Exponential Family Hybrid Semi-Supervised Learning

Arvind Agarwal
School of Computing
University of Utah
arvind@cs.utah.edu

Hal Daumé III
School of Computing
University of Utah
hal@cs.utah.edu

Abstract

We present an approach to semi-supervised learning based on an exponential family characterization. Our approach generalizes previous work on coupled priors for hybrid generative/discriminative models. Our model is more flexible and natural than previous approaches. Experimental results on several data sets show that our approach also performs better in practice.

1 Introduction

Labeled data on which to train machine learning algorithms is often scarce or expensive. This has led to significant interest in semi-supervised learning methods that can take advantage of unlabeled data [Cozman *et al.*, 2003; Zhu, 2005]. While it is straightforward to integrate unlabeled data in a generative learning framework [Nigam *et al.*, 2000], it is not so in a discriminative framework. Unfortunately, it is well-known both empirically and theoretically [Ng and Jordan, 2002] that discriminative approaches tend to outperform generative approaches when there is enough labeled data. This has led to many recent developments in *hybrid* generative/discriminative models that are able to leverage the power of both frameworks (see Section 4). One particular such example is the work of Lasserre et al. [2006], who describe a hybrid framework ("PCP") in which a generative model and discriminative model are jointly estimated, using a *prior* that encourages them to have similar parameters.

In this paper, we generalize the *parameter coupling prior*[1] (PCP) method [Lasserre *et al.*, 2006] to arbitrary distributions belonging to the *exponential family*. Unlike the PCP method, we do not restrict ourselves to the *Gaussian prior*, but instead choose a prior that is natural to the model. Other authors [Bouchard, 2007] have also noted the inappropriateness of the Gaussian prior to couple the generative and discriminative models. Our resulting approach for hybridizing discriminative and generative models is: (1) not restricted to a particular class of the models; (2) more flexible in choosing the way these two models can be combined; (3) enables us to achieve a closed form solution for the generative parameters, unlike PCP method, where one has to resort to the numerical optimization. We demonstrate our framework on using a Beta/Binomial conjugate pair on the text categorization problems addressed by Druck et al. [2007].

[1] Terms *parameter coupling prior* and *coupled prior* were introduced in [Druck *et al.*, 2007] and do not appear in [Lasserre *et al.*, 2006] though they refer to the framework introduced in [Lasserre *et al.*, 2006].

2 Background

In general, machine learning approaches to classification can be divided into two categories: *generative approaches* and *discriminative approaches*. Generative approaches assume that the data is generated though an underlying process. One simple example is document categorization: for each example $(\mathbf{x}, y)$, we first choose a category y, and then produce a document $\mathbf{x}$ conditioned on the category y. The goal in generative modeling is to approximate the joint distribution $p(\mathbf{x}, y)$ that represents this process. On the other hand, discriminative approaches do not assume any underlying process and directly model the probability of category given the document $p(y|\mathbf{x})$. Ng and Jordan [2002] compare these two approaches and show that while discriminative models are asymptotically better than generative models, generative models need less data to train.

In semi-supervised settings, one has access to lots of unlabeled data but only a small amount of labeled data. It is easy to see that unlabeled data is not directly useful in a discriminative setting but can be easily used in generative setting. However, since discriminative methods asymptotically tend to outperform generative methods [Ng and Jordan, 2002], this naturally leads to combining these two approaches and building a hybrid model that does better than the individual models. Earlier work [Bouchard and Triggs, 2004; Lasserre *et al.*, 2006; Druck *et al.*, 2007] has shown the efficacy of the hybrid approach.

2.1 Exponential Family and Conjugate Priors

For the sake of completeness, we briefly define the exponential family which we will use as the basis of our hybrid model. The exponential family is a set of distributions whose probability density function[2] can be expressed in the following

[2] "Density" can be replaced by "mass" in the case of discrete random variable.

form:

$$f(\mathbf{x};\theta) = h(\mathbf{x}) \exp(\langle \eta(\theta) T(\mathbf{x}) \rangle - A(\theta)) \quad (1)$$

Here $T(\mathbf{x})$ is sufficient statistics, $\eta(\theta)$ is a function of natural parameters θ, and $A(\theta)$ is a normalization constant (also known as *log-partition function*).

