Semi-supervised Learning of Naive Bayes Classifier with feature constraints

Nagesh Bhattu, D.V.L.N.Somayajulu
Department of Computer Science and Engineering
National Institute of Technology Warangal-506004
INDIA
nageshbhattu@gmail.com, soma@nitw.ac.in

ABSTRACT
Semi-supervised learning methods address the problem of building classifiers when labeled data is scarce. Text classification is often augmented by rich set of labeled features representing a particular class. As tuple level labling is resource consuming, semi-supervised and weakly supervised learning methods are explored recently. Compared to labeling data instances (documents), feature labeling takes much less effort and time. Posterior regularization (PR) is a framework recently proposed for incorporating bias in the form prior knowledge into posterior for the label. Our work focuses on incorporating labeled features into a naive bayes classifier in a semi-supervised setting using PR. Generative learning approaches utilize the unlabeled data more effectively compared to discriminative approaches in a semi-supervised setup. In the current study we formulate a classification method which uses the labeled features as constraints for the posterior in a semi-supervised generative learning setting. Our empirical study shows that performance gains are significant compared to an approach solely based on Generelized Expectation(GE) or limited amount of labeled data alone. We also show an application of our framework in a transfer learning setup for text classification. As we allow labeled data as well as labeled features to be used, our setup allows the presence of limited amount of labeled data on the target side of transfer learning where feature constraints are used for transferring knowledge from source domain to target domain.

KEYWORDS: Classification,Posterior Regularization.

Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology, pages 65–78,
COLING 2012, Mumbai, December 2012.

1 Introduction

Semi-supervised learning methods (O. Chapelle and Zien, 2006) address the difficulties of integration of information contained in labeled data and unlabeled data. Though labeled data is scarce, unlabeled data is abundantly available. The success of Semi-supervised methods is based on the premise of using the hidden structure of unlabeled data and aligning it with the limited amount of labeled data. Apart from unlabeled data, there are auxiliary forms of information in the form of labeled features. For example, sentiment analysis which focuses on sentiment classification is often leveraged with sentiment lexicon, which contains prior sentiment orientation of commonly occurring words. They can be used as prior knowledge in building the classifiers. Such auxiliary information has to be incorporated in the form of bias into the learning algorithm.

Though there have been efforts to build informative priors (Raina et al., 2006) or lexicon based classifiers in (Melville et al., 2009), they have been of limited success and often interact with the model in complex ways. Recently there have been research efforts from multiple perspectives addressing the same issue. Generalized Expectation Criteria (GE) is one such approach for regularizing the model based on rich set of constraints. GE allows us to specify global constraints which are allowed to be arbitrary combinations of features.

(Druck et al., 2008) used GE for building classifier solely based on labeled features. But one of the problems encountered while using GE criteria is, the model parameters and constraint parameters when exist together in a semi-supervised setup, increases the computational complexity of the algorithm. (Bellare et al., 2009) and (Druck and McCallum, 2010) addressed this issue using alternative projections approach, which is an extension of EM to discriminative learning methods. Another approach for parameter estimation which is developed simultaneously is Posterior Regularization framework (Ganchev et al., 2010) which is initially proposed for generative learning methods (Graça et al., 2007). In the initial framework in (Graça et al., 2007), constraints chosen were of limited expressibility (instance based).

1.1 Generative vs Discriminative

Generative approaches solve the inference problem modeling the joint distribution of dependent and independent variables. Discriminative methods directly model the conditional of the objective. Previous research by (Ng and Jordan, 2002) indicated that generative approaches outperform discriminative when limited amount of labeled examples are present and may under perform discriminative approaches given large amount of labeled data. Particularly in a semi-supervised setting as observed by (Nigam, 2001) where the unlabeled data is also taken while learning, generative methods shown promising results as they maximize both conditional as well model marginal together.

They proved this in the context of text classification using Multinomial Naive Bayes (MNB) approach. This view is further strengthened by recent work by (Su et al., 2011). In an another work (Druck and McCallum, 2010) has shown how discriminatively constrained generative models based on HMM, can benefit in a semi-supervised learning setup for sequence labeling task. We use PR framework for building a naive bayes classifier using feature labels as well as labeled tuples.

