The Hybrid Nested/Hierarchical Dirichlet Process and Its Application to Topic Modeling with Word Differentiation

Tengfei Ma and **Issei Sato** and **Hiroshi Nakagawa**
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{matf@r., sato@r., nakagawa@}dl.itc.u-tokyo.ac.jp

Abstract

The hierarchical Dirichlet process (HDP) is a powerful nonparametric Bayesian approach to modeling groups of data which allows the mixture components in each group to be shared. However, in many cases the groups themselves are also in latent groups (categories) which may impact the modeling a lot. In order to utilize the unknown category information of grouped data, we present the hybrid nested/ hierarchical Dirichlet process (hNHDP), a prior that blends the desirable aspects of both the HDP and the nested Dirichlet Process (NDP). Specifically, we introduce a clustering structure for the groups. The prior distribution for each cluster is a realization of a Dirichlet process. Moreover, the set of cluster-specific distributions can share part of atoms between groups, and the shared atoms and specific atoms are generated separately. We apply the hNHDP to document modeling and bring in a mechanism to identify discriminative words and topics. We derive an efficient Markov chain Monte Carlo scheme for posterior inference and present experiments on document modeling.

Introduction

Bayesian nonparametric models have drawn a lot of attention in the machine learning community recently. Among them the hierarchical Dirichlet process (HDP) (Teh et al. 2006) has achieved great success for modeling groups of data. It has been applied to various areas such as topic modeling and hidden Markov models. It constructs multiple DPs by sharing the base measure which is drawn by another DP. With this setting, the HDP allows different groups to share mixture components.

The basic assumption of the HDP is that each group distribution is conditionally independent based on the same base measure. However, this assumption ignores the latent correlation among groups. If we consider a group of data as an object; the objects are often organized into categories, such as documents in multi-corpora data and epileptic seizures (groups of channels) across patients (Wulsin, Litt, and Jensen 2012). Intuitively, objects within the same category should be more similar to each other than to those in other categories. These kinds of category information are useful for modeling data (Lacoste-Julien, Sha, and Jordan 2008; Ramage et al. 2009), and finding implicit category structure is also an important task (Yu, Yu, and Tresp 2005). Teh et al. (Teh et al. 2006) even demonstrate that ignoring the category information would result in much worse performance in the multi-corpora topic modeling task.

In this paper, we consider the case that the membership of the grouped data is unknown. Our aim is to develop a Bayesian nonparametric model which improves the HDP by uncovering the latent categories of the data groups and utilizing them for data modeling simultaneously. We borrow the idea from the nested Dirichlet process (NDP) (Rodriguez, Dunson, and Gelfand 2008) and construct the hybrid nested hierarchical Dirichlet Process (hNHDP). The NDP is an extension of the Dirichlet Process (DP) to solve the problem of ”clustering probability distributions and simultaneous multilevel clustering”. It allows us to simultaneously cluster groups and observations within groups. The NDP and HDP both generalize the DP to allow hierarchical structures, but they induce fundamentally different dependence among the group distributions. For the HDP, different group distributions share the same atoms but assign them different weights. NDP, on the other hand, leads to distributions that have either the same atoms with the same weights or completely different atoms and weights.

We integrate the advantages of the NDP and the HDP in our hybrid model. The group distributions are clustered as in the NDP. However, different from the NDP, our model allows distributions to share some atoms even when they belong to different clusters; and it clusters the groups using only the local components (i.e. unshared atoms). We define the distribution F_k for each cluster k as a mixture of two independently drawn DPs: G_0 which is shared by all clusters and G_k which is cluster-specific. Through some settings, we make F_k still a realization of the DP. Thus our model can be transferred into a special case of the NDP. At the same time, it can be also regarded as a special case of an alternative HDP (Müller, Quintana, and Rosner 2004), which we call LC-HDP (linear combination based HDP) in this paper. Moreover, the model are very flexible by allowing to set different base measures (H_0 and H_1) for G_0 and G_j. As a result, G_j can only include useful features for clustering. This setting in fact induces feature selection in the clustering process.

We then apply the hNHDP to the problem of topic modeling, which is a suitable case for illustrating its power. Our model assumes documents to be mixtures of topics and assigns documents into latent categories. It discriminates between global topics and local topics and automatically identifies local words for local topics. The local topics and local words could help explain the data structure as in the section of experiments. In summary, the contribution of this work includes:

- improving the HDP (as well as LC-HDP) by taking advantage of the latent category information of grouped data.
- improving the NDP by sharing mixture components among all groups and discriminating local components.
- developing a topic model that organize documents into latent categories by topic distributions and identifies category-specific words and topics for data analysis.

