A Hybrid Generative/Discriminative Approach to Semi-supervised Classifier Design

Akinori Fujino, Naonori Ueda, and **Kazumi Saito**
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
{a.fujino,ueda,saito}@cslab.kecl.ntt.co.jp

Abstract

Semi-supervised classifier design that simultaneously utilizes both labeled and unlabeled samples is a major research issue in machine learning. Existing semi-supervised learning methods belong to either generative or discriminative approaches. This paper focuses on probabilistic semi-supervised classifier design and presents a hybrid approach to take advantage of the generative and discriminative approaches. Our formulation considers a generative model trained on labeled samples and a newly introduced bias correction model. Both models belong to the same model family. The proposed hybrid model is constructed by combining both generative and bias correction models based on the maximum entropy principle. The parameters of the bias correction model are estimated by using training data, and combination weights are estimated so that labeled samples are correctly classified. We use naive Bayes models as the generative models to apply the hybrid approach to text classification problems. In our experimental results on three text data sets, we confirmed that the proposed method significantly outperformed pure generative and discriminative methods when the classification performances of the both methods were comparable.

Introduction

In conventional classifier design, a classifier is trained only on *labeled* samples. To obtain a better classifier with high generalization ability, a large amount of training samples are usually required. In practice, however, labeled samples are often fairly expensive to acquire because class labels are identified by experienced analysts. In contrast, *unlabeled* samples can often be inexpensively collected. For example, a large amount of unlabeled text samples are available from the web. Developing *semi-supervised* classifier design algorithms that learn from both labeled and unlabeled samples and take advantage of unlabeled samples is a major research issue in machine learning (Nigam et al. 2000; Grandvalet and Bengio 2005; Szummer and Jaakkola 2001; Amini and Gallinari 2002; Blum and Mitchell 1998; Zhu, Ghahramani, and Lafferty 2003). See (Seeger 2001) for a comprehensive survey.

Copyright © 2005, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Semi-supervised classifier design algorithms have been proposed for *generative* and *discriminative* classifiers. Generative classifiers learn the joint probability model, $P(\boldsymbol{x}, y)$, of input $\boldsymbol{x}$ and class label y, and make their predictions by using the Bayes rule to compute $P(y|\boldsymbol{x})$, and then taking the most probable label y. Unlabeled samples are dealt with a missing class label problem in mixture models (Nigam et al. 2000).

Discriminative classifiers, on the other hand, model posterior class probabilities $P(y|\boldsymbol{x})$ for all classes directly and learn mapping from $\boldsymbol{x}$ to y. Since $P(x)$ is not modeled in the discriminative approach, some assumptions are required to incorporate unlabeled samples into the model. Szummer and Jaakkola (2001) utilized the assumption that if two feature vectors are close, then class labels of both samples should be the same. Very recently Grandvalet and Bengio (2005) introduced *entropy regularizer* (ER) to semi-supervised learning. Utilizing the knowledge that unlabeled samples are beneficial for improving classification accuracy when samples are well separated among classes, Grandvalet and Bengio (2005) try to minimize the entropy of class posteriors.

It has been shown that discriminative classifiers often get better classification performance than generative classifiers. However, it has also been reported that when the number of labeled training samples is small, generative classifiers often obtain higher test set accuracy than discriminative classifiers (Ng and Jordan 2002). To take advantage of both approaches, in this paper we explore a *hybrid* model that is partly generative and partly discriminative.

In a *fully supervised* learning framework, such hybrid methods have recently been proposed (Tong and Koller 2000; Raina, Ng, and McCallum 2004). Tong and Koller (2000) propose a restricted Bayes classifier in which a Bayes optimal classifier is modified based on the maximum margin classification. They showed that hybrid classifier increased classification performance when training set contained samples with missing feature values; but the missing label problem has never considered. In (Raina, Ng, and McCallum 2004), every input feature vector is divided into R subvectors, each of which is modeled on the naive Bayes assumption; *weight* parameters to combine these *subgenerative* models are determined by maximizing class posterior likelihood. That is, the model combination is *discriminatively* performed. They applied the method to document

classification where each document consists of "subject" and "body" parts ($R = 2$) and experimentally showed that the hybrid classifier achieved more accurate classification with $R = 2$ than a pure generative ($R = 1$) classifier. Since the word distributions of "subject" and "body" may differ from each other, it is reasonable that such submodeling increases classification performance.

