Semi-supervised Document Classification with a Mislabeling Error Model

Anastasia Krithara[1], Massih R. Amini[2],
Jean-Michel Renders[1], and Cyril Goutte[3]

[1] Xerox Research Centre Europe, chemin de Maupertuis, F-38240, Meylan, France
Anastasia.Krithara@xrce.xerox.com
[2] University Pierre et Marie Curie, 104, avenue du President Kennedy,
75016 Paris, France
amini@poleia.lip6.fr
[3] National Research Council Canada, 283, boulevard Alexandre-Taché,
Gatineau, QC J8X 3X7, Canada
Cyril.Goutte@nrc-cnrc.gc.ca

Abstract. This paper investigates a new extension of the Probabilistic Latent Semantic Analysis (PLSA) model [6] for text classification where the training set is partially labeled. The proposed approach iteratively labels the unlabeled documents and estimates the probabilities of its labeling errors. These probabilities are then taken into account in the estimation of the new model parameters before the next round. Our approach outperforms an earlier semi-supervised extension of PLSA introduced by [9] which is based on the use of *fake labels*. However, it maintains its simplicity and ability to solve multiclass problems. In addition, it gives valuable information about the most uncertain and difficult classes to label. We perform experiments over the `20Newsgroups`, `WebKB` and `Reuters` document collections and show the effectiveness of our approach over two other semi-supervised algorithms applied to these text classification problems.

1 Introduction

In this paper we present a new semi-supervised variant of the Probabilistic Latent Semantic Analysis (PLSA) algorithm [6] for text classification in which a mislabeling error model is incorporated.

Semi-supervised learning (SSL) algorithms have widely been studied since the 1990s mostly thanks to Information Access (IA) and Natural Language Processing (NLP) applications. In these applications unlabeled data are significantly easier to come by than labeled examples which generally require expert knowledge for correct and consistent annotation [3,4,13,16,11,1]. The underlying assumption of SSL algorithms is, if two points are *close* then they should be labeled similarly, resulting in that the search of a decision boundary should take place in low-density regions. This assumption does not imply that classes are formed from single compact clusters, only that objects from two distinct classes are not likely

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 370–381, 2008.

to be in the same cluster. This *cluster assumption* has first been expressed by [12] who proposed a mixture model to estimate the generation probability of examples by using both the labeled and unlabeled data. Prediction (the classification of new examples) is done by applying Bayes rule. Many practical algorithms have been implemented within this generative framework and successfully been applied to text classification [13].

Following the cluster assumption, we propose a new algorithm that iteratively computes class labels for unlabeled data and estimates the class labeling error using a mislabeling error model.

The parameters of this mislabeling error model are estimated within a semi-supervised PLSA (ssPLSA) model by maximizing the data log-likelihood, taking into account the class labels and their corresponding error estimates over the unlabeled examples. This work generalizes the study in [2], where a mislabeling error model was also proposed for SSL of discriminative models in the case of binary classification problems. We further show why the generative assumption leading to the ssPLSA we propose is more likely to hold than the one which serves to develop the semi-supervised Naive Bayes (ssNB) model. The empirical results we obtained confirm the effectiveness of our approach on `20Newsgroups`, `WebKB` and `Reuters` document collections over the ssNB [13], the transductive Support Vector Machine (SVM) [8] and a previously developed ssPLSA model [9] in which fake labels are assigned to unlabeled examples.

In the remainder of the paper, we first briefly describe in section 2.2, the ssNB model proposed by [13] for text classification. Then in section 2.3, we present our extension of the aspect PLSA model for semi-supervised learning, in which we incorporate a mislabeling error. The previously developed ssPLSA model with fake labels is presented in the same section. The experiments we conducted are described in section 3. Finally, in section 4, we discuss the outcomes of this study and we also draw some pointers for the continuation of this research.

2 Semi-supervised Generative Models for Document Classification

This section presents two probabilistic frameworks for modeling the nature of documents in the case where a partially labeled training set is available. Each framework defines a generative model for documents and encompasses different probabilistic assumptions for their generation and their labeling. The ultimate aim of each framework is to assign a label to an unseen document.