1.2 Transfer Learning

Transfer Learning approaches also address the issue of learning from unlabeled data using labeled data of a related domain. Such approaches are very effective in the context of widespread use of social networks which require text classification as the basic primitive. (Dai et al., 2007) has shown a method of transferring naive bayes classifier using KL-Divergence of source and target domains. A detailed survey of transfer learning approaches is addressed in (Pan and Yang, 2010). In this work we will go through an approach for transfer learning where transfer happens through feature constraints from labeled source domain to target domain. Our framework allows seamless integration of domain knowledge in the form of feature constraints, labeled data as well other domains which have abundant labeled data. Our contributions include constraining naive bayes with feature expectations and simplifying the transfer learning problem seamlessly in a semi-supervised setup.

2 Related Work

Feature prior induction into the model has been studied by (Druck et al., 2008). Work done by (Liu et al., 2004) is one of the earlier efforts for using labeled features in (classification) sentiment analysis. They use paradigm words for each class and prepare a model document for each class and compare geometric similarity of unlabeled documents with these documents for training an EM algorithm. (Melville et al., 2009) used a similar approach where he made a generative assumption of all class specific features being equally likely to appear in their respective class specific documents and all the other words being equally likely to appear in any document. They call this Lexical Classifier and pool multinomials from both lexical classifier as well as MNB from limited amount of labeled data. The prior they induce is rather less intuitive and user does not have much ease in controlling the prior. In the present method discussed in this paper, the prior can be fine-tuned and gives best results when domain expert gives exact prior knowledge.

Topic models presented by (Blei et al., 2003) using Latent Dirichlet Allocation (LDA) are used for inferring the latent topic structure hidden in the document distribution. It is totally unsupervised approach, hence LDA topic based features can be used as prior knowledge for our algorithm. Recently (Lin and He, 2009) used topic models for sentiment analysis. They further used GE expressions to bias the classifier based on sentiment lexicon. It is very effective in the context of sentiment analysis, as the presence of certain words surely effects the sentiment of entire document in one way or other. But they used GE terms along with Sentiment enhanced LDA model with regularization. In a study by (Druck et al., 2009) they have shown how the optimization problem gets complicated with the presence of GE parameters and model parameters in a semi-supervised setup. They have used Dynamic Programming to compute the covariance among the GE parameters and model parameters from labeled data. (Lin and He, 2009) don't use such approach, it is not well defined way of integrating GE terms with the generative model. In a similar context, (Mann and McCallum, 2010) have shown label regularization can be safely added in a semi-supervised setting. But it's use is limited.

Addressing the difficulties of semi-supervised learning with GE, (Bellare et al., 2009) has introduced the method of alternate projections. They use EM algorithm in a discriminative setup. They take two kinds of projections I-Projection and M-Projection which are computationally intensive. But we prefer a generative approach so as to take the benefit of document-word distribution which are even present in a unlabeled corpus. (Su et al., 2011) points out that the traditional EM formulation as given by (Nigam, 2001) reduces the conditional likelihood

of the model learnt from the labeled data. They avoid this problem in a rather efficient way which avoids the iterative procedure of EM and prove their results over large of amount of text-classification datasets. There are other methods (Sindhwani et al., 2008) which make use of graph laplacian successfully in a semi-supervised setup using Co-Clustering. These methods are computationally intensive and our method is simpler as we just use a unified constrained learning. As co-clustering methods model the higher order co-occurrences well they show performance gains at the cost of more computation. As in the current work our emphasis is on modeling a unified framework for learning from features and labels, we don't consider laplacian regularization anymore. As Naive Bayes method is generative, (Druck et al., 2007) has first described the method for building a hybrid classifier based on generative and descriminative pair, to fix the bias of generative learning. (Fujino et al., 2008) has given a method for building a hybrid classifier maximizing the joint likelihood of generative and discriminative classifiers. But the method need not converge to a stationary point as the two objectives are different. In our current work, we rather use expectation constraints over unlabeled data. So it is possible to express the optimization function as a single objective function.

(Pan and Yang, 2010) has given a detailed survey of transfer learning methods. In one of their works (Pan et al., 2010) they show a novel method of constructing a graph of domain-independent and domain-dependent features and show how spectral clustering can be used for effective domain adaptation. As part of our transfer learning application we don't consider building a graph, and hence use simple features based on mutual information. Our results are comparable to that of (Blitzer, 2008) with out spending any extra effort in dimensionality reduction. (Ganchev et al., 2010) have also addressed the problem of transfer learning using multi-view learning using agreement constraints between the views. Our approach is to develop a semi-supervised framework where we have feature transfer from related domain in the form of expectation constraints and some labeled training data. (**?**) have shown a method for transfer learning using hybrid generative/discriminative framework. It again suffers from convergence issues as two different objectives are combined.