Background

The Hierarchical Dirichlet Process

The HDP (Teh et al. 2006) is a Bayesian nonparametric prior for modeling groups of data. It ensures that sets of group-specific DPs share the atoms. Suppose that we have observations organized into groups. Let x_{ji} denote the i^{th} observation in group j. All the observations are assumed to be exchangeable both within each group and across groups, and each observation is assumed to be independently drawn from a mixture model. Let $F(\theta_{ji})$ denote the distribution of x_{ji} with the parameter θ_{ji}, which is drawn from a group-specific prior distribution G_j. For each group j, the G_j is drawn independently from a DP, $DP(\alpha_0, G_0)$. To share the atoms between groups, the HDP model forces G_0 to be discrete by defining G_0 itself as a draw from another DP, $DP(\gamma, H)$. The generative process for HDP is represented as:

$$\begin{aligned} G_0 &\sim \text{DP}(\gamma, H), \\ G_j &\sim \text{DP}(\alpha_0, G_0) \text{ for each j}, \\ \theta_{ji} &\sim G_j \text{ for each j and i}, \\ x_{ji} &\sim F(\theta_{ji}) \text{ for each j and i}. \end{aligned} \tag{1}$$

Using the stick-breaking construction of Dirichlet processes, we can express G_0 as $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, where δ_{ϕ_k} is a probability measure concentrated at the atom ϕ_k. The atoms are drawn from the base measure H independently, and the probability measure for the weights $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$[1] are mutually independent. Because G_0 has support at the points $\{\phi_k\}$, each G_j necessarily has support at these points as well; and can thus be written as $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$, where the weights $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty} \sim \text{DP}(\alpha_0, \boldsymbol{\beta})$.

LC-HDP

Motivated by a similar problem to that of the HDP(Teh et al. 2006), Müller et. al. (2004) developed another hierarchical Dirichlet process (LC-HDP). They considered a model in which a coupled set of random measures F_j is defined as

$$\begin{aligned} F_j &= \epsilon G_0 + (1-\epsilon) G_j, \\ G_j &\sim \text{DP}(\gamma, H) \text{ for } j = 0, 1, ...J. \end{aligned} \tag{2}$$

where $0 \leq \epsilon \leq 1$ defines weights of the linear combination. This model provides an alternative approach to sharing atoms, in which the shared atoms are given the same stick-breaking weights in each of the groups. Its has an attractive characteristic that it can discriminate local components, which are useful for clustering.

The Nested Dirichlet Process

The NDP (Rodriguez, Dunson, and Gelfand 2008) is motivated by simultaneously clustering groups and observations within groups. It induces multi-level clustering, while the HDP can cluster only observations. In the NDP model, the groups are clustered by their entire distribution. Consider a set of distributions $\{G_j\}$, each for one group. If $\{G_j\} \sim \text{nDP}(\alpha, \gamma, H)$, it means that for each group j, $G_j \sim Q$ with $Q \equiv \text{DP}(\alpha \text{DP}(\gamma H))$. This implies that we can first define a collection of DPs

$$G_k^* \equiv \sum_{l=1}^{\infty} w_{lk} \delta_{\theta_{lk}^*} \text{ with } \theta_{lk}^* \sim H, (w_{lk})_{l=1}^{\infty} \sim \text{GEM}(\gamma)$$

and then draw the group specific distributions G_j from the following mixture

$$G_j \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*} \text{ with } (\pi_k)_{k=1}^{\infty} \sim \text{GEM}(\alpha)$$

The process ensures G_j in different groups can select the same G_k^*, leading to clustering of groups.

Although the NDP can also borrow information across groups, groups belonging to different clusters cannot share any atoms. For the NDP, the different distributions have either the same atoms with the same weights or completely different atoms and weights. This makes it different from the HDP.

The Hybrid Nested/Hierarchical Dirichlet Process

We propose the hybrid nested/hierarchical Dirichlet process (hNHDP) mixture model for groups of data. Our motivation is to improve the HDP by uncovering the latent categories of the groups. At the same time, we want to remain the properties of HDP. Following the setting of HDP, assume that we have M groups of data. Each group is denoted as $\mathbf{x}_j = x_{j1}, ..., x_{jN_j}$, where $\{x_{ji}\}$ are observations and N_j is the number of observations in group j. Each x_{ji} is associated with a distribution $p(\theta_{ji})$ with parameter θ_{ji}. For example, in topic modeling the distribution p is a multinomial distribution. We now describe the generative process of observations using the hNHDP model.