2.1 Notations

We assume that the training set is a collection of partially labeled documents $\mathcal{D} = \{d_1, \ldots, d_{N_d}\}$ containing words from the vocabulary $\mathcal{W} = \{w_1, \ldots, w_{N_w}\}$. D_l and D_u denote respectively the set of labeled and unlabeled documents in $\mathcal{D}$. All documents from D_l have a class label $y \in \mathcal{C} = \{y_1, ..., y_K\}$ and each document $d \in \mathcal{D}$ is represented by the vector of word frequencies $\boldsymbol{d} =< n(w, d) >_{w \in \mathcal{W}}$.

2.2 Naive Bayes Model

In this framework each document is assumed to be generated by a mixture model:

$$p(\boldsymbol{d}, \Theta) = \sum_{k=1}^{K} p(y_k \mid \Theta) p(\boldsymbol{d} \mid y_k, \Theta) \tag{1}$$

We further assume that there is an univocal correspondence between each class $y \in \mathcal{C}$ and each mixture component. A document d is therefore generated by first selecting a mixture component according to the prior class probabilities $p(y_k \mid \Theta)$, and then generating the document from the selected mixture component, with probability $p(\boldsymbol{d} \mid y_k, \Theta)$ (Figure 1 (a)).

The probability of a new document is the sum over all mixture components as the true class to which the document belongs to is unknown.

In the Naive Bayes model the co-occurrence of words within each document is assumed to be independent; this essentially corresponds to the *bag-of-words* assumption. From this assumption the probability of a document d given the class y_k can be expressed as

$$p(\boldsymbol{d} \mid y_k, \Theta) \propto \prod_{j=1}^{N_w} p_{jk}^{n(w_j, d)} \tag{2}$$

Where, p_{jk} is the probability of generating word w_j in class y_k. The complete set of model parameters consists of multinomial parameters for the class priors $p(y_k)$ and word generation probabilities p_{jk}:

$$\Theta = \{p(y_k) : y_k \in \mathcal{C}; p_{jk} : w_j \in \mathcal{W}, y_k \in \mathcal{C}\}$$

[13] propose to estimate Θ by maximizing the complete data log-likelihood using an Expectation-Maximization (EM) algorithm, and modulating the influence of unlabeled documents in the estimation of the log-likelihood using a weighting parameter λ. The algorithm we used in our experiments may be sketched out as follows (refer to [13] for further details). The initial set of Naive Bayes parameters $\Theta^{(0)}$ is obtained by maximizing the likelihood over the set of labeled documents $D_l \subset \mathcal{D}$. We then iteratively estimate the probability that each mixture component $y_k \in \mathcal{C}$ generates each document $d \in \mathcal{D}$ using the current parameters $\Theta^{(j)}$, and update the Naive Bayes parameters $\Theta^{(j+1)}$ by maximizing the complete-data log-likelihood in which the effect of unlabeled documents are moderated via a parameter $\lambda \in [0, 1]$. The complexity of this algorithm is $O(K \times M)$, where $M = \#\{(w, d) | n(w, d) > 0\}$.

2.3 Probabilistic Latent Semantic Analysis

The PLSA model introduced by Hoffmann [6] is a probabilistic model which characterizes each word in a document as a sample from a mixture model, where mixture components are conditionally-independent multinomial distributions.

This model, also known as the aspect model [14], associates an unobserved latent variable (called aspect, topic or component) $\alpha \in A = \{\alpha_1, ..., \alpha_L\}$ to each observation corresponding to the occurrence of a word $w \in \mathcal{W}$ within a document $d \in \mathcal{D}$. One component or topic can coincide with one class or, in another setting, a class can be associated to more than one component. Although originally proposed in an unsupervised setting, this latent variable model is easily extended to classification with the following underlying generation process:

- Pick a document d with probability $p(d)$,
- Choose a latent variable α according to its conditional probability $p(\alpha \mid d)$
- Generate a word w with probability $p(w \mid \alpha)$
- Generate the document class y according to the probability $p(y \mid \alpha)$

The final result of this generation process is the document class $y \in \mathcal{C}$ as well as words $w \in \mathcal{W}$ within it, while the latent variable α is discarded. Figure 1 depicts the generation processes for the aspect models and the Naive Bayes model introduced earlier.