3 Preliminaries

MNB method has been used for text classification because of it's simplicity. In a semi-supervised learning setup where learning has to be performed with limited labeled data and abundant unlabeled data Naive Bayes approach found it's application as it accounts for the marginal distribution over unlabeled data into it's objective. (Nigam, 2001) found that expectation maximization (EM) algorithm (Dempster et al., 1977) can be successfully applied in the semi-supervised context, treating missing labels of unlabeled data as latent variables of EM. We will now review the Naive Bayes approach.

3.1 Multinomial Naive Bayes

We assume $\mathscr{D}$ denote the set of documents and $\mathscr{V}$ defines the vocabulary of words. Let $\mathscr{Y}$ be set of labels. In the supervised learning setting, where $|\mathscr{D}|$ documents are given, MNB solves the following inference problem by maximum likelihood.

$$p(y|x) \propto p(x|y)p(y)$$

and $p(x|y)$ factors nicely into

$$p(x|y) = \prod_{w_i \in x} p(w_i|y)$$

The parameters of the model are computed as

$$p(w_i|y) = \frac{N_{yi} + \delta}{\sum_i N_{yi} + |\mathcal{V}|\delta}$$

where δ is used for avoiding zero probabilities (known as Lidstone smothing). If we disregard the smoothing, the above probability is simply the empirical ratio of a particular word's frequency compared to sum of all words frequencies appearing over the set of documents of the class.

$$p(w_i|y) = \frac{N_{yi}}{\sum_j N_{yj}}$$

$$N_{yi} = \sum_{t=1}^{|D|} f_{yi}^t$$

3.2 Semi-supervised learnig

Classification accuracy depends on the amount of training data available while building the classifier. In a semi-supervised learning setting we assume that in addition to the labeled training data $\mathscr{D}_{\mathscr{L}}$ we also have large amount of unlabeled data $\mathscr{D}_{\mathscr{U}}$ ($|\mathscr{D}_{\mathscr{L}}| << |\mathscr{D}_{\mathscr{U}}|$). As the joint likelihood of the labeled and unlabeled data is not in closed form, EM (Dempster et al., 1977) can be applied which results in the following iterative procedure. The algorithm starts by inferring model parameters from limited amount of labeled data and uses these parameters for inferring probabilistic labels for each of the unlabeled documents in E-step. M-step consists of inferring the parameters using these probabilistic labels for unlabeled document and labels for labeled documents.

$$Initial : \theta_L^0 = \arg\max_{\theta} \sum_{x \in \mathscr{D}_{\mathscr{L}}} log p_{\theta}(x.y)$$

$$EStep : \forall_{x \in \mathscr{D}_{\mathscr{L}} \cup \mathscr{D}_{\mathscr{U}}} computep_{\theta_t}(y|x)$$

$$MStep : \theta_{t+1} = \arg\max_{\theta} \sum_{x \in \mathscr{D}_{\mathscr{L}} U \mathscr{D}_{\mathscr{U}}} log P_{\theta_t}(x, y).$$

$$N_{yi} = \sum_{x \in \mathscr{D}_{\mathscr{U}}} f_i^x P_{\theta_t}(y|x)$$

The Expectation and Maximization steps are repeated till convergence. N_{yi} is denominator in deciding the probability of feature(word) f_i being in class y (For simplicity we have omitted

the terms from labeled data $\mathscr{D}_{\mathscr{L}}$). f_i^x gives the count of feature in document x. (Su et al., 2011) showed how conditional estimates of $P(y|x)$ from labeled data improves both accuracy and performance of naive bayes approach. θ_t gives the current estimates of the parameters for the model. The modified approach of (Su et al., 2011) computes N_{yi} is computed as follows

$$N_{yi} = P_{\theta_t^l}(y|x) \sum_{x \in \mathscr{D}_{\mathscr{U}} \cup \mathscr{D}_{\mathscr{L}}} f_i^x$$

Here θ_t^l are the parameters learnt from $\mathscr{D}_{\mathscr{L}}$.

