Semi-Supervised Learning with Explicit Misclassification Modeling

Massih-Reza Amini, Patrick Gallinari
University of Pierre and Marie Curie
Computer Science Laboratory of Paris 6
8, rue du capitaine Scott, F-75015 Paris, France
{amini, gallinari}@poleia.lip6.fr

Abstract

This paper investigates a new approach for training discriminant classifiers when only a small set of labeled data is available together with a large set of unlabeled data. This algorithm optimizes the classification maximum likelihood of a set of labeled-unlabeled data, using a variant form of the Classification Expectation Maximization (CEM) algorithm. Its originality is that it makes use of both unlabeled data and of a probabilistic misclassification model for these data. The parameters of the label-error model are learned together with the classifier parameters. We demonstrate the effectiveness of the approach on four data-sets and show the advantages of this method over a previously developed semi-supervised algorithm which does not consider imperfections in the labeling process.

1 Introduction

In many real-life applications, labeling training data for learning is costly, sometimes not realistic and often prone to error. For example, for many rapidly evolving data bases available via the web, there is not enough time to label data for different information needs. In some cases, like medical diagnosis or biological data analysis, labeling data may require very expensive tests so that only small labeled data sets may be available. In other cases, like object identification in images, noise is inherent in the labeling process.

The statistician and pattern recognition communities were the first to consider the problem of forming discriminant rules using either partially classified or random misclassified training data in order to cope with this type of situation.

More recently this idea has motivated the interest of the machine learning community and many papers now deal with this subject.The use of partially classified data for training, known as semi-supervised learning, has been the subject of intense studies since 1998 and more recently there has been a resurgence of interest for training with misclassified data also called learning in presence of label noise.

We consider here semi-supervised learning for classification. Most approaches to this problem make use of a mixture density model where mixture components are identified as classes. Labeled data are known to belong to exactly one mixture component whereas unlabeled data may belong to any components. Using the Expectation Maximization (EM) algorithm [Dempster et al., 1977], proposed approaches usually attempt to optimize the likelihood of the whole labeled-unlabeled data. Starting from an initial labeling, these approaches proceed by computing at each `E-step`, tentative labels for unlabeled data using the current parameters of the model and update in the `M-step`, these parameters using the estimated labels. Our departure point here is the work of [Amini and Gallinari-*a*, 2002] who proposed a semi-supervised discriminant algorithm using a variant form of the CEM algorithm [Celeux and Govaert, 1992]. Discriminative approaches which attempt to estimate directly posterior class probabilities are often considered superior to generative models which compute these posteriors after learning class conditional densities. Tests on different datasets in [Amini and Gallinari-*b*, 2002] led to the same conclusion for semi-supervised learning. Like for other methods, at each step of the algorithm, the model computes tentative labels for unlabeled data. We extend here the system by incorporating a model which takes into account label errors. This provides an unifying framework for semi-supervised learning and learning with label noise. To our knowledge, this form of model has not been studied yet. We detail the algorithm for the case of a logistic classifier and give a convergence proof for the general case. We then show experimentally that modeling the stochastic labeling noise, increases notably the performance, especially when only small labeled datasets are available. This paper is organized as follows; we first make a brief review of work on semi-supervised learning and learning in the presence of label noise (section 2). In section 3 we present the formal framework of our model and describe in section 4 the semi-supervised approach we propose. Finally we present a series of experiments on four data sets.

2 Related work

2.1 Learning with both labeled and unlabeled data

The idea of using partially labeled data for learning started in the statistician community at the end of 60's. The seminal paper by [Day, 1969] presents an iterative EM-like approach for learning the parameters of a mixture density under the assumption of multivariate normal components with a common covariance matrix for two group-conditional distributions.

Other iterative algorithms for building maximum likelihood classifiers from labeled and unlabeled data based on the same type of assumption followed [O'Neill, 1978] [McLachlan and Ganesalingam, 1982]. Some other authors have suggested updating procedures for no-normal group conditional densities using for example kernel methods for modeling mixture components [Murray and Titterington, 1978]. There has been considerably fewer work on discriminative approaches. In his fundamental paper on logistic regression, Anderson suggests modifying a logistic regression classifier to incorporate unlabeled data in order to maximize the likelihood function [Anderson, 1979].