One important property of the exponential family is the existence of conjugate priors. Given any member of the exponential family in Eq (1), the *conjugate prior* is a distribution over its *parameters* with the following form:

$$p(\theta|\alpha,\beta) = m(\alpha,\beta)\ \exp(\langle \eta(\theta), \alpha \rangle - \beta A(\theta))$$

Here $\langle a, b \rangle$ denotes the dot product of vectors a and b. Both α and β are hyperparameters of the conjugate prior. Importantly, function $A(\cdot)$ is the same between the exponential family member and the conjugate prior.

A second important property of the exponential family is the relationship between the log-partition function $A(\theta)$ and the sufficient statistics. In particular, we have:

$$\frac{\partial A}{\partial \theta} = \mathbb{E}_\theta \left[T(\mathbf{x}) \right] \quad (2)$$

2.2 Hybrid Model with Coupled Prior

We first define the problem and some of the notations that we will use through-out the paper. Our task is to learn a model that predicts a label y given an example $\mathbf{x}$. We are given the data $D = D_L \cup D_U$ where D_L represents the labeled data and D_U represents the unlabeled data. Each instance of the labeled data consists of a pair $(\mathbf{x}, y)$ where $\mathbf{x}$ is feature vector and y is the corresponding label. Each instance of unlabeled data consists of only feature vector $\mathbf{x}$. The $\mathbf{x}$s are M-dimensional feature vectors, and $\mathbf{x}_d$ denotes the d^{th} feature.

We now give a brief overview of the hybrid model presented by Lasserre et al. [2006]. The hybrid model is a mixture of discriminative and generative components, both of which have separate sets of parameters. These two sets of parameters (hence two models) are combined using a prior called *coupled prior*. Considering only one data point (the extension to multiple data points is straightforward and presented later), the model is defined as follows:

$$\begin{aligned} p(\mathbf{x}, y, \theta, \tilde{\theta}) &= p(\tilde{\theta}, \theta) p(y|\mathbf{x}, \theta) p(\mathbf{x}|\tilde{\theta}) \\ &= p(\tilde{\theta}, \theta) p(y|\mathbf{x}, \theta) \sum_{y'} p(\mathbf{x}, y'|\tilde{\theta}) \end{aligned}$$

Here θ is a set of discriminative parameters, $\tilde{\theta}$ a set of generative parameters, and $p(\tilde{\theta}, \theta)$ provides the natural coupling between these two sets of parameters. $p(y|\mathbf{x}, \theta)$ is the discriminative component; $p(\mathbf{x}|\tilde{\theta}) = \sum_{y'} p(\mathbf{x}, y'|\tilde{\theta})$ is the generative component.

The most important aspect of this model is the *coupled prior* $p(\tilde{\theta}, \theta)$, which interpolates the hybrid model between two extremes; generative model when $\theta = \tilde{\theta}$ and discriminative when θ is independent of $\tilde{\theta}$. In other cases, the goal of the coupled prior is to encourage the generative model and the discriminative model to have similar parameters. In earlier work [Lasserre *et al.*, 2006; Druck *et al.*, 2007], a Gaussian prior was used as the coupled prior:

$$p(\tilde{\theta}, \theta) \propto \exp\left[-\frac{1}{2\sigma^2} \left|\left| \tilde{\theta} - \theta \right|\right|^2 \right]$$

Unfortunately, the Gaussian prior is not always appropriate [Bouchard, 2007].

3 Exponential Family Hybrid Model

In this section, we provide a more general prior for the hybrid model that is not only mathematically convenient but also allows choosing a problem specific prior.

3.1 Exponential Family Generalization

First, we generalize the hybrid model defined in Section 2.2 for the distributions that come from the exponential family. In other words, all of the distributions (generative, discriminative and coupled prior) of the generalized hybrid model belong to the exponential family. We first provide the definitions of discriminative and generative models in terms of exponential family.