3.3 Learning From Labeled Features

As suggested by (Druck et al., 2008), it is far easier to label features than labelling documents. They have suggested a method for learning from features using Kullback Liebler (KL) constraints (GE) on feature expectations. Their approach is based on discriminative log-linear models. The objective is given as below.

$$G(f_k) = KL(\hat{f}_k, E_U(E_{p_\theta(y/x)} f_k))$$

$$\mathscr{O} = -\sum_k G(f_k) - \sum_i \frac{\theta_i^2}{2\sigma^2}$$

Here G is GE objective function which evaluates the expectation of conditional given a document has the constraint feature. But discriminative methods can not leverage the marginal word distributions over documents, as they directly maximize the likelihood of $p(y|x)$. In one of their later studies, (Druck and McCallum, 2010) have used both discriminative and generative approaches to get the advantages of both.

4 Expectation Maximization and Posterior Constraints

In this section we review an alternate view of expectation maximization as given in (Neal and Hinton, 1993). We are given a problem of modeling a distribution of x,z where x is observed data and z is unobserved or latent(here we use z instead of y for explicitly distinguishing observed and unobserved and also to be in sync with notation widely used in literature). In the document classification task, unobserved are missing labels for unlabeled documents. Given a sample $S = x_1, ..., x_n$ observed instances of data. EM maximizes the likelihood of $p_\theta(x)$ using two block ascent steps.

$$E : q^{t+1}(z|x) = \arg\min_{q(z|x)} KL(q(z|x)||p_{\theta^t}(z|x)) = p_{\theta^t}(z|x) \tag{1}$$

$$M : \theta^{t+1} = \arg\min_{\theta} E_S \left[\sum_z q^{t+1}(z|x) log\, p_\theta(x,z) \right]. \tag{2}$$

As suggested by (Ganchev et al., 2010) this view of EM allows us to add constraints over the posterior. Instead of directly using $p_{\theta^t}(z|x)$ we can constrain the posterior to some set $\mathscr{Q}$ (set

of constrained posteriors). Their work addressed the induction of instance based constraints for word alignment. The constrained E step looks as follows.

$$E : q^{t+1}(z|x) = \arg\min_{q(z|x)\in\mathcal{Q}} KL(q(z|x)||p_{\theta^t}(z|x)) \tag{3}$$

The affine constraints used by them are of the form $E_q[f(x,z)] \leq b$. Multiple such constraints are stacked into a vector $\mathbf{E_q}[\mathbf{f(x,z)}] \leq \mathbf{b}$ and it now looks like

$$E : q^{t+1}(z|x) = \arg\min_{q} KL(q(z|x)||p_{\theta^t}(z|x)) s.t \mathbf{E_q}[\mathbf{f(x,z)}] \leq \mathbf{b}. \tag{4}$$

It is proved in (Graça et al., 2007) that local maximum of this constrained EM is indeed local maximum of regularized likelihood. Instead of these constraints we can use any convex constraints which has the effect of changing the regularization term. If we use $L2$ constraints then it is equivalent to solving the same optimization problem with L2 regularization.

4.1 Feature Expectation Constraints

(Druck et al., 2008) has first shown the utility of feature constrained learning. The constraints specify our prior belief about the presence of certain words strongly biasing the class of the document. For example the presence of word 'puck' strongly indicative of document being about hockey. If we know that 90% of documents which contain 'puck' are of class 'hockey', then we can express our constraint as 0.90N documents should have class 'hockey' (where N is the number of all the documents which contain 'puck'). We can specify the same using the following L2 based constraint.

$$\frac{1}{2\beta}\left\|\hat{f_y} - \sum_j [f(w_j, y)]\right\|_2^2 \tag{5}$$

Here $\hat{f_y}$ is feature expectation of a feature f and summation on the right runs over all the documents counting the number of documents containing feature f. It's conjugate is $-\mu'\hat{f_y} + \frac{\beta}{2}\left\|\mu\right\|_2^2$. The conjugates of various convex functions and fenchel's duality are well treated in (Dudik, 2007).

4.2 Modified EM approach

In our problem we have k L2 constraints of the above type and we use these constraints for learning from unlabeled data. The set $\mathcal{Q}$ represents the distributions constrained by these k constraints. So we have to find the auxiliary distribution $q(z|x)$ which has minimum KL divergence with $p_{\theta^t}(z|x)$ subjected to q restricted to $\mathcal{Q}$. This is similar to Maximum Entropy principle of discriminative approaches. This is called I-Projection and it is used in (Graça et al., 2007), (Druck and McCallum, 2010), (Bellare et al., 2009) in a similar setting. The dual form of complete objective is

$$\mu^{(t+1)} = \arg\max_{\mu} \mu'\hat{f} - \sum_z p_{\theta^t}(z|x) exp(\mu' f(z,x)) - \frac{\beta}{2}\left\|\mu\right\|_2^2 \tag{6}$$