The semi-supervised paradigm has been recently rediscovered by the machine learning community. Most papers propose mixture density models for combining labeled and unlabeled data for classification. [Miller and Uyar, 1996] consider a mixture density model where each class is described by several component densities. [Roth and Steinhage, 1999] propose kernel discriminant analysis as an extension to classical linear discriminant analysis. This framework can be used also for semi-supervised learning. [Nigam et al., 2000] propose a semi-supervised EM algorithm which is essentially similar to the one in [McLachlan, 1992] but makes use of naive Bayes estimator for modeling the different densities. They present empirical evaluation for text classification tasks. Some authors make use of discriminant classifiers instead of modeling conditional densities. For example [Joachims, 1999] propose a transductive support vector machine which finds parameters for a linear separator using both the labeled data in the training set and the current test data whose class is unknown. [Blum and Mitchell, 1998] introduce the co-training paradigm where each sample x is supposed to be described by two modalities. Two classifiers are then used, one for each modality, operating alternatively as teacher and student. This framework can be used for unsupervised and semi-supervised learning. Based on the multi-modal framework introduced by [Blum and Mitchell, 1998], [Muslea et al., 2002] propose to combine both active and semi-supervised learning. Since that, different authors have proposed semi-supervised learning schemes, they usually follow one of the above ideas.

2.2 Learning with imperfectly labeled data

Practical applications of pattern recognition, like eg image classification problems, have motivated in the early 80's some work on the problem of learning in presence of mislabeled data for fully supervised learning. [Chittineni, 1980] obtained error bounds on the performance of the Bayes and nearest neighbor classifiers with imperfect labels. [Krishnan, 1988] considered a 2-class classification problem for two group multivariate normal mixture when training samples are subject to random misclassification and derived the likelihood estimation of parameters. [Titterington, 1989] proposed a logistic-normal distribution model and worked out an EM algorithm for the estimation of its parameters. More recently, [Lawrence and Schölkopf, 2001] proposed an algorithm for constructing a kernel Fisher discriminant from training examples in the presence of label noise.

3 A semi-supervised probabilistic model for mislabeling

We consider here the problem of semi-supervised learning. We start from the Logistic-CEM algorithm described in [Amini and Gallinari-*a*, 2002]. It is a generic scheme in the sense that it can be used with any discriminant classifier estimating the posterior class probabilities. The classifier is first trained on labeled data, it then alternates two steps until convergence to a local maximum of the Classification Maximum Likelihood (CML) criterion [Symons, 1981]. Unlabeled data are first labeled using the output of the current classifier. Classifier parameters are then learned by maximizing the CML function computed using the known labels of labeled data and the current estimated labels for unlabeled data. In this algorithm, at each iteration, labels computed for unlabeled data are considered as desired outputs.

We suppose here that labels from the labeled dataset are correct and that at each step of this algorithm, labels computed for unlabeled data are subject to error. We propose to model the imperfection of these labels using a probabilistic formalism and to learn the semi-supervised classifier by taking into account its labeling errors according to this error model. Parameters of the error model and of the classifier will be learned simultaneously in this new algorithm. Compared to the baseline algorithm [Amini and Gallinari-*a*, 2002], we explicitly take into account the fact that the current classifier is not optimally trained at each step, since many data labels are missing and are only estimated.

In the following we present the general framework and the learning criterion of our model.