Generative model:

$$p(\mathbf{x}, y|\tilde{\theta}) = h(\mathbf{x}, y) \exp(\langle \tilde{\theta}, T(\mathbf{x}, y) \rangle - A(\tilde{\theta})) \quad (3)$$

Discriminative model:

$$p(y|\mathbf{x}, \theta) = g(y) \exp(\langle \theta, T(\mathbf{x}, y) \rangle - B(\theta, \mathbf{x})) \quad (4)$$

Next, we break the coupled prior $p(\tilde{\theta}, \theta)$ into two parts; an independent prior on the discriminative parameters $p(\theta)$ and a prior on the generative parameters given discriminative parameters $p(\tilde{\theta}|\theta)$. This formulation lets us model the dependency of the generative component over the discriminative component. Our new hybrid model is now defined as:

$$p(\mathbf{x}, y, \theta, \tilde{\theta}) = \left[p(\theta) p(y|\mathbf{x}, \theta) \right]\ p(\tilde{\theta}|\theta)\ \left[\sum_{y'} p(\mathbf{x}, y'|\tilde{\theta}) \right] \quad (5)$$

For convenience and interpretability (later we will show that it also improves the performance), we choose the coupled prior $p(\tilde{\theta}|\theta)$ to be conjugate with the generative model.

Conjugate prior:

$$p(\tilde{\theta}|\theta) = m(\theta)\ \exp(\langle \tilde{\theta}, \alpha(\theta) \rangle - \beta(\theta) A(\tilde{\theta})) \quad (6)$$

Here, $\alpha(\cdot)$ and $\beta(\cdot)$ are user-defined functions that map the discriminative parameters θ into hyperparameters for the conjugate prior. We discuss suitable choices of these functions in Section 3.4.

Substituting the exponential definitions of generative model Eq (3), discriminative model Eq (4), and coupled prior Eq (6) in Eq (5), and taking a log, we obtain a log joint probability of data and parameters:

$$\begin{aligned} L = \log p(\mathbf{x}, y, \theta, \tilde{\theta}) = \quad (7) \\ &\log p(\theta) + \\ &\log m(\theta) + \langle \tilde{\theta} \alpha(\theta) \rangle - \beta(\theta) A(\tilde{\theta}) + \\ &\log g(y) + \sum_{(\mathbf{x},y) \in D_L} \left[\langle \theta, T(\mathbf{x}, y) \rangle - B(\theta, \mathbf{x}) \right] + \\ &\sum_{\mathbf{x} \in D} \log \sum_{y'} \left[h(\mathbf{x}, y') \exp(\langle \tilde{\theta}, T(\mathbf{x}, y') \rangle - A(\tilde{\theta})) \right] \end{aligned}$$

Note that here discriminative part is defined only for labeled data while generative part is defined for both labeled and unlabeled data.

3.2 Parameter Optimization

We perform parameter optimization by a coordinate descent method, alternating between optimizing the discriminative parameters θ and optimizing the generative parameters $\tilde{\theta}$.

For the generative parameters, we take the partial derivative of the log probability in Eq (7) with respect to $\tilde{\theta}$:

$$\frac{\partial L}{\partial \tilde{\theta}} = \alpha(\theta) - \beta(\theta)A'(\tilde{\theta}) + \sum_{\mathbf{x}\in D}\sum_{y'} p(y'|\mathbf{x},\tilde{\theta})(T(\mathbf{x},y') - A'(\tilde{\theta}))$$

Here, $p(y'|\mathbf{x},\tilde{\theta})$ is the probability based on the parameters estimated in the last iteration $p(y'|\mathbf{x},\tilde{\theta}_{old})$. Substituting this in the above equation and setting it equal to zero, we obtain:

$$A'(\tilde{\theta}) = \frac{\sum_{\mathbf{x}\in D}\sum_{y'} p(y'|\mathbf{x},\tilde{\theta}_{old})T(\mathbf{x},y') + \alpha(\theta)}{N+\beta(\theta)} = \frac{\hat{\mathbb{E}}_{\mathbf{x}\sim D}\,\mathbb{E}_{y\sim\tilde{\theta}_{old}}(T(\mathbf{x},y')) + \alpha(\theta)}{N+\beta(\theta)} \quad (8)$$

Here $A'(\tilde{\theta})$ denotes the partial derivative of $A(\tilde{\theta})$ with respect to $\tilde{\theta}$. As discussed in Section 1, choosing a conjugate prior gives us a closed form solution for $A'(\tilde{\theta})$. From Eq (2), we know that $A'(\tilde{\theta})$ is equivalent to the expected sufficient statistics of the generative model.