The generation of a word w within a document d can then be translated by the following joint probability model:

$$P(w, d) = p(d) \sum_{\alpha \in A} p(w \mid \alpha) P(\alpha \mid d) \tag{3}$$

for unlabeled data and, for labeled data:

$$P(w, d, y) = p(d) \sum_{\alpha \in A} p(w \mid \alpha) P(\alpha \mid d) P(y|\alpha) \tag{4}$$

This model overcomes some simplifying assumptions of Naive Bayes in two important ways. First, it relaxes the assumption that a class y is associated to a single topic. In PLSA, the number of topics $|A|$ may be larger than the number of classes K. The second and crucial difference is that in Naive Bayes, all words must be generated from the same topic (eq. 2). This requires the use of clever smoothing strategies to counter the fact that some words that are unrelated to a topic may appear by coincidence in a document from that topic. On the other hand, in PLSA, a topic is drawn independently from $p(\alpha \mid d)$ each time a new word is generated in a document. This provides a much more natural way to handle unusual words or multi-topicality.

Semi-supervised PLSA with Fake Labels. As the aspect PLSA model characterizes the generation of the co-occurrence between a word w and a document d, for learning the semi-supervised models we have to form two other labeled $\mathcal{Z}_l$ and unlabeled $\mathcal{X}_u$ training sets from D_l and D_u. We consider now each observation as a pair $x = (w, d)$ such that observations in $\mathcal{Z}_l$ are assigned to the same class label than the document d they contain.

We recall that we still characterize the data using a mixture model with L latent topic variables α, under the graphical assumption of aspect models (that

d and w are independent conditionally to a latent topic variable α). In this case the model parameters are

$$\Lambda = \{p(\alpha \mid d), p(y \mid \alpha), p(w \mid \alpha), p(d) : \alpha \in A, d \in \mathcal{D}, w \in \mathcal{W}\}$$

Krithara et al. [9], introduced a semi-supervised variant of PLSA, following the work of [5], where additional *fake* labels were introduced for the unlabeled data. The motivation for the latter was to try to solve the problem of the unlabeled components (components which contain only unlabeled examples). The lack of labeled examples in these components can lead to arbitrary class probabilities, and as a result, to arbitrary classification decisions. So all labeled examples in $\mathcal{Z}_l$ are kept with their real class labels and all unlabeled examples in $\mathcal{X}_u$ are assigned a new *fake* label $y = 0$.

The model parameters Λ are obtained by maximizing the complete data log-likelihood,

$$\mathcal{L}_1 = \sum_{x \in \mathcal{Z}_l \cup \mathcal{X}_u} \log p(x, y) = \sum_{x \in \mathcal{Z}_l \cup \mathcal{X}_u} \log p(w, d, y) \tag{5}$$

using the Expectation-Maximization algorithm. [9] showed how the EM iterations could be implemented via a single multiplicative update.

Once the model parameters are obtained, each new document d_{new} must first be "folded in" the model, by maximizing the likelihood on the new document using EM, in order to obtain the posteriors $P(\alpha|d_{new})$. We then need to distribute the probability associated with the *fake* label $y = 0$, on the "true" labels:

$$\forall y \neq 0,\ P(y|d_{new}) \propto \sum_{\alpha} P(\alpha|d_{new})P(y|\alpha) + \mu \sum_{\alpha} P(\alpha|x)P(y{=}0|\alpha) \tag{6}$$

with $\mu << 1$. This model corresponds to the graphical model in figure 1(b). A new document d is then assigned the class with maximum posterior probability. The complexity of this algorithm is $O(2 \times |A| \times M)$ where, as before, $M = \#\{(w, d)|n(w, d) > 0\}$.

Semi-supervised PLSA with a Mislabeling Error Model. In this section we present a new version of a semi-supervised PLSA model in which a misclassification error is incorporated. We assume that the labeling errors made by the generative model for unlabeled data come from a stochastic process and that these errors are inherent to semi-supervised learning algorithms. The idea here is to characterize this stochastic process in order to reduce the labeling errors computed by the classifier for unlabeled documents in the training set.