3.1 General framework

We suppose that each example belongs to one and only one class and that there are available a set of n labeled examples, D_l and a set of m unlabeled examples, D_u. A discriminant classifier is to be trained on the basis of these $n+m$, d-dimensional feature vectors[1] $x \in \mathbb{R}^d$. For each labeled example x_i in D_l, let c_i and $t_i = \{t_{ki}\}_k$ be respectively the class label and the indicator vector class associated to x_i :

$$\forall i \in \{1,...,n\}, \forall k, c_i = k \Leftrightarrow t_{ki} = 1 \text{ and } \forall h \neq k,\ t_{hi} = 0$$

We suppose that for each unlabeled example x_i in D_u, there exists a perfect and an imperfect label respectively denoted c_i and $\widetilde{c}_i$. We propose to model the imperfections in the labels by the following probabilities:

$$\forall k, \forall h, \alpha_{kh} = p(\widetilde{c} = k \mid c = h) \tag{1}$$

Which are subject to the constraint:

$$\forall h,\ \sum_k \alpha_{kh} = 1 \tag{2}$$

In order to simplify the presentation, we consider in the following a two-class classification problem.

$$\forall i \in \{1,...,n\}, \forall i' \in \{n+1,...,n+m\},\ \ c_i, \widetilde{c}_{i'} \in \{1,2\}$$

This is not restrictive since we can easily extend the analysis to multi-class cases.

[1] Components of x could also be discrete

3.2 Classification Maximum Likelihhod estimation in presence of label noise

The baseline semi-supervised discriminant algorithm has been conceived as an extension to classification problems of the CEM algorithm proposed by [Celeux and Govaert, 1992] for clustering. CEM has been first proposed for learning the parameters of gaussian mixtures using as training criterion the classification maximum likelihood criterion initially described in [Symons, 1981].

McLachlan has extended CML and CEM to the case where both labeled and unlabeled data are used for learning ([McLachlan, 1992], page 39). The CML criterion in this case is the complete-data likelihood and writes:

$$V_c = \prod_{i=1}^{n}\prod_{k=1}^{2}\{p(x_i, c = k)\}^{t_{ki}} \times \prod_{i=n+1}^{n+m}\prod_{k=1}^{2}\{p(x_i, \widetilde{c} = k)\}^{\widetilde{t}_{ki}}$$

The algorithm proposed by [McLachlan, 1992] optimizes this criterion for a mixture of normal densities. In their algorithm for training discriminant classifiers on labeled and unlabeled data, [Amini and Gallinari-*a*, 2002] make use of the following form of the log-classification-likelihood:

$$\begin{aligned} L_c = {} & \sum_{i=1}^{n}\sum_{k=1}^{2} t_{ki} \log\, p(c = k \mid x_i) + \\ & \sum_{i=n+1}^{n+m}\sum_{k=1}^{2} \widetilde{t}_{ki} \log\, p(\widetilde{c} = k \mid x_i) + \sum_{i=1}^{n+m} \log\, p(x_i) \end{aligned}$$

This writing makes apparent the posterior probabilities which are directly estimated in their algorithm instead of the conditional densities in [McLachlan, 1992]. As no assumptions are made on the distributional nature of data, maximizing L_c is equivalent to the maximization of L'_c ([McLachlan, 1992], page 261).

$$\begin{aligned} L'_c = {} & \sum_{i=1}^{n}\sum_{k=1}^{2} t_{ki} \log\, p(c = k \mid x_i) + \\ & \sum_{i=n+1}^{n+m}\sum_{k=1}^{2} \widetilde{t}_{ki} \log\, p(\widetilde{c} = k \mid x_i) \end{aligned} \qquad (3)$$

Let us now introduce our misclassifiction model. For that, we will express the $p(\widetilde{c} = k \mid x_i)$ in the second summation in (3) as a function of the mislabeling probabilities and of the posterior probability of correct labels $p(c = k \mid x_i)$. Consider

$$p(x_i, \widetilde{c} = k) = \sum_{h=1}^{2} p(x_i, \widetilde{c} = k, c = h)$$

$p(x_i, \widetilde{c} = k, c = h)$ can be decomposed with regard to the conditional class probabilities:

$$p(x_i, \widetilde{c} = k, c = h) = p(x_i \mid \widetilde{c} = k, c = h) \times p(\widetilde{c} = k, c = h)$$