Having solved for the generative parameters $\tilde{\theta}$, we now solve the hybrid model for discriminative parameters θ.

$$\frac{\partial L}{\partial \theta} = \frac{\partial \log p(\theta)}{\partial \theta} + \frac{\partial \log m(\theta)}{\partial \theta} + \tilde{\theta}\alpha'(\theta) - \beta'(\theta)A(\tilde{\theta}) + \sum_{(\mathbf{x},y)\in D_L}(T(y,\mathbf{x}) - B'(\theta,\mathbf{x})) \quad (9)$$

There is no closed form solution to the above expression therefore we solve it using numerical methods. In our implementation, we use stochastic gradient descent.

3.3 Hybrid Multiple Binomial Model

In this section, we see how this hybrid model can be applied in practice. We first choose a generative model that is suitable to our application. We next choose the coupled prior conjugate to the generative model. Since later on, we intend to use the hybrid model for the document classification task, we use a *naive Bayes*[3] (NB) model for the generative part and *logistic regression* for the discriminative part, akin to the study of Ng and Jordan [2002]. The generative part of our model (naive Bayes) is given by:

$$p(y,\mathbf{x}|\pi,v) = p(y|\pi)p(\mathbf{x}|y,v) = \prod_k \pi_k^{1_{\{y=k\}}} \prod_d v_{yd}^{\mathbf{x}_d}(1-v_{yd})^{1-\mathbf{x}_d} \quad (10)$$

[3]It should be noted that "naive Bayes" classifiers come into (at least) two different versions: the "multivariate Bernoulli version" and the "multinomial version" [Mccallum and Nigam, 1998]. Because of its generality, in our implementation, we use multivariate Bernoulli.

Here, $\sum_{y'}\pi_{y'} = 1$ and $0 \le v_{yd} \le 1$. $1_{\{y=k\}}$ is an indicator function that takes value 1 if $y = k$ and 0 otherwise.

The discriminative part is:

$$p(y|\mathbf{x},w,b) = \frac{1}{Z_{\mathbf{x}}}\exp\left(b_y + \sum_d \mathbf{x}_d w_{yd}\right) \quad (11)$$

Where $Z_{\mathbf{x}} = \sum_{y'}\exp\left(b_{y'} + \sum_d \mathbf{x}_d w_{y'd}\right)$ is a normalization constant. Note here that since these models form generative/discriminative pair, number of parameters is same in both models. It is easy to see that there is one-to-one relationship between these two sets of parameters. b_y in the discriminative model behaves similar to π_y in the generative model, and w_{yd} behaves similar to v_{yd}. Since w_{yd} and v_{yd} are the parameters that capture most of the information, we use coupled prior to couple these sets of parameters and do not couple b_y and π_y. It is important to note the difference between the canonical parameters of the exponential family representation of the model and the mean parameters. In the generative(or discriminative) model, $\tilde{\theta}_{yd}$ (or θ_{yd}) denote the canonical parameters while v_{yd} (or w_{yd}) denote the mean parameters.

Having defined the appropriate discriminative and generative models, now we can get equivalent exponential family forms of these models. First we show the exponential form of the generative model. The generative model in Eq (10) can be broken into two parts: one is class probability $p(y|\pi)$ and other class conditional probability $p(\mathbf{x}|y,v)$. Since the parameters of these distributions are independent, we can get their exponential representations separately. Considering the class conditional probability for one feature, Eq (10) can be written in the following form:

$$p(\mathbf{x}_d|y,v_{yd}) = \exp\left(\mathbf{x}_d \log\frac{v_{yd}}{1-v_{yd}} + \log(1-v_{yd})\right)$$

Comparing this with Eq (3) gives $\tilde{\theta}_{yd} = \log\frac{v_{yd}}{1-v_{yd}}$; $A(\tilde{\theta}_{yd}) = \log(1+e^{\tilde{\theta}_{yd}})$ and $T(y,\mathbf{x}) = \mathbf{x}_d$. Substituting these along with the appropriate conjugate prior in Eq (8) gives us a closed form solution for $A'(\tilde{\theta}_{yd})$, which, in the naive Bayes model is equal to v_{yd}.

$$A'(\tilde{\theta}_{yd}) = v_{yd} = \frac{\sum_{\mathbf{x}\in D} p(y|\mathbf{x},\tilde{\theta}_{old})\mathbf{x}_d + \alpha(\theta)}{N+\beta(\theta)} \quad (12)$$

In other words, v_{yd} is the normalized expected count of the d_{th} feature in class y, with smoothing parameters that are controlled by the coupled prior hyperparameters $\alpha(\theta)$ and $\beta(\theta)$.