We assume that for each unlabeled example $d \in D_u$, there exists a perfect, true label y, and an imperfect label $\tilde{y}$, estimated by the classifier. We model the stochastic nature of the labeling by the following probabilities:

$$\forall (k, h) \in \mathcal{C} \times \mathcal{C}, \beta_{kh} = p(\tilde{y} = k|y = h) \tag{7}$$

with the constraint that $\forall h, \sum_k \beta_{kh} = 1$.

In this case, the new extension of the aspect model to unlabeled documents can be expressed by the graphical model represented in figure 1(c).

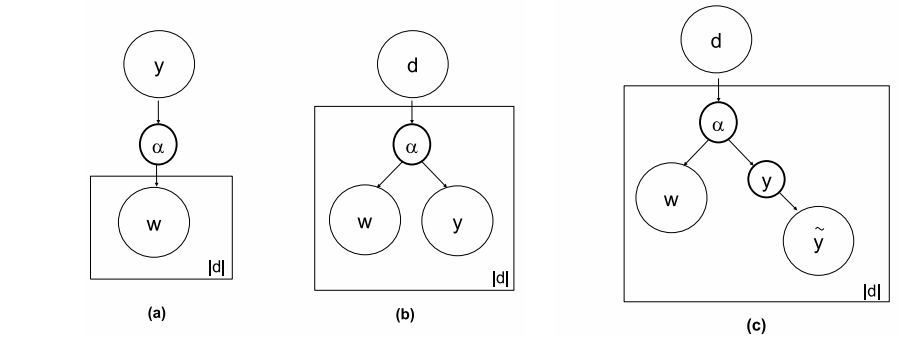

Fig. 1. Graphical model representation of the Naive Bayes model (a), PLSA/aspect models for labeled (b) and unlabeled (c) documents. The "plates" indicate repeated sampling of the enclosed variables.

The underlying generation process associated to this second latent variable model for unlabeled documents is:

- Pick a document d with probability $p(d)$,
- Choose a latent variable α according to its conditional probability $p(\alpha \mid d)$
- Generate a word w with probability $p(w \mid \alpha)$
- Generate the *latent* document class y according to the probability $p(y \mid \alpha)$
- The imperfect class label $\tilde{y}$ is generated with probability $\beta_{\tilde{y}|y} = p(\tilde{y} \mid y)$

With this new graphical model, the joint probability between an unlabeled example $x \in \mathcal{X}_u$ and its imperfect class label estimated by the classifier can be expressed as

$$\forall x \in \mathcal{X}_u, p(w, d, \tilde{y}) = p(d) \sum_{\alpha \in A} p(w|\alpha) p(\alpha|d) \sum_{y \in \mathcal{C}} \beta_{\tilde{y}|y} p(y|\alpha)$$

With this formulation it becomes apparent that for each unlabeled document, the imperfect class probabilities estimated by the classifier is weighted over all possible true classes (i.e. $p(\tilde{y} \mid \alpha) = \sum_y p(\tilde{y} \mid y) p(y|\alpha)$). This lessens the possibility that the classifier makes a mistake over the document class as it aggregates the estimates over all true classes.

The model parameters

$$\Phi = \{p(\alpha \mid d), p(w \mid \alpha), p(d), \beta_{\tilde{y}|y} : d \in \mathcal{D}, w \in \mathcal{W}, \alpha \in A, y \in \mathcal{C}, \tilde{y} \in \mathcal{C}\}$$

are estimated by maximizing the log-likelihood

$$\begin{aligned} \mathcal{L}_2 = & \sum_{d \in D_l} \sum_w n(w,d) \log \sum_\alpha p(d) p(w|\alpha) p(\alpha|d) p(y|\alpha) \\ & + \sum_{d \in D_u} \sum_w n(w,d) \log \sum_\alpha p(d) p(w|\alpha) p(\alpha|d) \sum_y p(\tilde{y}|y) p(y|\alpha) \end{aligned} \tag{8}$$

Algorithm 1. Semi-Supervised PLSA with mislabeling error model

Input :
- A set of partially labeled documents $\mathcal{D} = D_l \cup D_u$,
- Training sets $\mathcal{Z}_l$ and $\mathcal{X}_u$ formed from D_l and D_u,
- Random initial model parameters $\Phi^{(0)}$.
- $j \leftarrow 0$

repeat
- Re-estimate model parameters using multiplicative update rules (9–11)
- $j \leftarrow j + 1$

until *convergence of* $\mathcal{L}_2$ *(eq. 8)* ;