Following [Chittineni, 1980] we make the assumption that the density of an example, given its true label, does not depend on its imperfect label:

$$p(x_i \mid \widetilde{c} = k, c = h) = p(x_i \mid c = h)$$

It then comes:

$$p(x_i, \widetilde{c} = k) = \sum_{h=1}^{2} p(x_i \mid c = h) \times p(c = h) \times p(\widetilde{c} = k \mid c = h)$$

Using Bayes rule and (1), we get:

$$p(x_i, \widetilde{c} = k) = p(x_i) \times \sum_{h=1}^{2} \{\alpha_{kh}\, p(c = h \mid x_i)\} \qquad (4)$$

$p(\widetilde{c} = k \mid x_i)$ can then be derived from $p(c = k \mid x_i)$

$$\forall x_i \in D_u,\ p(\widetilde{c} = k \mid x_i) = \sum_{h=1}^{2} \{\alpha_{kh}\, p(c = h \mid x_i)\} \qquad (5)$$

From (5), (3) becomes

$$\begin{aligned} L'_c = {} & \sum_{i=1}^{n}\sum_{k=1}^{2} t_{ki} \log\, p(c = k \mid x_i) + \\ & \sum_{i=n+1}^{n+m}\sum_{k=1}^{2} \left\{ \widetilde{t}_{ki} \log \left(\sum_{h=1}^{2} \alpha_{kh}\, p(c = h \mid x_i) \right) \right\} \end{aligned} \qquad (6)$$

4 Updating a discriminant function on the basis of labeled and imperfect labeled data

We now present an iterative discriminant CEM algorithm for learning a classifier for semi-supervised learning, which incorporates the mislabeling error model. The training criterion is (6). For simplification, we consider a simple logistic classifier [Anderson, 1982], but the algorithm can be easily adapted for training any discriminant classifier. Consider C_β a logistic classifier with parameters $\{\beta_j\}_{j \in \{0,\dots,d\}}$. The output of C_β for input $x \in \mathbb{R}^d$ is $G(x) = \dfrac{1}{1 + exp(-(\beta_0 + \beta^t x))}$. After β have been learned, $G(x)$ and $1 - G(x)$ are respectively used to estimate $p(c = 1 \mid x)$ and $p(c = 2 \mid x)$. Let, $\pi^{(p)}$ be the current partition for the unlabeled data, $\alpha^{(p)}$, $\beta^{(p)}$ the parameters for the misclassification model and the logistic classifier at iteration p of the algorithm. The learning criterion (6) is a function of α, β and π. An iterative approach is then adopted for its maximization (algorithm 1). Parameters β are first initialized by training the classifier on the labeled dataset D_l. Two steps are then iterated until the convergence of criterion L'_c. In the first step, the classifier is considered as an imperfect supervisor for the unlabeled data. Its outputs, $G(x)$ and $1 - G(x)$, are used to compute the posterior imperfect class probabilities $p(\widetilde{c} = k \mid x)$ for each x in D_u and $k \in \{1, 2\}$, x is then labeled according to the maximum imperfect output. In the second step, the parameters of the error model and of the classifier are updated using the imperfect labels obtained in the previous step as well as the labeled data. We adopted in this step, a gradient algorithm to maximize (6). An advantage of this method is that it only requires the first order derivatives at each iteration. In the following lemma, we provide a proof of convergence of the algorithm to a local maximum of the likelihood function for semi-supervised training.

```
Algorithm 1 Semi-supervised learning with imperfect labels
Initialization : Train a discriminant logistic classifier $C_{\beta^{(0)}}$
                 over $D_l$
For $p \geq 0$, iterate until the convergence of $L'_c$
  1. Apply $C_{\beta^{(p)}}$ on $D_u$, estimate the imperfect class posterior probabilities using the output $G^{(p)}$ of the classifier and get an imperfect label for each $x_i \in D_u$ :
       $\forall x_i \in D_u,\ \widetilde{c}_i^{(p+1)} = \underset{k}{argmax}\, p(\widetilde{c}^{(p)} = k \mid x_i)$
     Let $\pi^{(p+1)}$ be the new partition obtained from this classifier for the unlabeled data.
  2. a) Maximize $L'_c\left(\pi^{(p+1)}, \alpha^{(p)}, \beta^{(p)}\right)$ with respect to $\alpha^{(p)}$ subject to constraints $\forall k, \forall h, \alpha_{kh}^{(p+1)} \in [0,1]$ and $\forall h, \sum_k \alpha_{kh}^{(p+1)} = 1$
     b) Maximize $L'_c\left(\pi^{(p+1)}, \alpha^{(p+1)}, \beta^{(p)}\right)$ with respect to $\beta^{(p)}$
```