Here β is a regularization constant. μ is a vector of parameters for the constraints. $\hat{f}$ is a vector of feature expectations. Here at $\mu*$ of the above objective function, the distribution $q_{\mu*}$ where $q_\mu(z|x) \propto p_{\theta'}(z|x)exp(\mu' f(z,x))$ gives the optimal constrained distribution we are looking for. The gradient of this objective is $\hat{f} - E_{q_\mu}[f(x,z)] - \beta\mu$. We solve this using L-BFGS which is a general purpose unconstrained optimization procedure. The complexity of each step of the EM algorithm is the cost of normal E step using MNB and the cost of constrained objective over the unlabeled data. If the unlabeled data is more L-BFGS might take longer time to converge to a optimal solution. Stochastic gradient descent method can be used in our objective for faster convergence. The M Step of the algorithm remains simple as we are using MNB. Computing the updated parameters of the M step uses $q^{t+1}(z|x)$ instead of $p(z|x)$.

5 Transfer Learning using Constraints

There are few efforts of transfer learning (Pan et al., 2010) (Dai et al., 2007) (Tan et al., 2009). In this section we will see how our method of constrained naive bayes is useful in building a simple scheme for transfer learning(TL). Finding the domain independent features is the key for the success of transfer learning approach. In a transfer learning setup we have labeled data only on the source side and the objective is to transfer the knowledge of inference from source task to target task(where no labeled data is available). In many TL tasks low dimensional embedding is factored out from the labeled source data and unlabeled target data. Letter this embedding is used along with source-domain classifier to do the target classification. In our framework we can figure out useful features for transfer learning and compute the feature expectations of these features in the source domain. If the features selected are informative for transfer learning, the feature expectations will behave similarly in the target domain. This allows us to transfer knowledge in a semi-supervised setup. So we can use the feature constraints learnt from a related domain to be used in target domain where only limited amount of supervision is available in the form of labeled features or labeled instances.

5.1 Method

Let $\mathscr{D}_\mathscr{S}$ and $\mathscr{D}_\mathscr{T}$ be source and target domains of our interest. $\mathscr{D}_\mathscr{S}$ consists of labeled documents (x,y) where $x \in \mathscr{X}$ document space and $y \in \mathscr{Y}$ label space. The target domain consists of documents (x) where $x \in \mathscr{X}$. Though they come from the same document space (vocabulary). The mutual information(MI) of a feature f_i with respect to the labels $\mathscr{Y}$ is

$$MI(f_i) = \sum_{y \in \mathscr{Y}} p(f_i, y) ln\left(\frac{p(f_i, y)}{(p(f_i) * p(y))}\right) \tag{7}$$

We select features of highest mutual information which occur in both domains. Though there are other methods of selecting informative features, we found this method giving better results.

6 Experiments

We evaluate our algorithm on 5 datasets which are previously used by Gregory Druck at.al (Druck et al., 2008). We further divide the dataset into binary classification problems. For datasets involving more than two classes we use multi-class classification. We use 65/35

[1]http://www.umass.edu/ mccallum/code-data.html
[2]http://www.cs.waikato.ac.nz/ml/weka/
[3]http://cs.cmu.edu/ webkb

Table 1: Datasets Description

NameOfDataset	Description	
20 newsgroups(20NG) [1]	20000 instances	20 classes
Ohscal [2]	111162 instances	10 classes
SRAA	73,218 instances	4 classes
WebKB [3]	4199 instances	4 classes
News3	9,558 instances	44 classes

training/test split through out our experiments. We took 20 news group, sraa and webkb from the same source as that of (Druck et al., 2008). Ohscal and News3 represent classification problems with large number of classes. They are taken from [2] and [3] respectively. We use information gain to simulate a human-expert providing us the labeled features.

6.1 BaseLine Approaches

We use EM-MNB as the base-line approach for semi-supervised learning. We also use GE-FL as the baseline classifier based on Druck's implementation. A semi-supervised learning algorithm leveraging labeled features and labeled data is called EM-FL (EM with Feature Labaling). The labeled features for the classification are learnt using mutual information. 1 to 16 labeled features are used for each class.