Inspired by hybrid modeling, we present a new semi-supervised hybrid classifier design method using both labeled and unlabeled samples. In our formulation, a generative model is trained on only a small amount of labeled samples. Clearly, the trained classifier has high bias. Thus, we newly introduce a model for *bias correction* that belongs to the same model family as the trained generative model. The parameters of the bias correction model are estimated by using training samples. Then, these models are discriminatively combined based on the *maximum entropy* (ME) principle (Berger, Della Pietra, and Della Pietra 1996). The use of the ME principle has already seen in (Nigam, Lafferty, and McCallum 1999), but they did not deal with the unlabeled data problem.

Using naive Bayes models as the generative and bias correction models, we apply the proposed method to text classification. Using three test collections, we experimentally show that the hybrid approach can also be effective in semi-supervised settings and also give discussion when the hybrid approach outperforms the pure generative and discriminative approaches.

Conventional Approaches

Semi-supervised Learning

In multiclass (K classes) classification problems, one of K classes $y \in \{1, ..k, ..K\}$ is assigned to a feature vector $\boldsymbol{x}$ by a classifier. In semi-supervised learning, the classifier is trained on not only labeled sample set $D_l = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, but also unlabeled sample set $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$. Usually, M is much greater than N. We require a framework to incorporate unlabeled samples without class labels y into classifiers. First, we briefly review the conventional approaches in the followings.

Generative Approach

Generative classifiers learn a joint probability model $P(\boldsymbol{x}, y|\Theta)$, where Θ is a model parameter. The class posteriors $P(y|\boldsymbol{x}, \Theta)$ for all classes are computed using the Bayes rule after parameter estimation. The class of $\boldsymbol{x}$ is determined as y that maximizes $P(y|\boldsymbol{x}, \Theta)$. The joint probability model is designed according to classification tasks: for example, a multinomial model for text classification or a Gaussian model for continuous feature vectors.

In the probabilistic framework, unlabeled samples are dealt with the missing class labels in mixture models (Dempster, Laird, and Rubin 1977). That is, $\boldsymbol{x}_m \in D_u$ is drawn from the marginal generative distribution $P(\boldsymbol{x}|\Theta) = \sum_{k=1}^K P(\boldsymbol{x}, k|\Theta)$. Model parameter Θ is computed by maximizing the posterior $P(\Theta|D)$ (MAP estimation). According to the Bayes rule, $P(\Theta|D) \propto P(D|\Theta)P(\Theta)$, the objective function of MAP estimation is given by

$$J(\Theta) = \sum_{n=1}^N \log P(\boldsymbol{x}_n, y_n|\Theta) + \sum_{m=1}^M \log \sum_{k=1}^K P(\boldsymbol{x}_m, k|\Theta) + \log P(\Theta). \quad (1)$$

Here, $P(\Theta)$ is a prior over the parameters. The value of Θ that maximizes $J(\Theta)$ is obtained by using Expectation-Maximization (EM) algorithm.

The estimation of Θ is affected by the number of unlabeled samples used with labeled samples. In other words, when $N \ll M$, model parameter Θ is estimated as almost unsupervised clustering because the second term on the RHS of Eq. (1) becomes much more dominant than the first term. Then, training the model by using unlabeled samples might not be useful for classification accuracy if mixture model assumptions are not true for actual classification task. To mitigate the problem, Nigam et al. (2000) introduced a weighting parameter λ that decreases the contribution of the unlabeled samples to the parameter estimation (EM-λ). Weighting parameter $\lambda \in [0, 1]$ is multiplied in the second term on the RHS of Eq. (1). The parameter value is determined by cross-validation so that the leave-one-out labeled samples are correctly classified as much as possible.

Discriminative Approach

Discriminative classifiers directly model posterior class probabilities $P(y|\boldsymbol{x})$ for all classes. In multinomial logistic regression, the posterior class probabilities are modeled as

$$P(k|\boldsymbol{x}, W) = \frac{\exp(\boldsymbol{w}_k \cdot \boldsymbol{x})}{\sum_{k'=1}^K \exp(\boldsymbol{w}_{k'} \cdot \boldsymbol{x})}, \ \forall k, \quad (2)$$

where $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K\}$ is a set of unknown model parameters. $\boldsymbol{w}_k \cdot \boldsymbol{x}$ represents the inner product of $\boldsymbol{w}_k$ and $\boldsymbol{x}$.