Lemma 1. *The modified CML criterion, L'_c, increases for every sequence $\left(\pi^{(p)}, \alpha^{(p)}, \beta^{(p)}\right)$ of Algorithm 1 and the sequence $L'_c\left(\pi^{(p)}, \alpha^{(p)}, \beta^{(p)}\right)$ converges to a stationary point.*

Proof : we first show that L'_c is increasing.

- Since $\forall x_i \in D_u, \widetilde{c}_i^{(p+1)} = k \Leftrightarrow p(\widetilde{c}^{(p+1)} = k \mid x_i) \geq p(\widetilde{c}^{(p+1)} = k' \mid x_i)$ for $k \neq k'$ (step 1), we have
$$L'_c(\pi^{(p+1)}, \alpha^{(p)}, \beta^{(p)}) \geq L'_c(\pi^{(p)}, \alpha^{(p)}, \beta^{(p)})$$
- and, since $\left(\pi^{(p+1)}, \alpha^{(p+1)}, \beta^{(p)}\right)$ and $\left(\pi^{(p+1)}, \alpha^{(p+1)}, \beta^{(p+1)}\right)$ maximize iteratively L'_c (step 2), we have
$$L'_c(\pi^{(p+1)}, \alpha^{(p+1)}, \beta^{(p+1)}) \geq L'_c(\pi^{(p+1)}, \alpha^{(p)}, \beta^{(p)})$$

Finally as there is a finite number of partitions of the example into 2-classes, the increasing sequence $L'_c\left(\pi^{(p)}, \alpha^{(p)}, \beta^{(p)}\right)$ takes a finite number of values and thus, converges to a stationary value. ■

In the following section we will present results on four datasets, using a baseline logistic classifier trained with the updating scheme presented above.

5 Experiments

5.1 Data sets

In our experiments we used the `Spambase`, `Credit screening` and `Mushroom` collections from the UCI repository[2] as well as the `Computation and Language (Cmp_lg)` collection of TIPSTER SUMMAC[3] for text summarization. Table 1 summarizes the characteristics of these datasets. We removed 37 samples with missing attributes from the `Credit` data set. For summarization, we adopt the text-span extraction paradigm which amounts at selecting a subset of representative document sentences. This is a classification problem where sentences are to be classified as relevant or non relevant for the extracted summary. All four problems are two-class classification tasks, `Email Spam` and `Cmp_lg` have continuous attributes, `Credit` vectors are mixed continuous and qualitative, `Mushroom` attributes are qualitative.

Table 1: Characteristics of the datasets and percentage of the smallest of the two classes in the datasets.

Data set	size	Attributes	Smallest class (%)
`Credit`	690(-37)	15	44.5%
`Email spam`	4601	57	39.4%
`Mushroom`	8124	22	48.2%
`Cmp_lg`	28985	5	10%

For each dataset, we ran 3 algorithms - the semi-supervised learning algorithm using the label imperfections, the baseline semi-supervised algorithm [Amini and Gallinari-*a*, 2002] and a fully supervised logistic classifier trained only on the available labeled data. We compared the performance of these algorithms on 20 runs of arbitrary training and test splits, by varying the proportion of labeled-unlabeled data on the training set.