As in the NDP, we first consider the set of distributions $\{F_k\}$ for different clusters. For each cluster (latent category) k, we model F_k as a combination of two components, G_0

[1] Here GEM stands for Griffiths, Engen, and McCloskey(2002). We say $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ if we have $\beta_k = \beta_k' \prod_{k=1}^{k-1}(1-\beta_k')$ for $k = 1, ..., \infty$, where $\beta_k' \sim \text{Beta}(1, \gamma)$.

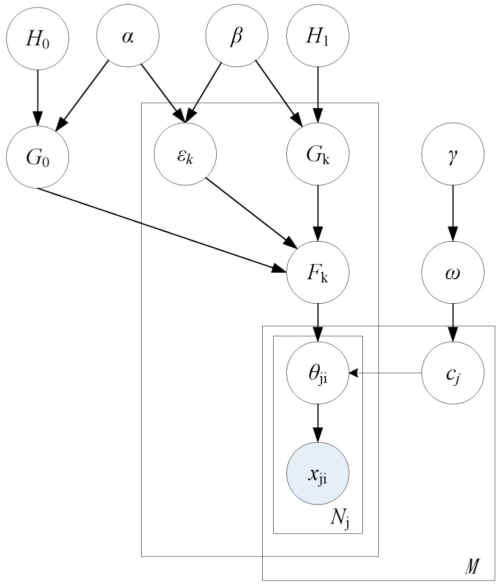


Figure 1: Graphical model representation of hNHDP.

and G_k. This setting is similar to that for the LC-HDP, but we impose some additional restrictions on the parameters. The combination weight ϵ_k is changed for each cluster, and the two components are drawn from DPs with different base measures.

$$\begin{aligned} G_0 &\sim \text{DP}(\alpha, H_0), \\ G_k &\sim \text{DP}(\beta, H_1) \text{ for each } k, \\ \epsilon_k &\sim \text{Beta}(\alpha, \beta) \text{ for each } k, \\ F_k &= \epsilon_k G_0 + (1-\epsilon_k) G_k \text{ for each } k. \end{aligned} \tag{3}$$

After getting the cluster-specific distributions, we assign the group distributions F'_j to the set $\{F_k\}$. This hierarchy is the same as that of the NDP.

$$F'_j \sim \sum_{k=1}^{\infty} \omega_k \delta_{F_k} \tag{4}$$

where $\boldsymbol{\omega} = \{\omega_k\} \sim \text{GEM}(\gamma)$. This is equal to selecting a cluster label k for a group and then assigning F_k to the group as its distribution. Then we generate observations using the following process.

- For each object $\mathbf{x}_j$,
  - Draw a cluster label $c_j \sim \boldsymbol{\omega}$;
  - For each observation x_{ji}
    - $\theta_{ji} \sim F_{c_j}$;
    - $x_{ji} \sim p(x_{ji}|\theta_{ji})$.

A graphical model representation is shown in Figure 1.We can also define the hNHDP mixture model in another way. For each group, the observations are independently drawn from the distribution $\mathcal{P}_j(\cdot) = \int p(\cdot|\theta) d(F'_j(\theta))$ where F'_j is drawn from the hNHDP prior as above and $x_{ji} \sim \mathcal{P}_j$ for each i in group j.

Model Properties

The hNHDP has some interesting properties:

- (1) F_k is still a sample from a DP.
- (2) F'_j can share atoms that are generated from G_0.

In (Lin, Grimson, and Fisher III 2010), the authors proposed a new construction of DPs by three operations based on existing ones. This construction is also used to derive a coupled mixture model for groups of data (Lin and Fisher 2012). Here we cite one of the operations: the superposition.

Superposition 1 *Let $D_k \sim DP(\alpha_k B_k)$ for $k = 1, ..., K$ be independent DPs and $(c_1, ..., c_k) \sim Dir(\alpha_1, ..., \alpha_k)$. Then the stochastic convex combination of these DPs remains a DP,*

$$c_1 D_1 + \cdots + c_k D_k \sim DP(\alpha_1 B_1 + \cdots + \alpha_k B_k).$$

From the set of equations in (3), we can infer that cluster-specific distribution F_k in the hNHDP model is still a realization of DP.

$$F_k \sim \text{DP}(\alpha H_0 + \beta H_1). \tag{5}$$

With this form, the hNHDP can be transferred into a special NDP. However, the generative process of (5) is not the same as that of (3) because G_0 is only sampled once in (3). If we directly use form (5) for each cluster, H_0 will generate different atoms for each cluster.