As one approach to incorporate unlabeled samples into the discriminative classifiers, minimum entropy regularizer (ER) was introduced (Grandvalet and Bengio 2005). The conditional entropy is used as a measure of class overlap. By minimizing the conditional entropy, the classifier is trained to separate unlabeled samples as well as possible. In other words, this method is based on the principle that classes should be well separated to take advantage of unlabeled samples because the asymptotic information content of unlabeled samples decreases as classes overlap.

Applying minimum ER to multinomial logistic regression, we estimate W to maximize the following conditional log-likelihood and regularizer:

$$J(W) = \sum_{n=1}^N \log P(y_n|\boldsymbol{x}_n, W) + \lambda \sum_{m=1}^M \sum_{k=1}^K P(k|\boldsymbol{x}_m, W) \log P(k|\boldsymbol{x}_m, W) + \log P(W). \quad (3)$$

Here, λ is a weighting parameter and $P(W)$ is a prior over the parameter W.

Hybrid Approach

As the mention in introduction, we propose the classifier based on the discriminative combination of a generative model and a bias correction model. In this section, we present our formulation of the classifier and the method for parameter estimation.

Generative Model & Bias Correction Model

Let $P(\boldsymbol{x}|k,\theta_k)$ be a class conditional generative model for the kth class. Here, $\Theta = \{\theta_k\}_{k=1}^K$ denotes a set of model parameters over all classes. In our formulation, each generative model is trained by using labeled set D_l. Θ is computed using MAP estimation: $\max_\Theta\{\log P(D_l|\Theta) + \log P(\Theta)\}$. Assuming Θ is independent of class probability $P(y)$, we can derive the objective function for Θ estimation as

$$J(\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log P(\boldsymbol{x}_n|k,\theta_k) + \log P(\Theta). \quad (4)$$

Here, $P(\theta_k)$ is a prior over the model parameter θ_k. z_{nk} is a class indicator variable of the nth labeled sample $(\boldsymbol{x}_n, y_n)$ ($z_{nk}=1$ if $y_n = k$, $z_{nk}=0$ otherwise).

In semi-supervised learning settings, the number of labeled samples is often small. Then, the trained generative models often have high bias. In order to obtain a better classifier with smaller bias, we newly introduce another class conditional generative model, called *bias correction model*, to decrease bias. The bias correction model belongs to the same model family as the generative model that we assume in some application, but a set of parameters Ψ of the bias correction model is different from Θ. $\Psi = \{\psi_k\}_{k=1}^K$ is obtained by using training samples based on MAP estimation:

$$J(\Psi) = \sum_{m=1}^{M}\sum_{k=1}^{K} u_{mk}\log P(\boldsymbol{x}_m|k,\psi_k) + \log P(\Psi), \quad (5)$$

where u_{mk} is a class indicator variable of the mth unlabeled sample $\boldsymbol{x}_m$. Unlike z_{nk} in Eq. (4), u_{mk} as well as ψ_k is unknown and should be estimated.

Discriminative Combination

Here, we estimate u_{mk} in a *discriminative* manner to correct bias associated with the classifier with $\hat{\Theta}$ estimated by labeled samples. More specifically, using the maximum entropy (ME) principle (Berger, Della Pietra, and Della Pietra 1996), we discriminatively combine the generative model and the bias correction model.

The ME principle is a framework for obtaining a probability distribution, which prefers the most uniform models that satisfy any given constraints. Let $R(k|\boldsymbol{x})$ be a target distribution that we wish to specify using the ME principle. A constraint is that the expected value of log-likelihood w.r.t. the target distribution $R(k|\boldsymbol{x})$ is equal to the expected value of log-likelihood w.r.t. the empirical distribution $\tilde{P}(\boldsymbol{x},k) = \frac{1}{N}\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n, k-y_n)$ of the training samples as

$$\sum_{\boldsymbol{x},k}\tilde{P}(\boldsymbol{x},k)\log P(\boldsymbol{x}|k,\hat{\theta}_k) = \sum_{\boldsymbol{x},k}\tilde{P}(\boldsymbol{x})R(k|\boldsymbol{x})\log P(\boldsymbol{x}|k,\hat{\theta}_k), \quad (6)$$

where $\tilde{P}(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n)$ is the empirical distribution of $\boldsymbol{x}$. The equation of the constraint for $\log P(\boldsymbol{x}|k,\psi_k)$ can be represented as the same form as Eq. (6). We also restrict $R(k|\boldsymbol{x})$ so that it has the same expected value for the class indicator variable $z_{k'}$ as seen in the training data, where $z_{k'}=1$ if $\boldsymbol{x}$ belongs to the k'th class, $z_{k'}=0$ otherwise, such that

$$\sum_{\boldsymbol{x},k}\tilde{P}(\boldsymbol{x},k)z_{k'} = \sum_{\boldsymbol{x},k}\tilde{P}(\boldsymbol{x})R(k|\boldsymbol{x})z_{k'}, \forall k'. \quad (7)$$