6.2 Comparison with Base Line Approaches

Together with the labeled features we also use some labeled examples. We used mallet library for our implementation. It has GE-FL and EM-MNB implemented in it. We vary the number of labeled examples among 1,2,4,8,16 per class. As the number of labeled examples increases, our algorithm's performance asymptotically approaches that of EM-MNB (unless the labeled features contain some information not expressed in the labeled tuples) . But when labeled examples are scarce, our method takes the benefit of labeled features and performs all the time better than EM-MNB by large amount. Each of the experiments are repeated for 10 runs and the average accuracy is reported. On 20 News and WebKB datasets, EM-FL fared significantly better than GE-FL. On sraa and new3 datasets, GE-FL fared better than EM-FL. The reason for this behavior is, GE-FL learns it's model based on limited number of feature constraints. The parameters of the model for features not in the constraint set estimated based on their co-occurrence with the constraint features. In our model, we have to cope with labeled features as well as labeled examples, so when the feature constraints carry more information GE-FL outperforms our approach. The two parameters of EM-FL are weight for unlabeled data and gaussian prior variance. The optimal parameters are found by using a grid search of possible values for these parameters with values between 0.1 to 1.5 with in a span of 0.1 and optimal values are used for each of our experiments. For the ohscal dataset EM algorithm fared well. This is because of the amount of unlabeled data available for learning. But EM-FL is also closer and infact fared better than EM when the number of labeled examples per class is lesser.

[4] http://mallet.cs.umass.edu

Dataset	Algo	Number of Labeled Examples per class				
		1	2	4	8	16
20 News	EM-FL	71.5(1.51)	71.59(1.53)	71.83(1.35)	71.92(1.39)	**72.53(1.53)**
	EM	14.29(5.15)	19.22(3.62)	21.02(4.38)	25.5(7.47)	29.98(6.45)
	GE-FL	62.9(1.53)				
news3	EM-FL	70.7(2.41)	70.65(2.06)	71.0(1.84)	71.08(1.63)	71.73(1.24)
	EM-FL	21.64(6.44)	29.15(6.68)	36.04(5.33)	47.75(5.38)	57.3(2.91)
	GE-FL	**75.39(2.53)**				
ohscal	EM-FL	57.19(0.84)	57.35(0.74)	57.95(0.68)	58.4(1.11)	59.22(1.06)
	EM	38.48(7.24)	46.53(5.27)	51.01(4.1)	57.34(2.26)	**59.88(2.13)**
	GE-FL	57.28(0.64)				
sraa	EM-FL	94.94(1.25)	94.98(1.23)	94.93(1.22)	94.91(1.24)	94.73(1.43)
	EM	58.02(25.38)	69.79(15.26)	79.44(12.76)	85.84(6.5)	91.43(1.52)
	GE-FL	**99.43(0.04)**				
webkb	EM-FL	86.12(0.81)	86.15(0.75)	86.09(0.8)	86.01(1.03)	**86.36(0.88)**
	EM	45.55(23.4)	42.03(16.24)	52.08(20.22)	43.16(9.07)	55.08(18.67)
	GE-FL	82.53(1.06)				

Table 2: Performance Results for varying number of labeled samples

6.2.1 Varying Unlabeled Data

Dataset	Algo	fraction of unlabeled examples used				
		0.1	0.2	0.3	0.4	0.5
20 News	EM-FL	50.01(4.85)	59.47(1.84)	**64.61(2.18)**	**67.01(1.36)**	**69.72(0.94)**
	EM	11.14(2.8)	18.55(3.67)	23.27(6.84)	27.53(2.94)	31.38(5)
	GE-FL	**56.36(1.19)**	**59.63(1.84)**	60.42(1.07)	60.94(0.52)	62.69(0.99)
SRAA	EM-FL	79.67(0.48)	82.49(0.43)	88.27(0.33)	92.05(0.71)	94.71(0.43)
	EM	58.36(2.74)	72.56(8.84)	79.44(0.34)	79.43(0.46)	79.67(0.4)
	GE-FL	**99.18(0.1)**	**99.35(0.1)**	**99.32(0.12)**	**99.37(0.04)**	**99.44(0.08)**
news3	EM-FL	41.92(1.94)	54.05(2.84)	60.98(1.13)	64.26(0.73)	67.99(1.59)
	EM	42.25(2.59)	41.06(3.21)	42.84(2.54)	45.27(2.6)	48.17(2.3)
	GE-FL	**56.1(5.52)**	**64.79(1.11)**	**68.53(1.56)**	**67.61(2.05)**	**71.02(0.81)**
ohscal	EM-FL	**62.43(2.153)**	**65.21(1.526)**	**64.98(0.957)**	**65.03(0.996)**	**62.83(2.016)**
	EM	61.54(2.14)	63.59(1.577)	63.02(1.22)	64.19(1.2)	61.48(2.039)
	GE-FL	49.284(2.91)	53.55(2.052)	54.21(1.377)	57.74(0.906)	55.5(0.568)
webkb	EM-FL	73.76(3.1)	**79(4.48)**	**81.93(1.5)**	**84.04(1.03)**	**85.17(1.49)**
	EM	73.76(8.25)	65.59(7.26)	74.95(1.28)	78.31(2.16)	74.26(3.8)
	GE-FL	**77.07(2.74)**	78.75(1.23)	81.05(1.28)	81.05(0.6)	81.05(1.5)