Now we consider the relationship between the hNHDP and the LC-HDP. We ignore the clustering structure of the hNHDP, and focus on only the group-specific distributions $\{F'_j\}$. For each group j, F'_j can be written as $\epsilon_k G_0 + (1 - \epsilon_k) G'_j$ where $k = c_j$ is the cluster label and $G'_j = G_k$. If ϵ_k is same for all k,the hNHDP degenerates into a special LC-HDP. It also indicates that F'_j can share atoms generated from a global component G_0.

Application to Topic Modeling

Motivation

Topic modeling is always an important application of the Bayesian nonparametric models. It is also suitable to illustrate the motivation and the power of the hNHDP, for category information is very useful for topic modeling. For example, the discriminative LDA(Lacoste-Julien, Sha, and Jordan 2008) and the labeled LDA(Ramage et al. 2009), which utilized category (or label) information of documents, had better predictive performance than the general unsupervised LDA (latent Dirichlet allocation (Blei, Ng, and Jordan 2003)). Now consider the HDP for topic modeling. In the case of modeling documents in multiple corpora (each corpus is a category), Teh et al. (Teh et al. 2006) demonstrated that a 2-level HDP that treats documents from different corpora in the same way performs much worse than a 3-level HDP that considers the category information of documents . All these studies proved the advantage of discriminating documents from different categories in text modeling. This stimulated us to take consideration of the document structure and to utilize it.

In practice , however, we often do not know category information in the topic modeling task. In this case, we could apply the hNHDP model to topic modeling to discover both latent categories and latent topics. We assume that word is an observation and that a document is a group. We generate the parameter distribution F'_j for each document j using the generative process in Section , where the distribution $p(\theta_{ji})$ is set as a multinomial distribution with parameter θ_{ji}. The base measures H_0 and H_1 are set as Dirichlet distributions over words.

hNHDP for Topic Modeling with Word Differentiation

Feature selection is an critical process in clustering analysis,. By selecting a subset of efficient features, feature selection can improve the text clustering efficiency and performance (Liu et al. 2003). Feature selection has already been successfully used in the Dirichlet mixture models for clustering (Kim, Tadesse, and Vannucci 2006; Yu, Huang, and Wang 2010; Wang and Blei 2009). So we also want to integrate feature selection into our model to reduce the dimension of topics and avoid noisy words in clustering. Our solution is word differentiation as follows.

Assuming that the size of the vocabulary is V, we bring in a binary vector $\mathbf{q^1} = (q_1^1, ..., q_V^1)$ to select discriminative words and separate the vocabulary into two disjoint sets. If $q_v^1 = 1$, the word v is regarded as discriminative and included only in local topics. Otherwise, v is regarded as global and included only in global topics. In this way we get two disjoint base measures, H_0 and H_1, for the hNHDP.

- For each word v
  - $q_v^0 \sim$ Bernoulli(π).
  - $q_v^1 = 1 - q_v^0$.
- $H_0 = \text{Dir}(\eta \mathbf{q^0})$;
- $H_1 = \text{Dir}(\eta \mathbf{q^1})$.

Here, $\mathbf{q^0}$ and $\mathbf{q^1}$ are binary vectors $\mathbf{q^0} = (q_1^0, ..., q_V^0)$ and $\mathbf{q^1} = (q_1^1, ..., q_V^1)$. For the parameter π, we set $\pi \sim$ Beta(α, β) to conform with the hyperparameters of ϵ_k. In practice, we may not require that the q_v^0 is uncertain if we already know the feature words; and it is also possible that $q_v^1 \neq 1 - q_v^0$. However, these cases are not discussed in this paper.