Output : A generative classifier with parameters $\Phi^{(j)}$

using an EM-type algorithm. Joining the E and M steps in a single multiplicative update, we get:

$$p^{(j+1)}(w|\alpha) = p^{(j)}(w|\alpha) \times \left[\sum_{d \in D_l} n(w,d) \frac{p^{(j)}(\alpha|d) p(y|\alpha)}{p^{(j)}(w,y|d)} + \sum_{d \in D_u} n(w,d) \frac{p^{(j)}(\alpha|d) \sum_y p(y|\alpha) \beta^{(j)}_{\tilde{y}|y}}{p^{(j)}(w,\tilde{y}|d)} \right] \tag{9}$$

$$p^{(j+1)}(\alpha|d) = p^{(j)}(\alpha|d) \sum_w n(w,d) p^{(j)}(w|\alpha) \times \begin{cases} \frac{p(y|\alpha)}{p^{(j)}(w,y|d)}, \forall d \in D_l \\ \frac{\sum_y p(y|\alpha)\beta^{(j)}_{\tilde{y}|y}}{p^{(j)}(w,\tilde{y}|d)}, \forall d \in D_u \end{cases} \tag{10}$$

$$\beta^{(j+1)}_{\tilde{y}|y} = \beta^{(j)}_{\tilde{y}|y} \sum_w \sum_{d \in D_u} n(w,d) \frac{p^{(j)}(w,y|d)}{p^{(j)}(w,\tilde{y}|d)} \tag{11}$$

where $p^{(j)}(w,y|d) = \sum_\alpha p^{(j)}(\alpha|d) p^{(j)}(w|\alpha) p(y|\alpha)$, and
$p^{(j)}(w,\tilde{y}|d) = \sum_\alpha p^{(j)}(\alpha|d) p^{(j)}(w|\alpha) \sum_y p(y|\alpha) \beta^{(j)}_{\tilde{y}|y}$.

Note that the mislabeling probabilities are estimated over the unlabeled set.

In this new version of the semi-supervised PLSA algorithm, $P(y|\alpha)$ is fixed, and its values depend on the value of latent topic variable α. The overall number of topics, $|A|$, is given, and in addition, the number of latent topics α per class is also known. During initialization, we set $P(y|\alpha) = 0$ for all latent topic variables α which do not belong to the particular class y. This algorithm (Algorithm 1, above) is also an EM-like algorithm, and the iterative use of equations 9, 10 and 11 corresponds to alternating the `E-step` and `M-step`. Convergence is therefore guaranteed to a local maximum of the likelihood.

The complexity of this algorithm is $O(|A| \times M \times K)$, which is comparable with the previous algorithms, as the number of latent variables $|A|$ is generally set to a relatively low value.

3 Experiments

In our experiments we used two collections from the CMU World Wide Knowledge Base project, `WebKB` and `20Newsgroups`[1], and the widely used text classification collection `Reuters` − 21578. For each dataset, we ran 4 algorithms: the two flavours of semi-supervised PLSA presented above (with mislabeling error model and with fake labels), as well as the semi-supervised Naive Bayes and the transductive Support Vector Machine (TSVM) algorithm [7]. For the latter, we performed a one class vs. all TSVM for all existing classes using the SVM-light package of Joachims [7]. We used the linear kernel and we have optimized, for each of the different ratio of labeled-unlabeled documents in the training set, the cost parameter C by cross-validation. All performance reported below were averaged over 10 randomly chosen labeled, unlabeled and test sets.

The `20Newsgroups` dataset is a commonly used document classification collection. It contains 20000 messages collected from 20 different Usenet newsgroups. The `WebKB` dataset contains web pages gathered from 4 different university computer science departments. The pages are divided into seven categories. In this paper, we focus on the four most often used categories: student, faculty, course and project, all together containing 4196 pages. Finally, the `Reuters` dataset consists of 21578 articles and 90 topic categories from the Reuters newswire. We selected the documents which belong only to one class, and in addition we only kept the classes which contain at least 100 documents. This gave us a base of 4381 documents belonging to 7 different classes.

