2004s-28

Learning from Partial Labels with Minimum Entropy

Yves Grandvalet, Yoshua Bengio

Série Scientifique
Scientific Series

Montréal
Mai 2004

CIRANO
Centre interuniversitaire de recherche
en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de la Recherche, de la Science et de la Technologie, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère de la Recherche, de la Science et de la Technologie, and grants and research mandates obtained by its research teams.

Les organisations-partenaires / The Partner Organizations

PARTENAIRE MAJEUR
. Ministère du développement économique et régional et de la recherche [MDERR]

PARTENAIRES
. Alcan inc.
. Axa Canada
. Banque du Canada
. Banque Laurentienne du Canada
. Banque Nationale du Canada
. Banque Royale du Canada
. Bell Canada
. BMO Groupe Financier
. Bombardier
. Bourse de Montréal
. Caisse de dépôt et placement du Québec
. Développement des ressources humaines Canada [DRHC]
. Fédération des caisses Desjardins du Québec
. GazMétro
. Hydro-Québec
. Industrie Canada
. Ministère des Finances du Québec
. Pratt & Whitney Canada Inc.
. Raymond Chabot Grant Thornton
. Ville de Montréal

. École Polytechnique de Montréal
. HEC Montréal
. Université Concordia
. Université de Montréal
. Université du Québec à Montréal
. Université Laval
. Université McGill
. Université de Sherbrooke

ASSOCIE A :
. Institut de Finance Mathématique de Montréal (IFM^2)
. Laboratoires universitaires Bell Canada
. Réseau de calcul et de modélisation mathématique [RCM^2]
. Réseau de centres d'excellence MITACS (Les mathématiques des technologies de l'information et des systèmes complexes)

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.
This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Learning from Partial Labels with Minimum Entropy

Yves Grandvalet[*], *Yoshua Bengio* [†]

Résumé / *Abstract*

Cet article introduit le régularisateur à entropie minimum pour l'apprentissage d'étiquettes partielles. Ce problème d'apprentissage incorpore le cadre non supervisé, où une règle de décision doit être apprise à partir d'exemples étiquetés et non étiquetés. Le régularisateur à entropie minimum s'applique aux modèles de diagnostics, c'est-à-dire aux modèles des probabilités postérieures de classes. Nous montrons comment inclure d'autres approches comme un cas particulier ou limité du problème semi-supervisé. Une série d'expériences montrent que le critère proposé fournit des solutions utilisant les exemples non étiquetés lorsque ces dernières sont instructives. Même lorsque les données sont échantillonnées à partir de la classe de distribution balayée par un modèle génératif, l'approche mentionnée améliore le modèle génératif estimé lorsque le nombre de caractéristiques est de l'ordre de la taille de l'échantillon. Les performances avantagent certainement l'entropie minimum lorsque le modèle génératif est légèrement mal spécifié. Finalement, la robustesse de ce cadre d'apprentissage est démontré : lors de situations où les exemples non étiquetés n'apportent aucune information, l'entropie minimum retourne une solution rejetant les exemples non étiquetés et est aussi performante que l'apprentissage supervisé.

Mots clés : apprentissage discriminant, apprentissage semi-supervisé, entropie minimum.

This paper introduces the minimum entropy regularizer for learning from partial labels. This learning problem encompasses the semi-supervised setting, where a decision rule is to be learned from labeled and unlabeled examples. The minimum entropy regularizer applies to diagnosis models, i.e. models of the posterior probabilities of classes. It is shown to include other approaches to the semi-supervised problem as particular or limiting cases. A series of experiments illustrates that the proposed criterion provides solutions taking advantage of unlabeled examples when the latter convey information. Even when the data are sampled from the distribution class spanned by a generative model, the proposed approach improves over the estimated generative model when the number of features is of the order of sample size. The performances are definitely in favor of minimum entropy when the generative model is slightly misspecified. Finally, the robustness of the learning scheme is demonstrated: in situations where unlabeled examples do not convey information, minimum entropy returns a solution discarding unlabeled examples and performs as well as supervised learning.