Finite Approximation

Samplers for the NDP based on Pólya Urns are in general infeasible(Rodriguez, Dunson, and Gelfand 2008). Our hNHDP encounters the similar problem, so here we use a finite approximation. In Bayesian statistics, the Dirichlet-multinomial allocation (DMA) has often been applied as a finite approximation to the DP (Yu, Yu, and Tresp 2005; Yu, Huang, and Wang 2010). It takes the form $G_N = \sum_{l=1}^{N} \pi_l \delta_{\theta_l}$, where $\pi = (\pi_{\mathbf{1}}, ..., \pi_{\mathbf{N}})$ is an N-dimensional vector distributed as a Dirichlet distribution $Dir(\alpha/N, ...\alpha/N)$. In our inference step, we approximate ω in (4) by a finite Dirichlet distribution

$$\omega \sim \text{Dir}(\gamma/K, ...\gamma/K). \tag{6}$$

The G_0 and G_1 are also approximated by the DMA.

$$G_0 = \sum_{l=1}^{L} w_{l0} \delta_{\theta_{l0}}$$
$$G_k = \sum_{l=1}^{L} w_{lk} \delta_{\theta_{lk}} \tag{7}$$

where $(w_{10}, ..., w_{L0}) \sim \text{Dir}(\alpha/L, ...\alpha/L)$ and $(w_{1k}, ..., w_{Lk}) \sim \text{Dir}(\beta/L, ...\beta/L)$. If we set K and L large, the DMA can give a good approximation in our model.

Inference

We use the Gibbs sampling method to infer the posterior of parameters. The inference proceeds through the following steps.

Sampling the cluster indicators c_j.

As we use the DMA approximation for ω in (6), the probability of cluster assignments conditioned on other variables can be calculated as

$$P(c_j = k | c_{-j}, ...) \propto \frac{m_k - 1 + \gamma/K}{M - 1 + \gamma} *$$
$$\prod_{x_{ji}} \sum_{l=1}^{L} \left(\epsilon_k w_{l0} P(x_{ji}|\theta_{l0}) + (1 - \epsilon_k) w_{lk} P(x_{ji}|\theta_{lk})\right),$$

where $p(x_{ji}|\theta_{lk}) = \theta_{lk}^v, v = x_{ji}$. M is the number of documents, and m_k is the number of documents assigned to cluster k.

Since the global words and local words are disjoint, we can re-write the upper equation as

$$P(c_j = k | c_{-j}, ...) \propto \frac{m_k - 1 + \gamma/K}{M - 1 + \gamma} *$$
$$\prod_{x_{ji} \in A_0} \sum_{l=1}^{L} \left(\epsilon_k w_{l0} P(x_{ji}|\theta_{l0})\right) *$$
$$\prod_{x_{ji} \in A_1} \sum_{l=1}^{L} \left((1 - \epsilon_k) w_{lk} P(x_{ji}|\theta_{lk})\right),$$

where $A_0 := \{v | q_v^0 = 1\}$ and $A_1 := \{v | q_v^1 = 1\}$ are the sets of global words and local words.

Sampling topic assignment z_{ji} for each word x_{ji}.

$$\begin{aligned} P(z_{ji} = t_{lk} | c_j = k, ...) &\propto (1 - \epsilon_k) * w_{lk} P(x_{ji}|\theta_{lk}) \\ P(z_{ji} = t_{l0} | c_j = k, ...) &\propto \epsilon_k * w_{l0} P(x_{ji}|\theta_{l0}), \end{aligned} \tag{8}$$

where t_{lk} and t_{l0} are topic indices.

Sampling the weights $\{w_{l0}\}$ and $\{w_{lk}\}$ for G_0 and G_1.

$$(w_{1k}, ..., w_{Lk}) \sim \text{Dir}(\beta/L + n_{1k}, ..., \beta/L + n_{Lk}) \tag{9}$$

$$(w_{10}, ..., w_{L0}) \sim \text{Dir}(\alpha/L + n_{10}, ..., \alpha/L + n_{L0}), \tag{10}$$

where n_{lk} is the number of words assigned to topic t_{lk}.

Sampling θ_{lk} and θ_{l0}

$$(\theta_{lk} | ...) \sim \text{Dir}(\eta q_1^1 + n_{lk}^1, ..., \eta q_V^1 + n_{lk}^V) \tag{11}$$

$$(\theta_{l0}|...) \sim \mathrm{Dir}(\eta q_1^0 + n_{l0}^1, ..., \eta q_V^0 + n_{l0}^V), \tag{12}$$

n_{lk}^v is the count when the word v assigned to topic lk.