Next we solve for π_y by directly optimizing the objective function Eq (10) with respect to π_y with the given constraints. This gives us $\pi_y = \frac{\sum_{\mathbf{x}\in D} p(y|\mathbf{x},\tilde{\theta}_{old})}{N}$ which is the normalized expected number of examples in class y.

Having solved for generative parameters, we now solve for the discriminative parameters. Ideally, we would like to first get an equivalent exponential form of Eq (11) and then solve it using Eq (9). Since Eq (9) is only defined for discriminative parameters that are coupled (w), we can not use Eq (9) unless we break Eq (11) into two exponential forms separate for w and b and, it is not clear how to do so. Therefore, we solve for discriminative parameters directly, without converting Eq (11) into exponential form. It is important to note here

that mean parameters w in Eq (11) is equal to the canonical parameters θ_{yd}. We place Gaussian prior $p(\theta) = N(\theta|0, \sigma^2)$ on $w = \theta$ and an improper uniform prior on b. Taking derivatives, we obtain:

$$\frac{\partial L}{\partial w_{yd}} = -\frac{w}{\sigma^2} + \frac{\partial \log m(w_{yd})}{\partial w_{yd}} + \langle \tilde{\theta}_{yd}\alpha'(w_{yd})\rangle - \beta'(w_{yd})A(\tilde{\theta}_{yd}) + \sum_{(\mathbf{x},y')\in D_L} \left\{ 1_{\{\mathbf{x}_d=1\}} - \frac{1}{Z_{\mathbf{x}}} \exp(b_{y'} + \sum_d \mathbf{x}_d w_{y'd}) 1_{\{\mathbf{x}_d=1\}} \right\}$$

$$\frac{\partial L}{\partial b_y} = \sum_{(\mathbf{x},y')\in D_L} \left\{ 1_{\{y=y'\}} - \frac{1}{Z_{\mathbf{x}}} \exp(b_y + \sum_d \mathbf{x}_d w_{yd}) \right\}$$

3.4 Conjugate Beta Prior

Recall that our conjugate prior crucially depends on two functions: $\alpha(\theta)$ and $\beta(\theta)$ that "convert" the discriminative parameters θ into a prior on the generative parameters $p(\tilde{\theta}|\theta)$. In the case of the binomial likelihood, the conjugate prior is Beta. Exponential form of Beta prior is defined as:

$$p(\tilde{\theta}_{yd}|\theta_{yd}) = m(\theta_{yd}) \exp(\tilde{\theta}_{yd}\alpha(\theta_{yd}) - \beta(\theta_{yd})A(\tilde{\theta}_{yd}))$$

Where $m(\theta_{yd}) = \frac{\Gamma(\beta(\tilde{\theta}_{yd})+2)}{\Gamma(\alpha(\tilde{\theta}_{yd})+1)\Gamma(\beta(\theta_{yd})-\alpha(\theta_{yd})+1)}$ and $A(\tilde{\theta}_{yd} = \log(1 + e^{\tilde{\theta}_{yd}})$.

We select the function $\alpha(\theta_{yd})$ and $\beta(\theta_{yd})$ to be such that: (1) the *mode* of the conjugate prior is θ_{yd} and (2) the *variance* of the conjugate prior is controllable by the hyperparameter γ. As noted from Figure 1, as γ goes to ∞, variance goes to 0 and prior forces generative parameters to be equal to the discriminative parameters (pure generative model) and as γ goes to 0, variance goes to ∞ which implies the independence between generative and discriminative parameters (pure discriminative model). Other values of γ interpolate between these two extremes. Thus, we choose $\alpha(\theta_{yd}) = \gamma/(1 + e^{-\theta_{yd}})$ and $\beta(\theta_{yd}) = \gamma$. This gives mode of $p(\tilde{\theta}_{yd}|\theta_{yd})$ at θ_{yd} with the variance that decreases in γ, as desired.

It is important to note that our choice of hyperparameters for the conjugate prior is not specific to this example, but holds true in general. In the general case, let A be the log-partition function associated with the generative model, then, the conjugate prior hyperparameters should be $\alpha(\theta) = \gamma A'(\theta)$ and $\beta(\theta) = \gamma$. This gives us the mode of conjugate prior at θ with the variance that decreases in γ. In the beta/binomial hybrid model, $A'(\theta) = A'(w) = 1/(1+e^{-w})$. Also note that in the beta/binomial example, $A'(\theta)$ is also the transformation function T that transforms the discriminative mean parameters w to the generative mean parameters v.