For text summarization, we represent each sentence using a continuous version of features proposed by [Kupiec et al., 1995]. This characterization has given good results in previous work [Amini and Gallinari-*b*, 2002]. Each sentence i, with length $l(i)$ is characterized by $\vec{x}_i = (\varphi_1^i, \varphi_2^i, \varphi_3^i, \varphi_4^i, \varphi_5^i)$, where $\varphi_1^i = \frac{l(i)}{\sum_j l(j)}$ is the normalized sentence length, $\varphi_2^i = \frac{\text{frequency of cue words in } i}{l(i)}$ is the normalized number of cue words in sentence i, $\varphi_3^i = \frac{\text{frequency of acronyms in } i}{l(i)}$ is the normalized number of acronyms in i, φ_4^i is the position indicator feature (beginning, middle or end of the document) and φ_5^i is the normalized number of terms within a generic query q and i. q corresponds to the most frequent words in the training set.

5.2 Evaluation measures

For the three UCI datasets there is approximately the same proportion of examples for each class (table 1). We used as performance criterion the percentage of good classification (PGC) defined as:

$$\text{PGC} = \frac{\text{\# of examples in the test set well classified by the system}}{\text{\# of examples in the test set}}$$

For text summarization, we followed the SUMMAC evaluation by using a 10% compression ratio. Hence, for each document in the test set we have formed its summary by selecting the top 10% sentences having higher score with respect to the output of the classifier. For evaluation we compared these extractive sentences with the desired summary of each document. The desired extractive summaries were generated from the abstract of each article using the text-span alignment method described in [Banko et al., 1999]. Since the collection is not well balanced between positive and negative