All datasets were pre-processed by removing the email tags and other numeric terms, discarding the tokens which appear in less than 5 documents, and by removing a total of 608 stopwords from the `CACM` stoplist[2]. We used the microaverage F-score measure to compare the effectiveness of the semi-supervised algorithms. To this end, for each generative classifier, $\mathcal{G}_f$, we first compute its microaverage precision P and recall R by summing over all the individual decisions it made on the test set:

$$R(\mathcal{G}_f) = \frac{\sum_{k=1}^{K} \theta(k, \mathcal{G}_f)}{\sum_{k=1}^{K} (\theta(k, \mathcal{G}_f) + \psi(k, \mathcal{G}_f))}$$

$$P(\mathcal{G}_f) = \frac{\sum_{k=1}^{K} \theta(k, \mathcal{G}_f)}{\sum_{k=1}^{K} (\theta(k, \mathcal{G}_f) + \phi(k, \mathcal{G}_f))}$$

Where, $\theta(k, \mathcal{G}_f)$, $\phi(k, \mathcal{G}_f)$ and $\psi(k, \mathcal{G}_f)$ respectively denote the true positive, false positive and false negative documents in class k found by $\mathcal{G}_f$. The F-score measure is then defined as [10]:

$$F(\mathcal{G}_f) = \frac{2P(\mathcal{G}_f)R(\mathcal{G}_f)}{P(\mathcal{G}_f) + R(\mathcal{G}_f)}$$

[1] http://www.cs.cmu.edu/∼webkb/

[2] http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/

3.1 Results

We first compare the systems in a fully supervised way, that is where 100% of the documents in the training set have their true labels and are used for training the classifiers. As there are no unlabeled training document to consider here there are no fakes or mislabeling errors to characterize, both semi-supervised PLSA models behave identically. This comparison hence gives an upper bound on the performance of each generative approach and also provides a first comparison between these frameworks. We also compared our results with the TSVM model using the SVM-light package [8]. The number of latent class variables we used in the PLSA model, $|A|$, was found by cross-validation on each data set. Table 1 sums up these results. As we can notice, in all 3 datasets, the performance of PLSA is slightly better than the Naive Bayes and SVM classifiers. These results corroborate with the intuition that the generative hypothesis, which leads to the construction of the PLSA model, is more efficient than the Naive Bayes document generation assumption (section 2.3).

Table 1. Comparison of the F-score measures between the Naive Bayes and PLSA generative models as well as the SVM classifier on `20Newsgroups`, `WebKB` and `Reuters` test sets. All classifiers are trained in a fully supervised way.

	`20Newsgroups`	`WebKB`	`Reuters`
System	F-score (%)	F-score (%)	F-score (%)
Naive Bayes	88.23	84.32	93.89
PLSA	$\|A\| = 40$	$\|A\| = 16$	$\|A\| = 14$
	89.72	**85.54**	**94.29**
SVM	88.98	85.15	89.50

Figures 2 and 3 (left) show the F-score measured over the test sets on all three data collections for semi-supervised learning at different ratio of labeled-unlabeled documents in the training set. 5% in the x-axis means that 5% of the training documents were labeled ($|D_l|$), the remaining 95% being used as unlabeled training documents ($|D_u|$). The ssPLSA with mislabeling consistently outperforms the three other models on these datasets. With only 5% of labeled documents in the training set, the F-score of the ssPLSA with mislabeling algorithm is about 15% higher than that of the ssPLSA with fake labels, on the `Reuters` dataset. Labeling only 10% of the documents allows to reach 93% F-score on `Reuters` while the 90% remaining labeled documents allows to reach the maximum performance level. The semi-supervised Naive Bayes model outperforms in the other hand the ssPLSA with fake labels on both datasets. This might be due to the fact that fake label parameterization makes it inappropriate to apply PLSA over both labeled and unlabeled documents.

The bad results of the TSVM in these experiments can be explained by the fact that the model was initially designed for 2-class classification problems and the one vs. all strategy does not give adequate recognition of classes.