Keywords: *discriminant learning, semi-supervised learning, minimum entropy.*

[*] Yves Grandvalet, Université de Technologie de Compiègne, France, tél. : +011 33(0)3 44 23 49 28, grandval@utc.fr

[†] Yoshua Bengio, Université de Montréal et CIRANO, (514) 343-6804, Yoshua.Bengio@cirano.qc.ca

1. Introduction

In the classical supervised learning classification framework, a decision rule is to be learned from a learning set $\mathcal{L}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$. Each example is described by a pattern $\mathbf{x}_i \in \mathcal{X}$ and by the response of a supervisor $y_i \in \Omega = \{\omega_1, \ldots, \omega_K\}$. This response variable is supposedly *the* correct class among the finite set of exclusive classes Ω.[1]

This paper proposes an estimation principle applicable to probabilistic classifiers when the learning set includes examples whose class is not precisely known. We consider the situation where the goal of learning is still to provide a decision rule at any point of the input space $\mathcal{X}$, but where, instead of answering the correct class, the supervisor only returns a subset of possible classes which is supposed to include the correct solution. This kind of information is sometimes a more faithful description of the true state of knowledge when labeling is performed by an expert. For example, in medical diagnosis, a physician is sometimes able to discard some diseases, but not to pinpoint the precise illness of his patient. Last but not least, some examples may not be labeled at all: in particular, semi-supervised learning[2] is a special case of partially labeled problem, where all examples are either precisely labeled or unlabeled, i.e. with labels allowing the whole Ω.

Partial labeling has been investigated in the frameworks of probability and Dempster-Shafer theories (Ambroise et al., 2001). Dempster-Shafer theory enables to reason on beliefs expressed on subsets of Ω without distributing them to singletons. Its description is out of the scope of this paper, which focuses on the probabilistic framework. The reader is referred to (Ambroise et al., 2001), where classifiers based on Dempster-Shafer theory are compared to probabilistic mixture models.

In the probabilistic framework, partial labels can be modeled as a missing data problem, which can be adressed by generative models such as mixture models thanks to the EM algorithm and extensions thereof (McLachlan, 1992; Ambroise et al., 2001). Generative models apply to the joint density of patterns and class (X, Y). They have appealing features: besides discrimination, they can be used for other tasks, such as outlier detection. However, they also have drawbacks. Their estimation is much more demanding than discriminative models, since the joint density model of (X, Y) is necessarily more complex than the conditional model of $(Y|X)$. This means that more parameters are to be estimated, resulting in more uncertainty in the estimation process. In addition, the fitness measure for joint density models is not discriminative, which means that better models are not necessarily better predictors of class labels. Finally, the generative model being more precise, it is more likely to be misspecified. These difficulties have lead to proposals aiming at exploiting partially labeled data in the framework of probabilistic classification by diagnosis models, i.e. models of the posterior class probabilities (Grandvalet, 2002; Jin & Ghahramani, 2003).

In this paper, we first formalize the partial labeling problem in the probabilistic framework. At this point, we emphasize the necessary assumptions pertaining to the generation of partial labels. The conditional likelihood shows that unlabeled data are not informative in the diagnosis paradigm. We thus look at theoretical results guiding the search for a sensible induction bias to be made in this setting. The latter is encoded by a prior distribution in the maximum a posteriori framework. The maximization of the posterior is then shown to be related to previous existing approaches such as self-learning or semi-supervised support vector machines. We finally demonstrate the performance of discriminant models trained with this criterion of on a series of semi-supervised problems, taylored to be favorable to a well-founded challenger: generative mixture models.

2. Derivation of the criterion

2.1. Likelihood

We first look at how the partial labeling problem fits the likelihood estimation principle. The learning set is now $\mathcal{L}_n = \{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$, where $\mathbf{z} \in \{0, 1\}^K$ denote the dummy variable representing partial labels. It is the indicator of the subset returned by the supervisor, where $z_k = 1$ means that ω_k is in the subset (i.e. is a possible label), whereas $z_k = 0$ means that the true label y is definitely not ω_k.

Besides the fact that the possible label always includes the true label (i.e. $P(\mathbf{z}|\mathbf{x}, \omega_k) = 0$ if $z_k = 0$), we assume that the labeling information is missing completely at random (i.e $(\forall(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, \forall \mathbf{z} : z_k = z_\ell = 1)$ $P(\mathbf{z}|\mathbf{x}, \omega_k) = P(\mathbf{z}|\mathbf{x}', \omega_\ell)$). Note that the first assumption (which could be relaxed) *does not* mean that the Bayes error is null: one may observe two identical patterns with incompatible partial labels provided $P(\omega_k|\mathbf{x})$ and $P(\omega_\ell|\mathbf{x})$ are strictly positive.