By maximizing the conditional entropy $H(R) = -\sum_{\boldsymbol{x},k}\tilde{P}(\boldsymbol{x})R(k|\boldsymbol{x})\log R(k|\boldsymbol{x})$ under these constraints, we can obtain the target distribution:

$$R(k|\boldsymbol{x},\hat{\Theta},\Psi,\Lambda) = \frac{P(\boldsymbol{x}|k,\hat{\theta}_k)^{\lambda_1}P(\boldsymbol{x}|k,\psi_k)^{\lambda_2}e^{\mu_k}}{\sum_{k'=1}^{K}P(\boldsymbol{x}|k',\hat{\theta}_{k'})^{\lambda_1}P(\boldsymbol{x}|k',\psi_{k'})^{\lambda_2}e^{\mu_{k'}}}, \quad (8)$$

where $\Lambda = \{\lambda_1,\lambda_2,\{\mu_k\}_{k=1}^K\}$ is a set of Lagrange multipliers. λ_1 and λ_2 represent the combination weights of the generative and bias correction models, and μ_k is the bias parameter for the kth class. The distribution $R(k|\boldsymbol{x},\hat{\Theta},\Psi,\Lambda)$ gives us the formulation of the discriminative classifier that consists of the trained generative model and the bias correction model.

The derived distribution given in Eq. (8) is used as estimate of u_{mk} in Eq. (5). The parameter Λ is estimated by maximizing the conditional likelihood of labeled sample set D_l. However, since D_l is used for estimating Θ and Λ, a biased estimator may be obtained. Thus, when estimating Λ, a leave-one-out cross-validation of the labeled samples is used. This cross-validation usually leads to $\lambda_2 \neq 0$. Let $\hat{\Theta}^{(-n)}$ be a generative model parameter estimated by using the labeled samples except $(\boldsymbol{x}_n, y_n)$. The objective function of Λ then becomes

$$J(\Lambda) = \sum_{n=1}^{N}\log R(y_n|\boldsymbol{x}_n,\hat{\Theta}^{(-n)},\Psi,\Lambda) + \log R(\Lambda), \quad (9)$$

where $R(\Lambda)$ is a prior over the parameters Λ. We used the Gaussian prior (Chen and Rosenfeld 1999) as $R(\Lambda) \propto \prod_j \exp\left(\frac{(\lambda_j - m_j)^2}{\sigma_j^2}\right)$. Clearly, the objective function shown in Eq. (5) (Eq. (9)) depends on Λ (Ψ) and therefore Ψ and Λ cannot be estimated independently. Thus, we alternatively and iteratively estimate them. One can see that these parameter estimations are performed both generatively and discriminatively.

Parameter Estimation Algorithm

As mentioned above, our method has three sets of parameters: Θ, Ψ, and Λ. We summarize the algorithm for estimating these model parameters in Fig. 1.

The generative model parameter Θ is estimated using only labeled samples. After the estimation of Θ, bias correction model parameter Ψ and discriminative combination

Given training sets: $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$

1. Initialize $\psi_k^{(0)}, \forall k$, $t \leftarrow 0$.
2. Compute $\hat{\Theta}$ and $\hat{\Theta}^{(-n)}$, $\forall n$ using Eq. (4).
3. Compute $\Lambda^{(0)}$ using Eq. (9) under fixed $\hat{\Theta}^{(-n)}$ and $\Psi^{(0)}$.
4. Performing below until $\frac{J(\Psi^{(t+1)}) - J(\Psi_u^{(t)})}{J(\Psi_u^{(t)})} < \epsilon$.
   Set $u_{mk}^{(t)} \leftarrow R(k|\boldsymbol{x}_m, \hat{\Theta}, \Psi^{(t)}, \Lambda^{(t)})$.
   Compute $\Psi^{(t+1)}$ using Eq. (5) under fixed $\Lambda^{(t)}$ and $\hat{\Theta}$.
   Compute $\Lambda^{(t+1)}$ using Eq. (9) under fixed $\Psi^{(t+1)}$ and $\hat{\Theta}^{(-n)}$.
   Set $t \leftarrow t + 1$.
5. Output a classifier $R(k|\boldsymbol{x}, \hat{\Theta}, \hat{\Psi}, \hat{\Lambda})$.