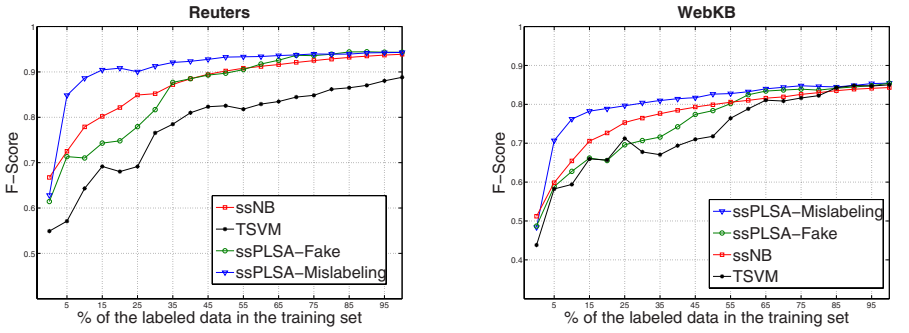

Fig. 2. F-Score (y-axis) vs. percentage of labeled training examples (x-axis), for the four algorithms on `Reuters` (left, $|A| = 14$) and `WebKB` (right, $|A| = 16$)

In order to evaluate empirically the effect of unlabeled documents for training the models we have also trained the PLSA model in a supervised manner using only the percentage of labeled documents in the training set. Figure 3 (right) shows these results on `20Newsgroups`. We can see that semi-supervised algorithms are able to take advantage from unlabeled data. For example, with 5% labeled data (corresponding to approximately 800 labeled documents with 40 documents per class), the fully supervised PLSA reaches 52.5% F-score accuracy while semi-supervised Naive Bayes and ssPLSA with fake labels achieve 63% and ssPLSA with mislabeling achieves 72%. This represents a 32% gain in F-score for the two former models.

Table 2. F-score for varying proportions of labeled-unlabeled training data, for semi-supervised Naive Bayes (ssNB), TSVM as well as semi-supervised PLSA with either the fake label (ssPLSA-f) or the mislabeling error model (ssPLSA-mem), and different numbers of latent topics $|A|$. Bold indicates statistically significantly better results, measured using a t-test at the 5% significance level.

		`20Newsgroups`				
		1%	5%	20%	40%	80%
$\|A\| = 20$	ssNB	51.45 ± 3.45	66.45 ± 0.67	75.65 ± 0.91	83.46 ± 0.46	87.98 ± 0.82
	ssPLSA-mem	53.69 ± 5.49	75.520 ± 0.22	81.59 ± 0.6	84.54 ± 0.3	87.76 ± 1.115
	ssPLSA-f	54.67 ± 4.11	75.48 ± 0.75	80.45 ± 1.09	78.86 ± 0.39	84.11 ± 0.93
$\|A\| = 40$	ssPLSA-mem	**53.52 ± 6.46**	**77.18 ± 0.66**	**82.89 ± 0.73**	**85.9 ± 0.85**	**89.04 ± 0.75**
	ssPLSA-f	**54.04 ± 6.98**	64.65 ± 3.54	67.61 ± 1.69	79.59 ± 0.28	**88.96 ± 0.64**
	TSVM	50.64 ± 1.79	54.37 ± 0.55	65.21 ± 0.75	71.31 ± 0.85	82.37 ± 1.03

One interesting aspect of our experimental results is that the behavior of the two ssPLSA variants is very different when the number of latent variables per class increases (Table 2).

For the *fake* label approach, the performance tends to decrease when more components are added to the model, and the variability of the results increases. Overall, this approach yields consistently lower performance than the "Mislabeling"

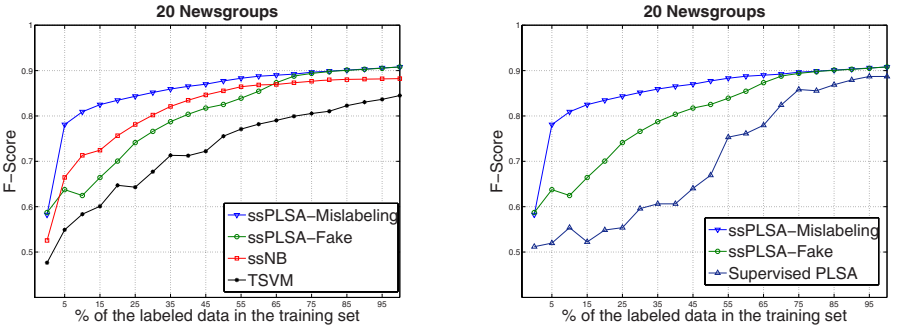

Fig. 3. Comparison of the ssPLSA models with the fully supervised PLSA (right) and with the other algorithms (left) for the `20Newsgroups` dataset ($|A| = 40$)

