comparison for label heterogeneity

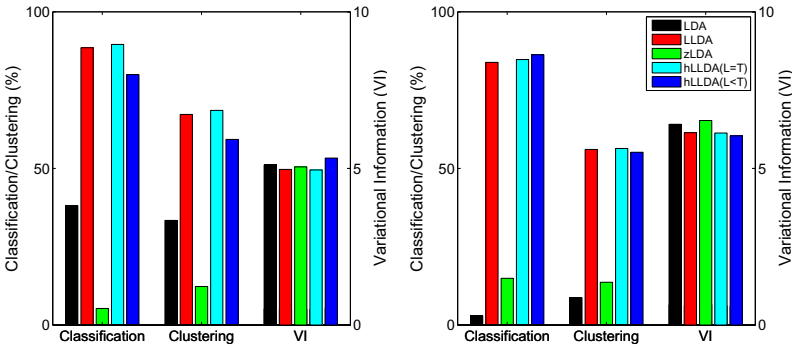


(a) Clustering F-measure



(b) Variational Information (VI)

Fig. 3. Clustering F-measure and VI (the lower the better) for partially labeled documents on *20News* (left) and *Delicious* (right). *PartialRatio* indicates the probability of each topic being included in the labels.



(a) 48% Labeled **(b)** 64% Labeled

Fig. 4. Performance comparison for partially labeled corpus on *Delicious*

Table 6. Number of topically irrelevant (Red) and relevant (Blue) words marked by users in Table 7. The more red words are, the lower the topic quality is. Similarly, the more blue words are, the higher the topic quality is.

	LLDA		***h*LLDA**	
	#RedWords	#BlueWords	#RedWords	#BlueWords
20News	15	11	2	35
Delicious	17	12	6	30

Partially Labeled Corpus: The goal is to find the labels for all the documents in the corpus when only a subset of the documents are labeled ($|\mathcal{D}_L| \ll |\mathcal{D}|$). We conduct the same experiments as for label heterogeneity using *Delicious*, but introduced unlabeled documents in the training data. Figure 4a and Figure 4b show the results when only the documents belonging to the first 5 topics (48% of the documents) and the first 10 topics (64% of the documents) are considered labeled respectively. As we can see, *h*LLDA outperforms both LDA and zLDA significantly in all cases. Further, the results show that *h*LLDA achieves a comparable performance to LLDA while using less than half of the labels and even better performance only with about 60% of the labels!

Quality of Discovered Topics

We compare the quality of topics discovered by *h*LLDA with partial labels and by LLDA with full labels. We ran *h*LLDA using only 10 topics as the documents labels and discovered 20 topics. To keep the amount of domain information the same, we split the data set into two subsets with 10 topics each and ran LLDA separately for each subset. Table 7 shows the discovered topics for *20News* (top) and *Delicious* (bottom): The first column shows the the true topics, and the second and the third columns show the

Table 7. Comparison of topics generated by LLDA with full labels and *h*LLDA with partial labels. Each row depicts a topic label and top five words for the topic discovered by the two algorithms. Words marked in red or blue show the differences between the two algorithms. The words in red indicate topically irrelevant words, and the words in blue denote relevant words for the topic.