Figure 1: Algorithm of learning model parameters in proposed method with naive Bayes model.

parameter Λ are estimated alternatively. First, $\Lambda^{(0)}$ is estimated using trained generative model parameter $\hat{\Theta}$ and initialized bias correction model parameter $\Psi^{(0)}$. Given parameter value $\Lambda^{(t)}$ estimated by using Eq. (9) in the tth step, the algorithm calculates conditional probabilities $R(k|\boldsymbol{x}, \hat{\Theta}, \Psi^{(t)}, \Lambda^{(t)})$ for unlabeled samples. Using the conditional probabilities for $u_{mk}^{(t)}$ in Eq. (5), we obtain $\Psi^{(t+1)}$ by maximizing $J(\Psi)$ w.r.t. Ψ. Using $\hat{\Theta}$ and $\Psi^{(t+1)}$, we obtain $\Lambda^{(t+1)}$ by maximizing $J(\Lambda)$ w.r.t. Λ. These updates are iteratively and alternatively performed until some convergence criterion is met.

Experiments

Test Collections

Empirical evaluation is done on three test collections that have been often used as bench mark tests of classifiers on text classification tasks. The first is the Reuters-21578 data set (Reuters) that consists of 135 topic categories from the Reuters newswire (Yang and Liu 1999). The ten most frequent categories were usually used, and we made a subset by selected articles that belonged to one of the ten categories. For single-labeled classification tasks, we removed multi-labeled articles. Since two of the ten categories contained few articles, eight categories **acq, crude, earn, grain, interest, money-fx, ship**, and **trade** that contained many single-labeled articles were used in our evaluation. On Reuters, the articles were divided into two groups by point in time, and there were 5,732 earlier articles and 2,430 later articles in the subset. In our experiments, the later articles were used for test samples, and the earlier articles were selected as labeled or unlabeled samples. The number of vocabularies was 21,505 in the subset after removing words included in a stoplist (Salton and McGill 1983).

The second is the WebKB data set (WebKB) that contains web pages from universities. This data set consists of seven categories, and each page belongs to one of the categories. Following the setup in (Nigam, Lafferty, and McCallum 1999), only four categories **course, faculty, project**, and **student** were used. There were 4,199 pages in the categories. We removed tags, links in the pages, and words included in the stoplist. The number of vocabularies in the data set was 26,389.

The third one is the 20 newsgroups data set (20news) that consists of 20 different UseNet discussion groups. Following the setup in (Nigam, Lafferty, and McCallum 1999), only five groups **comp.*** were used for our evaluation. There were 4,881 articles in the groups. We removed words included in the stoplist and vocabularies that only one page included. The number of vocabularies in the data set was 19,357.

Experimental Settings

For a text classification task, we used a naive Bayes (NB) model as generative models $P(\boldsymbol{x}|k, \theta_k)$ and a bias correction models $P(\boldsymbol{x}|k, \psi_k)$ using independent word-based representation, known as Bag-of-Words (BOW) representation. Let $\boldsymbol{x} = (x_1, ...x_i, ...x_V)$ represent the word-frequency vector of a document, where x_i denotes the frequency of the ith word in the document and V denotes the total number of words in the vocabulary included in the text data set. In a NB model, document $\boldsymbol{x}$ in the kth class is assumed to generated from a multinomial distribution

$$P(\boldsymbol{x}|k, \theta_k) \propto \prod_{i=1}^{V} (\theta_{ki})^{x_i}. \quad (10)$$

Here, $\theta_{ki} > 0$ and $\sum_{i=1}^{V} \theta_{ki} = 1$. θ_{ki} is the probability that the ith word appears in a document belonging to the kth class. $P(\boldsymbol{x}|k, \psi_k)$ is also given a multinomial distribution as well as $P(\boldsymbol{x}|k, \theta_k)$.