approach, which in addition seems less sensitive to varying numbers of components. Notice in Table 2 how, when the number of components per class is increased from 1 to 2 - corresponding respectively to $|A| = 20$ and $|A| = 40$ (`20Newsgroups`), the performance of the mislabeling approach increases slightly, but consistently. In addition, the variability of the results is mostly well contained and generally smaller than for the "fake label" approach. The results are similar for the other two datasets.

4 Conclusion

We have presented a new version of the semi-supervised PLSA algorithm, where a mislabeling error model is incorporated in the generative aspect model. Our model has been compared to two state-of-the-art semi-supervised Naive Bayes and TSVM models as well as a previously designed ssPLSA algorithm. Performances on the `20Newsgroups`, `WebKB` and `Reuters` datasets have shown promising results indicating decreases in the number of labeled documents used for training needed to achieve good accuracy, if an unlabeled document set is available. One of the advantages of our model is that it can be used directly to perform multiclass classification tasks, and as a result it is easily applicable to real world problems. A next step would be to try to combine the two presented variants of PLSA, that is the 'fake' label and the mislabeling error models. This combination would benefit from the advantages of each of the two versions and would hopefully improve the performance of our classifier. However, further experimental observations would be required to fully understand the behavior and the performance of these models.

Acknowledgments

The authors thank Nicolas Usunier for helpful comments. This work was supported in part by the IST Programme of the European Community, under the

PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

1. Amini, M.R., Gallinari, P.: The use of unlabeled data to improve supervised learning for text summarization. In: SIGIR, pp. 105–112 (2002)
2. Amini, M.R., Gallinari, P.: Semi-supervised learning with an explicit label-error model for misclassified data. In: Proceedings of the 18th IJCAI, pp. 555–560 (2003)
3. Blum, A., Mitchell, T.M.: Combining labeled and unlabeled data with co-training. In: COLT 1998, pp. 92–100 (1998)
4. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: EMNLP/VLC (1999)
5. Gaussier, E., Goutte, C.: Learning from partially labelled data - with confidence. In: Learning from Partially Classified Training Data - Proceedings of the ICML 2005 workshop (2005)
6. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd ACM SIGIR, pp. 50–57 (1999)
7. Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of ICML 1999, 16th International Conference on Machine Learning, pp. 200–209. Morgan Kaufmann Publishers, San Francisco (1999)
8. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the European Conference on Machine Learning (1998)
9. Krithara, A., Goutte, C., Renders, J.M., Amini, M.R.: Reducing the annotation burden in text classification. In: Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies (InSciT 2006), Merida, Spain (October 2006)
10. Lewis, D., Ringuette, M.: A comparison of two learning algorithms for text categorization. In: Proceedings of SDAIR 1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, pp. 81–93 (1994)
11. McLernon, B., Kushmerick, N.: Transductive pattern learning for information extraction. In: Proc. Workshop Adaptive Text Extraction and Mining (2006), Conf. European Association for Computational Linguistics
12. Miller, D.J., Uyar, H.S.: A mixture of experts classifier with learning based on both labelled and unlabeled data. In: Proc. of NIPS-(1997)
13. Nigam, K., McCallum, K.A., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39(2/3), 103–134 (2000)
14. Saul, L., Pereira, F.: Aggregate and mixed-order Markov models for statistical language processing. In: Proc of 2nd ICEMNLP (1997)
15. Si, L., Callan, J.: A semi-supervised learning method to merge search engine results. ACM Transactions on Information Systems 24(4), 457–491 (2003)
16. Slonim, N., Friedman, N., Tishby, N.: Usupervised Document Classification Using Sequentiel Information Maximization. In: SIGIR, pp. 129–136 (2002)
17. Zhang, T.: The value of unlabeled data for classification problems. In: ICML (2000)