[2] ftp://ftp.icsS.uci.edu/pub/machine-learning-databases/

[3] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/

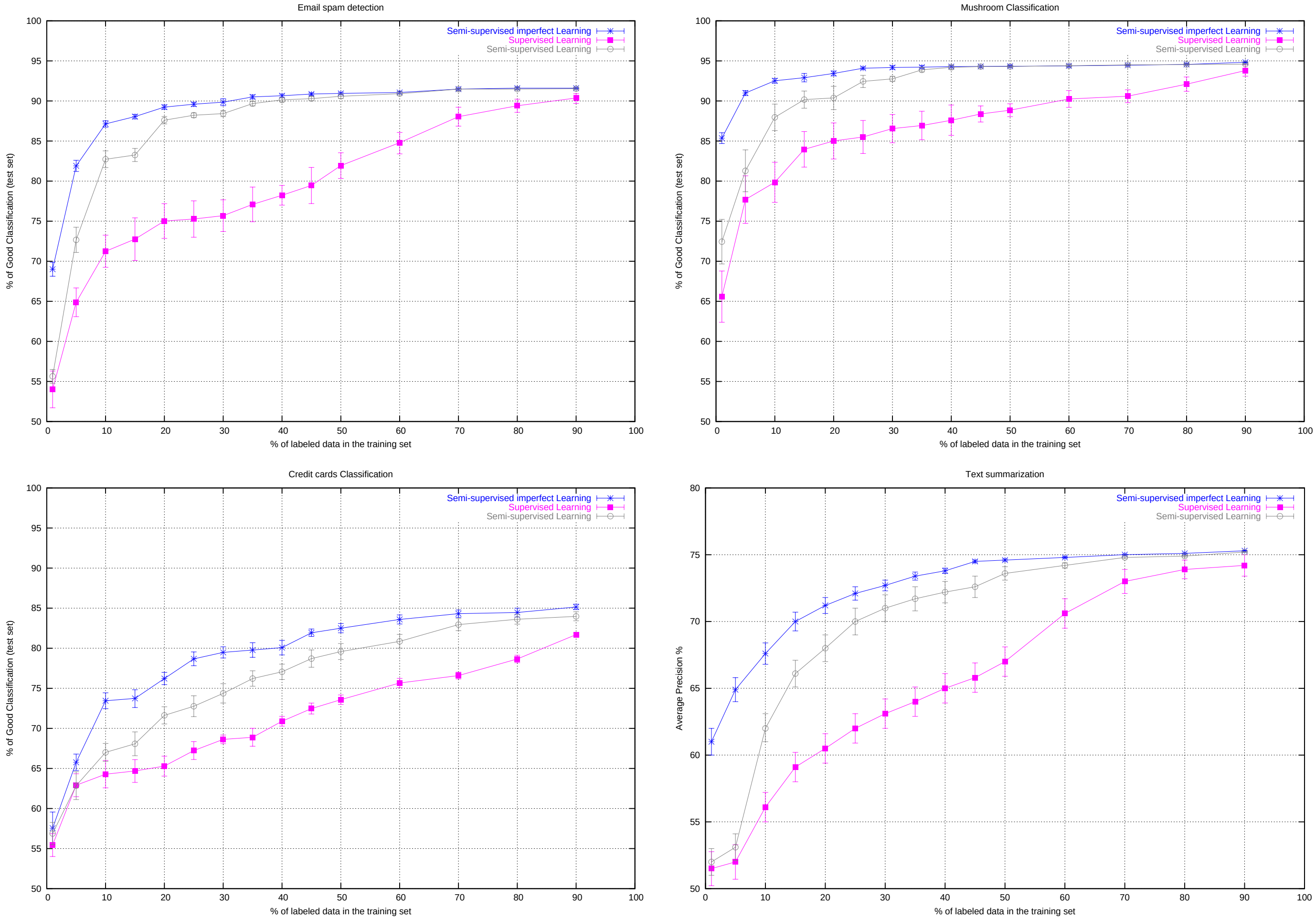

Figure 1: Performance curves for the 4 datasets showing, for a classical logistic classifier trained in a fully supervised scheme (square), the baseline semi-supervised Logistic-CEM algorithm (circle) and our semi-supervised algorithm with label imperfections (star). Each point represents the mean performance for 20 arbitrary runs. The error bars show standard deviati ons for the estimated performance.

examples, PGC is meaningless, we used the average precision (AP) measure for the evaluation. Let ξ, be the number of sentences extracted by the system which are in the target summaries, and σ, be the total number of sentences extracted by the system. The precision is defined as the ratio ξ/σ.

5.3 Results

For each dataset and each cross validation, 25% of examples are held aside as a test set. We vary the percentage of labeled-unlabeled data in the remaining 75% training set of the collections. Figure 1, shows the performance on the test sets for the four datasets as a function of the proportion of labeled data in the training set. On the x-axis, 5% means that 5% of data in the training set were labeled for training, the 95% remaining being used as unlabeled training data. Each experiment was carried on twenty paired trials of randomly selected training-test splits. On the y-axis, each point represents the mean performance for the 20 runs and the error bars correspond to the standard deviation for the estimated performance [Tibshirani, 1996].

All figures exhibit the same behavior. In all four datasets, semi-supervised algorithms are over the baseline classifier trained only on the labeled training data. For example, if we consider text summarization, using only 5% of labeled sentences, our algorithm allows to increase performance by 15% compared to a fully supervised algorithm trained on the 5%. 15% labeled data are needed to reach the same performance with the baseline method. Our model is uniformly better than the two reference models used in our experiments. It provides an important performance increase especially when there are only few labeled data available which is the most interesting situation in semi-supervised learning. For the `Credit card` problem, the dataset is small and semi-supervised learning allows for a smaller increase of performance than for other datasets where there are a lot of unlabeled data available. Semi-supervised learning allows to reach 'optimal' performance with about 50% of labeled data (only 30% for the `Mushroom` set).

6 Conclusion

We have described how to incorporate a label error model into an iterative semi-supervised discriminant algorithm. We have detailed a version of this general algorithm in the case of a simple logistic classifier, shown its convergence and proved empirically its efficiency on four datasets with different characteristics. The algorithm allows for an important performance increase compared to a reference efficient semi-supervised algorithm without mislabeling model. The main contribution of the paper is to provide a general framework for handling simultaneously semi-supervised learning and learning in the presence of label noise. The noise model we have used is simple and allows for efficient estimations in the semi-supervised setting. More sophisticated models have still to be investigated. However, our experience is that simple models do often perform better when only few labeled data are available.