As prior $P(\Theta)$ in Eq. (4), we use the following Dirichlet prior over Θ as $P(\Theta) \propto \prod_{k=1}^{K} \prod_{i=1}^{V} (\theta_{ki})^{\xi_k - 1}$. Dirichlet prior is also used for $P(\Psi)$ in Eq. (5). We tuned hyper parameters ξ_k by using leave-one-out cross-validation of labeled samples, to maximize the log likelihood of generative probabilities estimated for unseen samples with the help of EM algorithm. Since it is not an essential part of the method, we will omit the details of estimating the hyperparameters for the lack of space.

For our experiments, labeled, unlabeled, and test samples were selected randomly from each data set. We made ten different evaluation sets for each data set by random selection. 4,500 articles from earlier articles in Reuters were selected as unlabeled samples. 2,430 later articles were used as test samples. 1,000 and 2,500 web pages from WebKB were selected as test and unlabeled sets for each evaluation set. For 20news, 1,000 and 2,500 articles were selected as well as WebKB. After extracting test and unlabeled samples, labeled samples were selected from the remaining samples in each data set. Average classification accuracy over the ten evaluation sets was used to evaluate methods in each of the three data sets.

The proposed method was compared with two semi-supervised learning methods: naive Bayes with EM-λ (Nigam et al. 2000) and multinomial logistic regression with minimum entropy regularizer (MLR/MER) (cf. Grandvalet and Bengio 2005). The proposed method was also compared with two *supervised* learning methods: naive Bayes (NB) and multinomial logistic regression (MLR) classifiers (Nigam, Lafferty, and McCallum 1999). NB and MLR were only trained on labeled samples.

Table 1: Classification accuracies (%) on Reuters over various labeled data size.

Training set		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Proposed	EM-λ	MLR/MER	NB	MLR
16	0.0036	83.3 (4.5)	**86.1** (2.6)	73.1 (6.8)	70.1 (6.0)	67.4 (8.0)
32	0.0071	**89.7** (1.7)	89.4 (1.9)	82.1 (3.9)	80.1 (1.4)	80.9 (3.4)
64	0.014	**92.2** (0.7)	90.0 (1.6)	83.4 (5.0)	84.1 (1.5)	83.1 (4.6)
128	0.028	**92.8** (0.7)	91.4 (0.9)	88.5 (1.5)	88.1 (0.8)	87.8 (1.3)
256	0.057	**93.3** (0.6)	92.2 (0.8)	90.8 (0.8)	89.9 (1.3)	90.5 (0.8)
512	0.11	**94.0** (0.4)	93.1 (0.5)	93.3 (0.6)	92.4 (0.7)	93.0 (0.6)
1024	0.23	**94.6** (0.2)	93.7 (0.3)	**94.6** (0.3)	93.5 (0.5)	94.4 (0.3)

Table 2: Classification accuracies (%) on WebKB over various labeled data size.

Training set		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Proposed	EM-λ	MLR/MER	NB	MLR
8	0.0032	61.6 (6.1)	**61.9** (10.4)	52.8 (5.0)	53.5 (8.5)	52.4 (4.8)
16	0.0064	66.5 (4.5)	**68.2** (4.7)	53.2 (7.0)	59.3 (3.9)	53.2 (6.6)
32	0.013	**72.9** (3.0)	71.4 (3.0)	61.8 (5.3)	68.2 (3.1)	61.8 (5.0)
64	0.026	**76.9** (2.0)	74.3 (2.3)	69.6 (3.4)	72.7 (1.3)	69.0 (2.7)
128	0.051	**79.4** (1.6)	75.5 (2.3)	77.6 (2.2)	76.7 (1.8)	77.4 (2.1)
256	0.10	81.4 (1.6)	77.8 (1.6)	**83.1** (1.9)	79.4 (1.1)	83.0 (1.8)
512	0.20	83.2 (1.7)	79.1 (1.9)	**87.4** (1.3)	82.2 (1.2)	**87.4** (1.2)

In our experiments, for EM-λ, the value of weighting parameter λ was set by maximizing the leave-one-out cross-validation classification accuracy of the labeled samples, following the method in (Nigam et al. 2000). Note that in our experiments we selected the value from six candidate values of $\{0.01, 0.1, 0.25, 0.5, 0.75, 1\}$ to save computational time, but these candidate values were carefully selected via preliminary experiments. We used Dirichlet distribution for $P(\Theta)$, and its hyperparameter was set in a similar manner to λ.