From the above-mentioned assumptions, we derive

$$P(\omega_k|\mathbf{x}, \mathbf{z}) = \frac{z_k P(\omega_k|\mathbf{x})}{\sum_{\ell=1}^K z_\ell P(\omega_\ell|\mathbf{x})} \ , \qquad (1)$$

i.e. $P(\omega_k|\mathbf{x}, \mathbf{z})$ is the the Kullback-Leibler projection of

[1] Note that correct labels do not imply that the Bayes error is null; examples described by the same pattern $\mathbf{x}$ may have different labels. This diversity is not supposed to be generated by errors in supervisor's response, but to arise from the limited description of examples provided by the pattern $\mathbf{x}$.

[2] In the terminology used here, semi-supervised learning refers to generalizing (i.e. learning a decision rule) from labeled and unlabeled data. We do not consider the related but distinct problem of predicting labels on a set of predefined patterns.

$P(\omega_k|\mathbf{x})$ on the set of distribution compatible with $\mathbf{z}$.

Assuming independent examples, the conditional log-likelihood of $P(Z|X)$ on the observed sample is

$$L(\boldsymbol{\theta};\mathcal{L}_n) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i;\theta_k) \right) + h(\mathbf{z}_i) \quad , \quad (2)$$

where $h(\mathbf{z})$ is only affected by the missingness mechanism, regardless of $P(X,Y)$ and $f_k(\mathbf{x};\theta_k)$ is the model of $P(\omega_k|\mathbf{x})$ parameterized by θ_k, and $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^{K}$.

The conditional log-likelihood is referred to as “minimum commitment” by Grandvalet (2002) because it assumes minimal requirements on the distribution of Z; it is also named the EM model by Jin and Ghahramani (2003) since it can be optimized by the EM algorithm. This criterion is a concave function of $f_k(\mathbf{x}_i;\theta_k)$, and for simple models such as the ones provided by logistic regression, it is also concave in $\boldsymbol{\theta}$, so that the global solution can be obtained by the Newton-Raphson algorithm.

The diagnosis paradigm, where the conditional log-likelihood (2) is maximized, corresponds to maximizing the complete likelihood if no assumption whatsoever is made on $P(X)$ (McLachlan, 1992). In this framework, unlabeled data convey no information and thus do not affect the likelihood. More generally, any distribution with constant mass $m_i = \sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i;\theta_k)$ achieves the same value of the criterion.

In the Bayesian maximum a posteriori (MAP) framework, Seeger (2002) shows that unlabeled data are useless regarding discrimination when the priors on $P(X)$ and $P(Y|X)$ factorize. In other words, observing $\mathbf{x}$ does not inform about y, unless the modeler assumed that it should be the case. Hence, if we are willing to benefit from unlabeled examples in the diagnosis paradigm, we have to assume some relationship between $\mathbf{x}$ and y. In the Bayesian framework, this relationship is encoded by a prior distribution. There is however no such thing like a universally relevant prior knowledge. Here, we chose to look for an assumption which should be able to take advantage of unlabeled examples when the latter are known to be beneficial.

2.2. When Are Unlabeled Examples Informative?

There has been little theoretical work in the general setting of partial labeling. Even in the semi-supervised case, theory gives little backup for the numerous experimental evidences (e.g. (Ambroise et al., 2001; Bennett & Demiriz, 1999; Grandvalet, 2002; Jin & Ghahramani, 2003; Nigam et al., 2000; Nigam & Ghani, 2000)) showing that unlabeled examples can help the learning process.

Theory has been mostly developed at the two extremes of the statistical learning paradigm: in parametric statistics where examples are known to be generated from a known class of distribution, and in the distribution-free Structural Risk Minimization (SRM) or Probably Approximately Correct (PAC) frameworks. Semi-supervised learning, in the terminology used here, does not fit the distribution-free SRM or PAC frameworks: no positive statement can be made without distributional assumptions, since one can find distributions $P(X,Y)$, for which learning from labeled data is easy and unlabeled data are non-informative.In this regard, generalizing from labeled and unlabeled data differs from transductive inference, where the goal is to infer the label of known patterns from a set of labeled examples (Vapnik, 1998, chapter 8).