Table 3: Performance Results for varying fraction of unlabeled data

We have conducted a set of experiments by varying the number of unlabeled data samples from which the classifier is learnt using feature constraints. For each dataset, we vary the amount of training data from 0.1 to 0.5 in steps of 0.1, of the total amount of data. The amount of test data is fixed at 0.1 fraction of total data. The number of feature constraints are kept same as the previous experiment. We kept the number of labeled examples available for training EM-FL and EM algorithms as 512. As observed in the Table 3, EM-FL performs better than GE-FL in 3 datasets. GE-FL learns the classifier very accurately for the sraa dataset as observed before.

6.2.2 Performance

Training a GE-FL classifier from a large dataset such as news3 requires significant computation resources as it is based on global optimization requiring L-BFGS to be run on a parameter space equal to the number of features. Our approach consists of much lighter optimization where

the number of parameters for L-BFGS is much smaller, hence converges faster. (Fujino et al., 2008) have given that the number of steps for convergence of their objective function is 160, under similar conditions on webkb dataset, our objective converges in 15 steps. Training GE-FL classifier on news3 dataset took 110.78 seconds (on a system with 2GHz processor and 2GB RAM) where as training a EM-FL classifier takes 42.59 seconds.

6.3 TransferLearning

Table 4: Multi Domain Sentiment Dataset for Transfer Learning

NameOfDataset	Description	No of Classes	No of Features
Books [5]	2000 instances	2 classes	473,856
DVD	2000 instances	2 classes	
Electronics	2000 instances	2 classes	
Kitchen	2000 instances	2 classes	

We use Multi Domain Sentiment analysis dataset developed by (Blitzer, 2008). This dataset contains product reviews collected at amazon. There are 4 sets of reviews, each of which contains 2000 positive/negative documents. It is already pre-processed. We make 12 transfer learning tasks of these 4 datasets. Here B–D indicates the task is to use labeled data of Books dataset and learn a classifier for the domain DVD for which there is no labeled data. We use 100 features collected through mutual information from source domain and use them as the constraints for the target domain to learn a classifier. We just use mutual information which requires one pass through the dataset. In most cases our approach is comparable to that of (Blitzer, 2008). But we observe that our results are a notch behind that of the results obtained in (Pan et al., 2010). This is because our approach does not take the benefit of co-clustering which requires building the graph of co-occurrences. We conducted another experiment varying the

	B–D	D–B	B–E	E–B	B–K	K–B	D–E	E–D	D–K	K–D	E–K	K–E
Base	0.759	0.762	0.66	0.719	0.719	0.729	0.667	0.735	0.734	0.758	0.851	0.825
GE-FL	**0.773**	0.771	0.707	**0.767**	0.775	**0.735**	0.713	**0.781**	0.737	**0.795**	0.822	0.821
SCL	0.758	**0.797**	**0.759**	0.754	**0.789**	0.686	**0.741**	0.762	**0.814**	0.767	**0.859**	**0.868**

Table 5: Transfer Learning Results

number of labeled features. We varied them from 50-250 in steps of 50. We observed that too many features is not benefiting the classification accuracy and roughly 150-200 features are enough to improve significantly from the baseline. In all our experiments we averaged over the 10 runs of the same algorithm. The results are plotted and given in figure 1.