In Figure 1, we also compare the Beta prior (solid blue curves) to an "equivalent" logistic-Normal prior (dashed black curves) for four settings of γ. The logistic-Normal is parameterized to have the same mode and variance as the Beta prior. As we can see, for high values of γ (wherein the model is essentially generative), the two behave quite similarly. However, for more moderate settings of γ, the priors are qualitatively quite different.

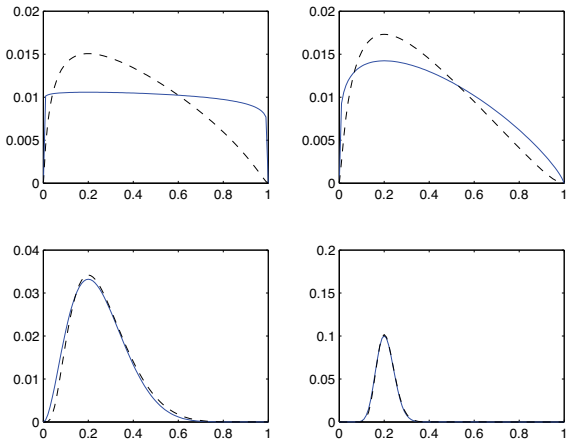

Figure 1: Effect of gamma on the Beta prior (solid curve) and logistic-Normal prior (dashed curve) for gamma=0.1, 1, 10, 100 (top-left, top-right, bottom-left, bottom-right) and for the transformed discriminative parameter $T(w) = 0.2$

4 Related Work

There have been a number of efforts to combine generative and discriminative models to obtain a hybrid model that performs better than either individually. Some of the earlier works [Raina *et al.*, 2003; Bouchard and Triggs, 2004] use completely different approaches to hybridize these models; Raina et al. [2003] present a model for the document classification task where a document is split into multiple regions and complementary properties of generative/discriminative models are exploited by training a large set of the parameters generatively and only a small set of parameters discriminatively. Bouchard and Triggs [2004] build a hybrid model by taking a linear combination of generative and discriminative model. This model is similar to the multi-conditional learning model presented by McCallum et al. [2006]. Jaakkola and Haussler [1999] describe a scheme in which the kernel of a discriminative classifier is extracted from a generative model. Though these models have shown to perform better than just the discriminative or generative model, none of them combine the hybrid model in natural way.

Our work builds on the work of Lasserre et al. [2006] and Druck et al. [2007], which are discussed in Section 2.2. Along these lines, Fujino et al. [2007] present another hybrid approach where a generative model is trained using a small number of labeled examples. Since the generative model has high bias, a generative "bias-correction" model is trained in a discriminative manner to discriminatively combine the bias-correction model with the generative model. Most of these work focus on the application and little on the theory of the hybrid model. There has been a recent work by Bouchard [2007] that presents a unified framework for the "PCP" model and the "convex-combination" model [Bouchard and Triggs, 2004], and proves performance properties.

Dataset	No. of Features	Dataset description
movie	24,841	classifies the sentiments of the review of the movies from IMDB as *positive* or *negative*
webkb	22,824	classifies webpages from university as *student, course, faculty* or *project*
sraa	77,494	classifies messages by the newsgroup to which they were posted: *simulated-aviation*, *real-aviation*, *simulated-autoracing*, *real-autoracing*

Table 1: Description of the datasets used in the experiments

Dataset	pcp-Gauss Mult	pcp-Gauss Bin	pcp-Beta Bin	% change
movie (10)	64.6	63.4 (3.2)	**68.3** (5.5)	+7.7%
movie (25)	68.6	69.0 (1.5)	**76.7** (1.2)	+11.1%
webkb (10)	72.5	73.7 (3.7)	**75.3** (2.9)	+2.2%
webkb (25)	76.7	83.8 (1.3)	**83.9** (1.6)	+1.1%
sraa (10)	**81.6**	67.7 (6.8)	79.1 (4.0)	+16.8%
sraa (25)	84.1	76.6 (3.5)	**86.1** (1.0)	+12.4%

Table 2: Comparative results for pcp with Gaussian prior and pcp with Beta prior. Parenthesized values denote the number of labeled examples per class and the standard deviation.