Sampling ϵ_k

$$(\epsilon_k|...) \sim \mathrm{Beta}(\alpha + \sum_{l=1}^{L} n_{l0}, \beta + \sum_{l=1}^{L} n_{lk}) \tag{13}$$

Sampling $\mathbf{q}$. For the word selection variable $\mathbf{q}$ (including q^0 and q^1)[2], we use the Metropolis-Hastings algorithm. In each step, we randomly select a word v and invert its q_v value. When q changes, the associated θ (i.e. the collection of θ_{lk} and θ_{l0}) should also be changed. As it is difficult to integrate out θ for the posterior distribution of q, we update q with θ together. The new candidates $\mathbf{q}^*$ and θ^* are accepted with probability

$$\min\left\{1, \frac{P(\mathbf{q}^{0*}, \theta^*|\mathbf{c}, \mathbf{X}, ...)P(\mathbf{q}^0, \theta|\mathbf{q}^{0*}, \theta^*)}{P(\mathbf{q}^0, \theta|\mathbf{c}, \mathbf{X}, ...)P(\mathbf{q}^{0*}, \theta^*|\mathbf{q}^0, \theta)}\right\} \tag{14}$$

This is equal to

$$\min\left\{1, \frac{P(\mathbf{X}|\mathbf{q}^{0*}, \theta^*, ...)P(q_v^{0*})}{P(\mathbf{X}|\mathbf{q}^0, \theta, , ...)P(q_v^0)}\right\} \tag{15}$$

where X is the collection of all the documents.

Sampling π

$$(\pi|...) \sim \mathrm{Beta}(\alpha + N_0, \beta + N_1). \tag{16}$$

N_0 and N_1 are the numbers of unique words identified as global and local ones respectively. Notice that $N_1 \neq \sum_{l,k} n_{l,k}$, because N_1 counts each word only once.

Experiments

Document Modeling on Real Data

We implemented the proposed hNHDP model on two real-world text datasets. The first one is the *NIPS* data, which is used in (Teh et al. 2006)[3]. This version of *NIPS* data collects NIPS articles from 1988–1999 and unifies the section labels in different years. It contains 13649 unique words and 1575 articles separated into nine sections: algorithms and architectures, applications, cognitive science, control and navigation, implementations, learning theory, neuroscience, signal processing, and vision sciences. The other dataset is "6 conference abstracts (*6conf*)", which contains abstracts from six international conferences (IJCAI, SIGIR, ICML, KDD, CVPR, and WWW) collected by (Deng et al. 2011). It has 11,456 documents and 4083 unique words. All words are stemmed and stop words are removed.

We compared our model with other models on training sets of various sizes as in (Lin and Fisher 2012). Each dataset was separated into two disjoint sets, one for training and the other for testing. We generated 6 pairs of training/testing datasets for NIPS data and 5 pairs for 6conf data. For all the experiments, we used the same setting of parameter settings for our model. We gave the hyperparameter γ a vague value $Gamma(0.1, 1)$ and set $\eta = 0.5$ for H_0 and H_1. The component numbers in DMA approximation are set as $K = 100$, $L = 30$. The other parameters were $\alpha = \beta = 1$. For each training set, we ran 1000 iterations and treated the first 500 as burn-in. In the initialization step, we used a simple feature selection method which ranks words by term variance quality (Dhillon, Kogan, and Nicholas 2004). We selected a random proportion of highest ranked words as discriminative words, while the others were set as global words. This allowed us to accelerate the convergence in the sampling process.

The models used for comparison were the following:

- `HDP-LDA` (Teh et al. 2006). We used the 2-level HDP mixture model, which ignores the group information of documents. Articles from different sections were not treated differently. We followed the parameter setting procedure given in (Teh et al. 2006). The concentration parameters for the two levels were given as: $\gamma \sim \mathrm{Gamma}(5, 0.1)$, $\alpha \sim \mathrm{Gamma}(0.1, 0.1)$. The base measure of the bottom level was a symmetric Dirichlet distribution over all words with parameters of 0.5.
- `NDP`. This model is based on the nested Dirichlet process. In its settings, G_0 did not exist and $F_k = G_k$. Since the NDP model does not share topics between clusters, it does not distinguish either local topics or local words.
- `hNHDP-nosel`. For this model, we used the same structure as for the proposed hNHDP model. The difference was that here we set $H_0 = H_1 \sim Dir(\eta)$. The base measures H_0 and H_1 were symmetric Dirichlet distributions over all words. In other words, this model does not differentiate words between global and local, while it remains to distinct global topics from local topics.

We evaluated all the models with the test-set perplexity, a standard metric for document modeling.