7 Conclusion & Future Work

Though the problem of adapting naive bayes approach with discriminative constraints has been addressed previously most of them are based on working with multi-objective optimization problem, which does not have theoretical convergence properties. In the current work, we instead used an objective function which learns both from labeled data and feature constraints

[5] http://www.cs.jhu.edu/ mdredze/datasets/sentiment/

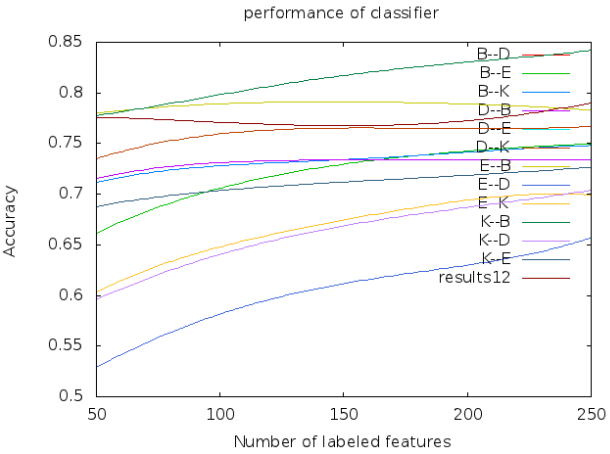

Figure 1: Performance of Feature transfer Learning

over unlabeled data and yet resulting in a single point solution. Our experimental results show how very few feature constraints (infact one cosntraint per class) can also help to improve the classifier by a significant margin over the base-line. From a computational efficiency point of view, our approach takes much lesser time compared to GE-FL as we use a combination of naive bayes and maximum entropy. It still remains an open question how to incorporate GE type of constraints in a semi-supervised setup. We have shown an application of our framework to the transfer learning problem. Empirical results show that it is competent with state of art approaches with out incurring extra computational burden.

References

Bellare, K., Druck, G., and McCallum, A. (2009). Alternating projections for learning with expectation constraints. In *UAI*, pages 43–50.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blitzer, J. (2008). *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University of Pennsylvania.

Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. (2007). Transferring naive bayes classifiers for text classification. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, AAAI'07, pages 540–545. AAAI Press.

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

Druck, G., Mann, G., and McCallum, A. (2008). Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 595–602, New York, NY, USA. ACM.

Druck, G., Mann, G. S., and McCallum, A. (2009). Semi-supervised learning of dependency parsers using generalized expectation criteria. In *ACL/AFNLP*, pages 360–368.

Druck, G. and McCallum, A. (2010). High-performance semi-supervised learning using discriminatively constrained generative models. In *ICML*, pages 319–326.

Druck, G., Pal, C., McCallum, A., and Zhu, X. (2007). Semi-supervised classification with hybrid generative/discriminative methods. In *KDD*, pages 280–289.

Dudik, M. (2007). *Maximum entropy density estimation and modeling geographic distributions of species*. PhD thesis, Princeton, NJ, USA. AAI3281302.

Fujino, A., Ueda, N., and Saito, K. (2008). Semisupervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):424–437.

Ganchev, K., Graça, J. a., Gillenwater, J., and Taskar, B. (2010). Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11:2001–2049.

Graça, J., Ganchev, K., and Taskar, B. (2007). Expectation maximization and posterior constraints. In *NIPS*.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *CIKM*, pages 375–384.

Liu, B., Li, X., Lee, W. S., and Yu, P. S. (2004). Text classification by labeling words. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI'04, pages 425–430. AAAI Press.

Mann, G. S. and McCallum, A. (2010). Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984.

Melville, P., Gryc, W., and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *KDD*, pages 1275–1284.

Neal, R. M. and Hinton, G. E. (1993). A new view of the em algorithm that justifies incremental and other variants. *Learning in Graphical Models*, pages 355–368.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 2(14):841–848.

Nigam, K. P. (2001). *Using unlabeled data to improve text classification*. PhD thesis, Pittsburgh, PA, USA. AAI3040487.

O. Chapelle, B. S. and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA,.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *WWW*, pages 751–760.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Raina, R., Ng, A. Y., and Koller, D. (2006). Constructing informative priors using transfer learning. In *ICML*, pages 713–720.

Sindhwani, V., Hu, J., and Mojsilovic, A. (2008). Regularized co-clustering with dual supervision. In *NIPS*, pages 1505–1512.

Su, J., Shirab, J. S., and Matwin, S. (2011). Large scale text classification using semisupervised multinomial naive bayes. In *ICML*, pages 97–104.

Tan, S., Cheng, X., Wang, Y., and Xu, H. (2009). Adapting naive bayes to domain adaptation for sentiment analysis. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 337–349, Berlin, Heidelberg. Springer-Verlag.