	Labels	**LLDA**(L=10,T=10) & **LLDA**(L=10,T=10)	***h*LLDA** (L=10,T=20)
20News	atheism	peopl, dont, god, moral, believ	peopl, god, dont, moral, believ
	space	space, launch, orbit, time, system	space, launch, orbit, system, time
	ms-windows	window, file, program, imag, run	window, file, driver, run, program
	baseball	game, team, plai, player, win	game, player, team, dont, hit
	forsale	drive, card, scsi, system, sale	sale, email, price, plea, drive
	christian	god, christian, peopl, believ, church	god, christian, peopl, believ, church
	autos	car, dont, bike, im, time	car, bike, dont, engin, im
	crypt	govern, kei, peopl, gun, encrypt	kei, encrypt, chip, govern, secur
	religion.misc	peopl, armenian, dont, jew, israel	god, peopl, dont, christian, moral
	med	medic, dont, health, peopl, drug	medic, effect, dont, disea, studi
	pc.hardware	drive, scsi, card, id, control	drive, card, scsi, mac, monitor
	hockey	game, team, plai, hockei, player	game, team, plai, hockei, win
	politics.guns	gun, peopl, dont, weapon, fire	gun, law, weapon, peopl, crime
	electronics	wire, ground, dont, circuit, power	power, wire, batteri, circuit, ground
	graphics	imag, file, graphic, program, format	-
	motorcycles	bike, dod, ride, dont, motorcycl	-
	politics.misc	peopl, dont, presid, govern, time	▹ presid, dont, peopl, govern, job ▹ parti, polit, vote, convent, univ
	mac.hardware	mac, appl, drive, monitor, system	-
	politics.mideast	armenian, peopl, israel, isra, turkish	▹ armenian, turkish, muslim, armenia, turk ▹ israel, isra, jew, arab, jewish
	windows.x	window, file, program, server, run	ile, imag, program, displai, window + fire, peopl, start, didnt, dont, children
Delicious	design	design, comment, repli, post, thank	design, comment, post, thank, repli
	software	file, softwar, download, support, web	file, download, softwar, window, free
	art	post, art, begin, map, comment	art, post, begin, artist, book
	education	learn, student, educ, talk, world	learn, student, educ, talk, world
	science	scienc, peopl, time, page, link	scienc, peopl, time, page, depress
	photography	photo, am, photographi, comment, jul	photo, am, photographi, post, photograph
	music	music, record, rock, band, de	music, record, rock, band, song
	business	xpng, twitter, busi, search, blog	busi, search, blog, inform, servic
	games	game, element, function, code, html	game, comment, articl, appl, app
	marketing	de, que, la, social, en	twitter, social, post, media, market
	shopping	tshirt, shop, de, product, top	ship, free, price, shop, offer
	typography	font, design, thank, type, comment	-
	graphics	icon, file, free, graphic, brush	-
	programming	code, function, post, file, page	▹ element, function, code, exampl, content ▹ python, tornado, thread, framework, server
	research	research, start, post, search, comment	-
	web	xpng, web, css, user, site	xpng, scalablesvg, xsvg, flash, arduino
	internet	de, que, le, da, la	-
	technology	comment, googl, technolog, inform, app	-
	reference	element, pdf, html, content, map	pdf, html, sheet, cheat, intel
	resources	repli, design, post, free, thank, web, site	+ stack, librari, sentenc, data, scholar + oct, plugin, jul, commentcont, jan + de, le, la, un, et + de, que, la, para, el + die, und, der, map, da

top 5 words discovered by LLDA and *h*LLDA respectively. We marked the topics that *h*LLDA did not find with '-' , and the topics *h*LLDA generated but do not exist in the data set with '+'. The topics with '▹' indicate that multiple topics were generated for one true topic. As we can see, *h*LLDA discovers topics very accurately with the first 10 topics matching very well with the true topics for both *20News* and *Delicious*. Further note that, for both *20News* and *Delicious* data sets, *h*LLDA discovered new latent topics even though no labels were provided for these topics. For example,*h*LLDA discovered 6 out of 10 latent topics for *20News*, such as *pc.hardware*, *hockey*, *politics.guns*, *electronics*, *politics.misc* and *windows.x*.