For MLR/MER, the value of weighting parameter λ in Eq. (3) was selected from eight candidate values of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 0.5, 1\}$ that were carefully selected via the preliminary experiments. For a fair comparison of the methods, the value of λ in MLR/MER, should be also determined using training samples, for example, using leave-one-out cross-validation of labeled samples (Grandvalet and Bengio 2005). We determined the value of λ that gave the best classification performance for *test* samples to examine the *potential ability* of MLR/MER because the computation cost to tune λ was quite high. For both MLR and MLR/MER, we fixed the values of hyperparameter in Gaussian prior that gives high average classification accuracy for test samples to see the potential ability of the methods.

Results and Discussion

We evaluated classification accuracy by changing the number of labeled samples. Tables 1-3 show the average of accuracies over the ten different evaluation sets on Reuters, WebKB, and 20news. Each number in parentheses in the Tables denotes the standard deviation of the ten evaluation sets. $|D_l|$ and $|D_u|$ represent the number of labeled and unlabeled samples.

In the supervised case, as reported in (Ng and Jordan 2002), generative (discriminative) classifiers obtained better classification performance than the discriminative (generative) ones when the number of the training samples was small (large). In our experiments using NB and MLR in supervised settings on Reuters and WebKB, we obtained similar results to those in (Ng and Jordan 2002).

Table 3: Classification accuracies (%) on 20news over various labeled data size.

Training set		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Proposed	EM-λ	MLR/MER	NB	MLR
10	0.0040	**52.2** (14.1)	40.7 (10.7)	42.7 (8.3)	31.7 (5.9)	37.6 (5.4)
20	0.0080	**63.5** (5.6)	51.4 (6.7)	45.2 (5.0)	41.8 (4.9)	44.6 (4.2)
40	0.016	**68.7** (2.8)	56.7 (6.3)	52.4 (5.4)	46.8 (2.9)	51.0 (3.7)
80	0.032	**72.8** (2.3)	59.4 (4.4)	59.3 (2.6)	53.5 (3.8)	59.0 (2.3)
160	0.064	**76.0** (1.5)	65.4 (4.4)	67.6 (2.7)	60.7 (2.7)	66.6 (2.1)
320	0.13	**78.3** (1.0)	69.4 (2.3)	73.7 (1.3)	68.4 (1.6)	72.7 (1.1)
640	0.26	**81.2** (1.3)	74.4 (1.4)	79.1 (1.4)	74.9 (1.8)	77.7 (1.2)
1280	0.51	**83.6** (1.2)	78.1 (1.9)	82.2 (1.4)	77.9 (1.9)	81.1 (1.4)

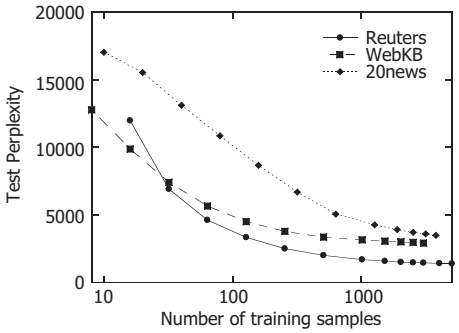

Figure 2: Test perplexities of naive Bayes models on three test collections.

However, in 20news, MLR outperformed NB even when the number of the training samples was small, which seems to be inconsistent with Ng and Jordan's report. To further investigate the result, we computed *test perplexity* $\mathcal{P}$ of the trained NB model on each test collection. $\mathcal{P}$ is a measure of how well the estimated model fits the test samples $\{\boldsymbol{x}_s, y_s\}_{s=1}^S$ not used in the training and is defined by

$$\mathcal{P} = \exp\left(-\frac{\sum_{k=1}^K \sum_{s=1}^S z_{sk} \sum_{i=1}^V x_{si} \log \hat{\theta}_{ki}}{\sum_{s=1}^S \sum_{i=1}^V x_{si}}\right), \quad (11)$$

where $\hat{\theta}_{ki}$ is an estimated parameter using the training data and z_{sk} is a class indicator ($z_{sk} = 1$ if $y_s = k$, $z_{sk} = 0$ otherwise). A smaller $\mathcal{P}$ value means better model fitness. As shown in Fig. 2, the $\mathcal{P}$ values of 20news were significantly larger than Reuters and WebKB when the number of training samples was less than 10^3. This indicates that the NB generative model did not fit the test data well when the training data size was small on 20news. In other words, if smaller $\mathcal{P}$ values were obtained for small $|D_l|$, NB would have outperformed MLR. Thus, in a supervised setting, generative classifiers can outperform discriminative ones when $|D_l|$ is small and the test perplexity of the estimated generative model is good enough. This viewpoint had better be added to conventional discussions on generative/discriminative classifiers in supervised settings.