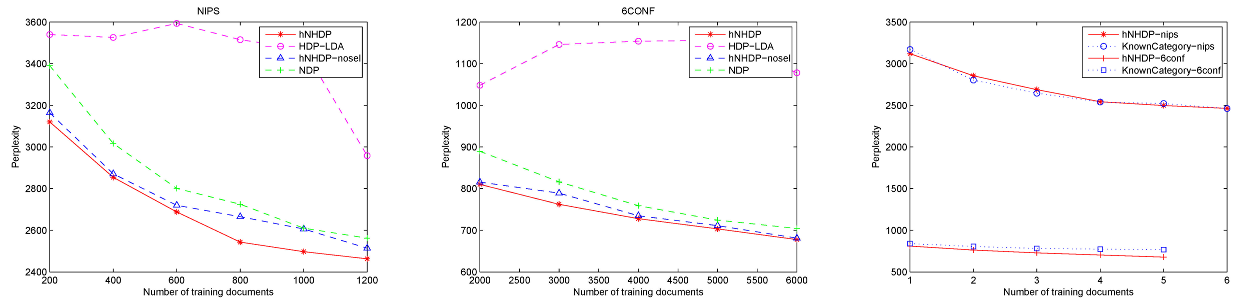
$$\mathrm{perplexity}(D_{test}) = \exp(-\frac{\sum_{d \in D_{test}} \log p(\mathbf{x}_d|D_{train})}{\sum_{d \in D_{test}} N_d}).$$

Figures 2(a) and 2(b) compare the perplexities on the two data sets. Our proposed model, hNHDP, achieved the best perplexities (lower is better) in all runs. Especially, it exceeded HDP-LDA by a large amount.

In addition, with the same setting for topic-word distributions (a symmetric Dirichlet distribution over all words), the hNHDP-nosel performed better than the NDP and HDP. The former demonstrates the advantage of sharing global topics and distinguishing local topics, while the latter indicates the effectiveness of taking advantage of the clustering structure of documents. We also noticed that the performance of the hHNDP was obviously better than that of the hHNDP-nosel only for some training sizes, while the two models got comparable results in other runs. This is reasonable because the vocabulary size was so large that the tail words of each topic may have contributed little. Thus, we may suppose that the global words in local topics and the local words in global topics contributed little to predictive performance, resulting in a result similar to the hNHDP's. Nevertheless, the hNHDP

[2] $\mathbf{q}^1$ is dependent on $\mathbf{q}^0$ via the equation $\mathbf{q}^1 = 1 - \mathbf{q}^0$, so we only need to consider $\mathbf{q}^0$ here.

[3] http://www.stats.ox.ac.uk/~teh/research/data/nips0_12.mat



(a) Results of document modeling on NIPS data.

(b) Results of document modeling on 6conf data.

(c) Comparison of hNHDP and hNHDP-knowncategory.

Figure 2: Perplexity Results on Real Data Sets

still achieved our aim by greatly reducing the model complexity without any performance decrease (in fact the performance increased a little). Its ability to extract local topics and local words is an additional benefit.

At last, we also wanted to know how well the latent categories found by hNHDP improves document modeling. So we developed another model, `hNHDP-knowncategory`, which assumes that the category labels of documents are known. It assigns real labels to documents in the hNHDP model and does not change them during iterations. Comparing the two models (Figure 2(c)), we found that the performance of hNHDP was comparable to `hNHDP-knowncategory` on NIPS data, while on 6conf data it was even better. These results indicate the efficiency of hNHDP for document modeling.

Clustering and Visualization

We then showed that hNHDP could find reasonable document structures and used them for visualization. We first present the clustering results on *6conf* data in Figure 3. Although the number of clusters inferred by our model is a little bigger than the real one, each conference has its specific clusters, which we can easily differentiate in the figure. Moreover, we could also find the connection between the conferences in Figure 3. CVPR is separate from the others, with only a little connection with ICML and IJCAI. ICML and KDD have a large overlap, but ICML has an additional cluster component. SIGIR and WWW also own the same major clusters although the cluster densities may differ. IJCAI is a comprehensive conference, so it includes several clusters shared by other conferences as well as a specific cluster.

We then extracted the typical topics in each cluster and matched them with corresponding conferences. The topics are shown in Table 1. The global topics/words and the local topics/words are easily distinguished in the table. The local topics conform to the features of different conferences. From Figure 3 and Table 1, we can see that both the clusters and the topics can be well explained and that they reveal the structure and features of the data. Although clustering is just a by-product of the hNHDP, we still made some quantitative evaluation of clustering results and demonstrated that hNHDP beat NDP. Limited to space, they are listed in supplemental materials. We also showed some simulated study of the model there to explain the iteration process.