We also examine the top 5 words for each topic: The words discovered by both algorithms are marked in black, and words discovered by only one algorithm are marked in red or blue– blue denoting relevant words and red denoting irrelevant words respectively. As we can see, *h*LLDA generates much more relevant (blue) words at the top and also extract more general words than LLDA, even when both cases were judged topically relevant. For instance, LLDA generates "drive", "card", "scsi" for topic *forsale*, while *h*LLDA produces "sale", "price", and "offer". The same trend is seen for *Delicious* data set, especially for topics *business*, *games* and *marketing*. Table 6 shows the total number of blue and red words generated by LLDA and *h*LLDA. As we can see, *h*LLDA produced much more relevant words and much fewer irrelevant words for both data sets, yielding 87% and 65% reduction in red words and 218% and 150% increase in blue words for *20News* and *Delicious* respectively. The results clearly show the effectiveness of *h*LLDA in handling partial labels.

6 Conclusion

We proposed *h*LLDA, a partially supervised topic model to deal with the heterogeneity and partialness of labels. Our algorithm is simple and flexible and can deal with different label types in a unified framework. Experimental results demonstrate the effectiveness of *h*LLDA for both label heterogeneity and label partialness problems. Experiments also validate that *h*LLDA can discover latent topics for which no label or side information was provided. Further, *h*LLDA produces comparable classification performance and much better clustering performance than existing semi-supervised models while using much smaller amount of labels.

In the future, we plan to incorporate additional type of label information such as partial or full taxonomy of topics [13]. Also, to further improve the performance of label prediction for partially labeled documents, we consider generating topic hierarchies such as Hierarchical Dirichlet Process (HDP) [23].

References

1. 20 Newsgroup, `http://qwone.com/~jason/20Newsgroups/`
2. Andrzejewski, D., Zhu, X.: Latent dirichlet allocation with topic-in-set knowledge. In: NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing (2009)
3. Andrzejewski, D., Zhu, X., Craven, M.: Incorporating domain knowledge into topic modeling via dirichlet forest priors. In: ICML (2009)

4. Andrzejewski, D., Zhu, X., Craven, M., Recht, B.: A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In: IJCAI (2011)
5. Blei, D.M., McAuliffe, J.D.: Supervised topic models. In: NIPS (2007)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR (2003)
7. Chang, J., Blei, D.M.: Relational topic models for document networks. Journal of Machine Learning Research - Proceedings Track 5 (2009)
8. Delicious, http://arvindn.livejournal.com/116137.html
9. Griffiths, T.L., Steyvers, M.: Proceedings of the National Academy of Sciences (2004)
10. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS (2004)
11. Ho, Q., Eisenstein, J., Xing, E.P.: Document hierarchies from text and links. In: WWW (2012)
12. Jagarlamudi, J., Daumé, H., Udupa, R.: Incorporating lexical priors into topic models. In: EACL, pp. 204–213 (2012)
13. Kang, D., Jiang, D., Pei, J., Liao, Z., Sun, X., Choi, H.J.: Multidimensional mining of large-scale search logs: a topic-concept cube approach. In: WSDM (2011)
14. Kim, D., Kim, S., Oh, A.: Dirichlet process with mixed random measures: a nonparametric topic model for labeled data. In: ICML (2012)
15. Lacoste-Julien, S., Sha, F., Jordan, M.I.: Disclda: Discriminative learning for dimensionality reduction and classification. In: NIPS (2008)
16. Lafferty, J.D., Blei, D.M.: Correlated topic models. In: NIPS (2005)
17. Meilă, M.: Comparing clusterings—an information based distance. J. Multivar. Anal. (2007)
18. Mimno, D.M., McCallum, A.: Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In: UAI, pp. 411–418 (2008)
19. Porter, M.F.: An algorithm for suffix stripping. Program: electronic library and information systems (1980)
20. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP (2009)
21. Ramage, D., Manning, C.D., Dumais, S.T.: Partially labeled topic models for interpretable text mining. In: KDD (2011)
22. Reuters-21578, http://kdd.ics.uci.edu/databases/reuters21578/
23. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association (2006)
24. Zhu, J., Ahmed, A., Xing, E.P.: Medlda: maximum margin supervised topic models for regression and classification. In: ICML (2009)