Let us investigate semi-supervised cases. First, the classification performances of EM-λ were better (worse) than or similar to MLR for all data sets when $|D_l|$ was small (large). That is, we confirmed that the characteristics of pure generative/discriminative approaches in supervised learning also hold in the semi-supervised learning, which seems reasonable.

Second, for EM-λ vs. the proposed method, we found that if MLR outperforms NB with good test perplexity in supervised cases, then the proposed method can outperform EM-λ. This is also an expected result from supervised settings.

Finally, for MLR/MER vs. the proposed method, for all data sets the proposed method outperformed MLR/MER except when there were many labeled samples. This result comes because MLR/MER tends to be overfitting to a small number of labeled samples. In contrast, the proposed method inherently has the nature of the generative model, mitigating such an overfitting problem. When many labeled samples are available such that the overfitting problem can be solved, it would be natural that a pure discriminative approach is better than a hybrid approach.

We summarize our experimental results in terms of processing time, under the condition that the hyperparameters of all methods were determined. The supervised learning method (NB or MLR) was clearly faster than the semi-supervised counterpart (EM-λ or MLR/MER) because the former learns a model with only labeled samples, while the latter additionally uses a relatively large number of unlabeled samples. The efficiency of the proposed method and EM-λ was almost comparable because the numbers of their training parameters are very similar. Recall that in the proposed method, we can analytically calculate Θ and the size of Λ is very small. In our experiments, MLR/MER required the largest processing time. This suggests that learning with the minimum entropy regularizer may generally require a large number of training iterations due to its relatively high nonlinearity. Finally note that in EM-λ and MLR/MER, we need to adequately determine a crucial weighting parameter λ by using some resampling techniques like cross-validation, which requires a substantial amount of processing time.

Conclusions

We proposed a new method for semi-supervised classifier design based on a hybrid of generative and discriminative approaches. The main idea is to introduce bias correction model with different parameterization to correct bias associated with the generative model trained on labeled samples. In our experiments using three actual data sets for text classification problems, we compared the classification performances of the proposed method with conventional pure generative and discriminative methods. We confirmed that the proposed hybrid method could significantly outperform both generative/discriminative approaches when the performances of the pure generative/discriminative approaches were similar. In other words, we can suggest that a hybrid method is useful when the discriminative classifier obtained similar or slightly better performance than the generative classifier. Although theoretical justification for our presented hybrid approach would be necessary, we believe that this paper still contains practically important results that would be valuable to both researchers and practitioners who are interested in classifier design using labeled and unlabeled samples.

References

Amini, M. R. and Gallinari, P. 2002. Semi-supervised logistic regression. In *Proceedings of 15th European Conference on Artificial Intelligence*, 390-394.

Berger, A., Della Pietra, S., and Della Pietra, V. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, **22(1)**, 39–71.

Blum A. and Mitchell, T. 1998. Combining labeled and unlabeled data with Co-Training. In *Conference on Computational Learning Theory 11*.

Chen, S. F. and Rosenfeld, R. 1999. A Gaussian prior for smoothing maximum entropy models, Technical Report, Carnegie Mellon University.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Grandvalet, Y. and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17* (pp. 529–536). Cambridge, MA: MIT Press.

Ng, A. Y. and Jordan, M. I. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14* (pp. 841–848). Cambridge, MA: MIT Press.

Nigam, K., Lafferty, J., and McCallum, A. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information filtering*, 61–67.

Nigam, K., McCallum, A., Thrun, S., and Mitchell T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, **39**, 103–134.

Raina, R., Shen, Y., Ng, Y., and McCallum, A. 2004. Classification with hybrid generative/discriminative models. In *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Salton, G. and McGill, M.J. 1983. *Introduction to Modern Information Retrieval*. New York: NcGraw-Hill.

Seeger, M. 2001. Learning with labeled and unlabeled data, Technical Report, University of Edinburgh.

Szummer, M. and Jaakkola, T. 2001. Kernel expansions with unlabeled examples. In *Advances in Neural Information Processing Systems 13* (pp. 626–632). Cambridge, MA: MIT Press.

Tong, S. and koller, D. 2000. Restricted Bayes optimal classifiers. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*, 658–664.

Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99)*, 42–49.

Zhu, X., Ghahramani, Z., and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine learning (ICML-2003)*, 912–919.