References

[Amini and Gallinari-*a*, 2002] M. R. Amini and P. Gallinari. Semi-supervised logistic regression. In *Proceedings of the* 15^{th} *European Conference on Artificial Intelligence*, pages 390–394, 2002. ECAI.

[Amini and Gallinari-*b*, 2002] M. R. Amini and P. Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the* 25^{th} *International ACM SIGIR Conference*, pages 105–112, 2002. SIGIR.

[Anderson, 1979] J. A. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

[Anderson, 1982] J. A. Anderson. Logistic discrimination. In *Handbook of Statistics*, P.R. Krishnaiah and L. Kanal (Eds.), 2:169–191, 1982.

[Banko et al., 1999] M. Banko, V. Mittal, M. Kantrowitz, J. Goldstein. Generating extraction-based summaries from hand-written done by text alignement. In *Proceedings of the Pacific Association for Computational Linguistics*, 1999. PACLING.

[Blum and Mitchell, 1998] A. Blum and T. Mitchell. Combining labeled and unlabeled data with Co-training. In *Proceedings of the Workshop on Computational Learning Theory*, pages 92–100, 1998. COLT.

[Celeux and Govaert, 1992] G. Celeux and G. Govaert. A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332, 1992.

[Chittineni, 1980] C. B. Chittineni. Learning with imperfect labeled patterns. *Pattern Recognition*, 12(5):281–291, 1980.

[Day, 1969] N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474, 1969.

[Dempster et al., 1977] A. Dempster, N. Laird and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, Series B, 39(1):1–38, 1977.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the* 16^{th} *International Conference on Machine Learning*, pages 200–209, 1999. ICML.

[Krishnan, 1988] T. Krishnan. Efficiency of learning with imperfect supervision. *Pattern Recognition*, 21(2):183–188, 1988.

[Kupiec et al., 1995] J. Kupiec, J. Pederson, F. A. Chen. Trainable Document Summarizer. In *Proceedings of the* 18^{th} *International ACM SIGIR Conference*, pages 68–73, 1995. SIGIR.

[Lawrence and Schölkopf, 2001] N. D. Lawrence and B. Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. In *Proceedings of the* 18^{th} *International Conference on Machine Learning*, pages 306–313, 2001. ICML.

[McLachlan, 1992] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc. New York, 1992.

[McLachlan and Ganesalingam, 1982] G. J. McLachlan and S. Ganesalingam. Updating a discriminant function in basis of unclassified data. *Communication on Statistics-Simulation and Computation*, 11(6):753–767, 1982.

[Miller and Uyar, 1996] D. Miller and H. Uyar. A mixture of experts classifier with learning based on both labeled and unlabeled data. Advances in Neural Information Processing Systems 9, pages 571–577, 1996. NIPS.

[Murray and Titterington, 1978] G. D. Murray and D. M. Titterington. Estimation problems with data from a mixture. *Applied Statistics*, 27(3):325–334, 1978.

[Muslea et al., 2002] I. Muslea, S. Minton and C. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the* 19^{th} *International Conference on Machine Learning*, pages 435–442, 2002. ICML.

[Nigam et al., 2000] K. Nigam, A. K. McCallum, S. Thrun, T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[O'Neill, 1978] T. J. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, 1978.

[Roth and Steinhage, 1999] V. Roth and V. Steinhage. Nolinear discriminant analysis using kernel functions. Advances in Neural Information Processing Systems 12, pages 568–574, 1999. NIPS.

[Symons, 1981] M. J. Symons. Clustering criteria and multivariate normal mixture. *Biometrics*, 37(1):35–43, 1981.

[Tibshirani, 1996] R. Tibshirani. A Comparison of Some Error Estimates for Neural Network Models *Neural Computation*, 8:152–163, 1996.

[Titterington, 1989] D. M. Titterington. An alternative stochastic supervisor in discriminant analysis *Pattern Recognition*, 22(1):91–95, 1989.