Table 1: Some example topics extracted from *6conf* data. Each column is a topic; for each topic the top 12 words are shown. The numbers in brackets are the typical cluster numbers of each conference shown in Figure 3.

Global Topics	Local Topics		
	CVPR(23)	ICML(16)	SIGIR(31)
task	imag	learn	retriev
predict	model	algorithm	queri
experiment	recognit	problem	inform
work	object	method	model
describ	base	search	document
fast	track	reinforc	base
solut	segment	gener	system
requir	motion	optim	relev
specif	shape	base	languag
context	visual	model	term
express	detect	function	data
categor	estim	plan	effect

Conclusions and Future Work

We proposed an extension to the HDP model for modeling groups of data by taking advantage of the latent category information of groups. The hNHDP model clusters the groups and also allows the clusters to share mixture components. The application of the hNHDP to topic modeling illustrates the power of the new prior. We identify both local topics and local words in the model and use them for clustering documents. Experiments on document modeling with real datasets proved the advantage of the hNHDP.

In addition to document modeling, the hNHDP can also be used for other applications, such as multi-level clustering of patients and hospitals. Moreover, the global exponents can be replaced by some context information, leading to some context-based models. Important future work includes enhancing the computation efficiency.

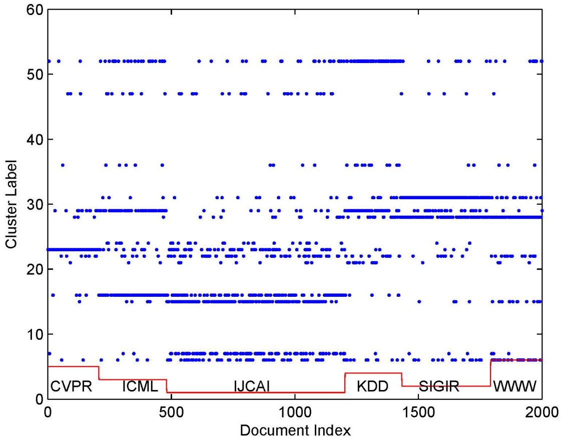


Figure 3: Visualization of the clustering results on *6conf* data. The red lines indicate the real labels, while the blue points indicate the clustering assignments.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Deng, H.; Han, J.; Zhao, B.; Yu, Y.; and Lin, C. X. 2011. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1271–1279. ACM.

Dhillon, I.; Kogan, J.; and Nicholas, C. 2004. Feature selection and document clustering. In *Survey of text mining*. Springer. 73–100.

Kim, S.; Tadesse, M. G.; and Vannucci, M. 2006. Variable selection in clustering via dirichlet process mixture models. *Biometrika* 93(4):877–893.

Lacoste-Julien, S.; Sha, F.; and Jordan, M. I. 2008. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, 897–904.

Lin, D., and Fisher, J. 2012. Coupling nonparametric mixtures via latent dirichlet processes. In *Advances in Neural Information Processing Systems 25*, 55–63.

Lin, D.; Grimson, E.; and Fisher III, J. W. 2010. Construction of dependent dirichlet processes based on poisson processes. *Neural Information Processing Systems Foundation (NIPS)*.

Liu, T.; Liu, S.; Chen, Z.; and Ma, W.-Y. 2003. An evaluation on feature selection for text clustering. In *ICML*, volume 3, 488–495.

Müller, P.; Quintana, F.; and Rosner, G. 2004. A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(3):735–749.

Pitman, J. 2002. Poisson-dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing* 11(5):501–514.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–256.

Rodriguez, A.; Dunson, D. B.; and Gelfand, A. E. 2008. The nested dirichlet process. *Journal of the American Statistical Association* 103(483).

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association* 101(476).

Wang, C., and Blei, D. M. 2009. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, 1982–1989.

Wulsin, D.; Litt, B.; and Jensen, S. T. 2012. A hierarchical dirichlet process model with multiple levels of clustering for human eeg seizure modeling. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 57–64.

Yu, G.; Huang, R.; and Wang, Z. 2010. Document clustering via dirichlet process mixture model with feature selection. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 763–772.

Yu, K.; Yu, S.; and Tresp, V. 2005. Dirichlet enhanced latent semantic analysis. In *In Conference in Artificial Intelligence and Statistics*, 437–444.