5 Experiments

5.1 Experimental Setup

In this section, we show empirical results of our approach and compare them with the existing (and most related to our method) state-of-the-art semi-supervised methods [Druck *et al.*, 2007]. In order to have a fair comparison, we use experimental setup of Druck et al. [2007] and perform experiments only for the datasets where PCP model have shown to perform best, There are three such datasets: *movie, sraa* and *webkb*. Description of these datasets is given in Table 1.

Although all of the examples in these datasets are labeled, we perform experiments by taking a subset of dataset as labeled and treating the rest of the examples as unlabeled. We use either 10 or 25 labeled examples from each class and vary unlabeled examples from 0 to a maximum of 1000. Number of unlabeled examples are same in each class. We show our results for two sets of experiments: (1) we show how performance varies as we vary the number of unlabeled examples; (2) we show how performance varies with respect to λ. Here λ normalizes the $\gamma \in [\infty, 0]$ in the range of $[0, 1]$ using $\gamma = ((1-\lambda)/\lambda)^2$. Now $\lambda = 0$ corresponds to the pure generative case while $\lambda = 1$ corresponds to the pure discriminative case. As in the work of Druck et al. [2007], the success of the semi-supervised learning depends on the quality of the labeled examples, therefore we choose five random labeled sets and report the average on them. In our results, we report the percentage classification accuracy which is the ratio of number of examples correctly classified to the total number of test examples.

5.2 Results and Discussion

Results on the above mentioned three datasets are presented in Table 2. Table shows the results for the PCP model with the Gaussian prior (PCP-Gauss) and with the Beta prior (PCP-Beta). Since PCP-Beta uses the binomial version of NB, we reimplemented the PCP-Gaussfor the binomial version of NB and compare the results with it. Though we also show the results for PCP-Gauss multinomial [Druck *et al.*, 2007], a fair comparison would be to compare only binomial models. %change is the change in PCP-Beta with respect to the PCP-Gauss binomial version. As we see, PCP-Beta performs better than PCP-Gauss binomial in all experiments and better than PCP-Gauss multinomial in all experiments except sraa(10). Compared to PCP-Gauss binomial, PCP-Beta performs significantly better on sraa and movie datasets.

Comparing multinomial and binomial versions of PCP-Gauss, we see that for movie and webkbb datasets, binomial version performs better (or almost equal) than the multinomial while for sraa dataset, multinomial performs better. We conjecture that reason for this behavior could be because sraa has a large number of features and feature independence assumption is less violated in multinomial NB than in binomial NB. When datasets do not have too many features, binomial version tends to perform better because binomial NB accounts for both presence and absence of the features, in contrast to multinomial NB which only accounts for the presence of the features.

Figure 2 and Figure 3 show the results for accuracy vs. λ for different number of unlabeled examples for *sraa* and *movie* datasets respectively. Remember that $\lambda = 0$ is the purely generative model and $\lambda = 1$ is the purely discriminative model. In both of these figures, we see that as we increase the number of unlabeled examples, performance improves. In *sraa*, we observe that increasing the number of unlabeled examples results in the shifting of optimal λ (λ^*) towards rights. We get an optimal $\lambda^* = 0.2$ for a fully supervised model while for 1000 unlabeled examples, we get $\lambda^* = 0.5$. All the curves in this experiment are uni-modal which means that there is a unique value of λ where hybrid model performs best.

Unfortunately, these nearly-perfectly shaped curves are not common to all settings. We do not observe it in the other dataset (Figure 3). There are values of λ where a fully supervised model performs better than the best semi-supervised model. This experiment emphasizes the need for choosing the right value of λ and also shows the importance of the hybrid model. If we do not choose the right value of λ, we might end up hurting the model by using the unlabeled data. We also observe that *movie* dataset gives us a bi-modal curve in contrast to the uni-modal curve obtained in the *sraa*. We see that curve is a uni-modal in the supervised setting but as we introduce unlabeled examples, the curves not only become bi-modal but also shift towards the left-hand side (best accuracy is achieved close to the generative end). This naturally suggests that generative model is actually affecting the hybrid model in a positive manner and exploiting the strength of the unlabeled examples.

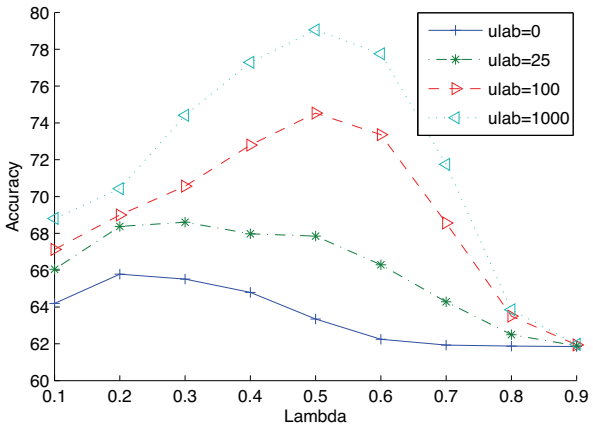

Figure 2: Results for sraa dataset for different number of unlabeled examples. Number of labeled examples=10.

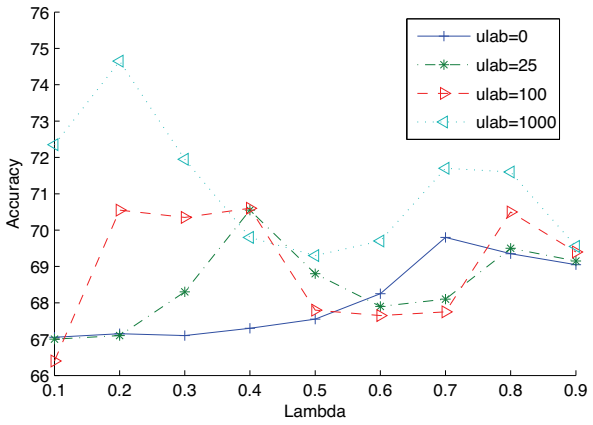

Figure 3: Results for movie dataset for different number of unlabeled examples, Number of labeled examples=25.

6 Conclusion and Future Work

We have presented a generalized "PCP" hybrid model for the exponential family distributions and have experimentally shown that the prior conjugate to the generative model is more appropriate than the Gaussian prior. In addition to the performance advantage, the conjugate prior also gives us a closed form solution for the generative parameters. In the future, we aim at interpreting these results in a theoretical way and answer questions like: (1) Under what conditions will the hybrid model perform better than both the generative and discriminative models? (2) What is the optimal value of γ? (3) Is a PAC-style analysis of the hybrid model possible for the finite sample case as opposed the asymptotic analysis mostly found in the literature?

References

[Bouchard and Triggs, 2004] G. Bouchard and Bill Triggs. The tradeoff between generative and discriminative classifiers. In *IASC International Symposium on Computational Statistics*, pages 721–728, Prague, August 2004.

[Bouchard, 2007] Guillaume Bouchard. Bias-variance tradeoff in hybrid generative-discriminative models. In *Proceedings of the Sixth International Conference on Machine Learning and Applications*, pages 124–129, Washington, DC, USA, 2007. IEEE Computer Society.

[Cozman *et al.*, 2003] Fabio Gagliardi Cozman, Ira Cohen, Marcelo Cesar Cirelo, and Escola Politcnica. Semi-supervised learning of mixture models. In *20th International Conference on Machine Learning*, pages 99–106, 2003.

[Druck *et al.*, 2007] Gregory Druck, Chris Pal, Andrew McCallum, and Xiaojin Zhu. Semi-supervised classification with hybrid generative/discriminative methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–289, New York, NY, USA, 2007. ACM.

[Fujino *et al.*, 2007] Akinori Fujino, Naonori Ueda, and Kazumi Saito. A hybrid generative/discriminative approach to text classification with additional information. *Inf. Process. Manage.*, 43(2):379–392, 2007.

[Jaakkola and Haussler, 1999] Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1999.

[Lasserre *et al.*, 2006] Julia A. Lasserre, Christopher M. Bishop, and Thomas P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 87–94, Washington, DC, USA, 2006. IEEE Computer Society.

[Mccallum and Nigam, 1998] A. Mccallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI Workshop on "Learning for Text Categorization"*, 1998.

[McCallum *et al.*, 2006] Andrew McCallum, Chris Pal, Greg Druck, and Xuerui Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 433–439, 2006.

[Ng and Jordan, 2002] Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[Nigam *et al.*, 2000] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, V39(2):103–134, May 2000.

[Raina *et al.*, 2003] Rajat Raina, Yirong Shen, Andrew Y. Ng, and Andrew McCallum. Classification with hybrid generative/discriminative models. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.