In parametric statistics, some theoretical studies have shown the benefit of unlabeled examples in the parametric setting, either for specific distributions (O’Neill, 1978), or for general mixtures of the form $P(\mathbf{x}) = pP(\mathbf{x}|\omega_1) + (1-p)P(\mathbf{x}|\omega_2)$, when the estimation problem is essentially reduced to the one of estimating the mixture parameter p (Castelli & Cover, 1996). These studies confirm what intuition suggests: the (asymptotic) information content of unlabeled examples decreases as classes overlap.[3] Thus, the assumption that classes are somewhat separated is sensible if we expect to take advantage of unlabeled examples.

The conditional entropy of class labels conditioned on the observed variables

$$H(Y|X,Z) = -E_{XYZ}[\log P(Y|X,Z)] \quad , \qquad (3)$$

is a measure of class overlap. It has the advantage of being invariant to the parameterization of the model.

In the Bayesian framework, assumptions are encoded by means of a prior on the model parameters. Stating that we expect a high conditional entropy does not uniquely define the form of the prior distribution, but the latter can be derived by resorting to the maximum entropy principle.[4] Let $(\boldsymbol{\theta},\boldsymbol{\psi})$ denote the joint model parameters, the maximum entropy prior verifying $E_{\Theta\Psi}[H(Y|X,Z)] = c$, where c is the constant quantifying how small the entropy should be on average, takes the form

$$P(\boldsymbol{\theta},\boldsymbol{\psi}) \propto \exp\left(-\lambda H(Y|X,Z)\right) \quad , \qquad (4)$$

where λ is the positive Lagrange multiplier corresponding to the constant c.

[3] This statement appears explicitly in (O’Neill, 1978), and is also formalized, though not stressed in (Castelli & Cover, 1996), where the Fisher information for unlabeled examples at the estimate $\hat{p}$ is clearly a measure of the overlap between class conditional densities: $I_u(\hat{p}) = \int \frac{(P(\mathbf{x}|\omega_1)-P(\mathbf{x}|\omega_2))^2}{\hat{p}P(\mathbf{x}|\omega_1)+(1-\hat{p})P(\mathbf{x}|\omega_2)}\, d\mathbf{x}$.

[4] Here, maximum entropy refers to the construction principle which enables to derive distributions from constraints, not to the content of priors regarding entropy.

This prior requires a model of the joint distribution $P(X, Y, Z)$ when the choice of the diagnosis paradigm is motivated by the possibility to limit modeling to conditional probabilities. The additional modeling can be avoided by applying the plug-in principle, i.e. by replacing the expectation with respect to (X, Z) by the average over the training sample. This substitution can be interpreted as "modeling" $P(X, Z)$ by its empirical distribution.

$$H_{\mathrm{emp}}(Y|\mathcal{L}_n) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} P(\omega_k|\mathbf{x}_i, \mathbf{z}_i) \log P(\omega_k|\mathbf{x}_i, \mathbf{z}_i) \ . \qquad (5)$$

This empirical measure is invariant to the parameterization of the model of conditional probabilities. When plugged in for $H(Y|X, Z)$ in (4), it defines an empirical prior (i.e whose form is partly defined from data, see Berger (1985) for other examples) on parameters $\boldsymbol{\theta}$.

2.3. Minimum Entropy Criterion

Recalling that $f_k(\mathbf{x}; \theta_k)$ denotes the model of $P(\omega_k|\mathbf{x})$, the model of $P(\omega_k|\mathbf{x}, \mathbf{z})$ (1) is defined as follows:

$$g_k(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \frac{z_k f_k(\mathbf{x}; \theta_k)}{\sum_{\ell=1}^{K} z_\ell f_\ell(\mathbf{x}; \theta_\ell)} \ .$$

From now on, we drop the reference to parameters in functions f_k and g_k to lighten notation. The MAP estimate is the maximizer of the posterior distribution, i.e. the maximizer of

$$\begin{aligned} C(\boldsymbol{\theta}, \lambda; \mathcal{L}_n) =& \ L(\boldsymbol{\theta}; \mathcal{L}_n) - \lambda H_{\mathrm{emp}}(Y|\mathcal{L}_n) \\ =& \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i)\right) + \\ & \lambda \sum_{i=1}^{n}\sum_{k=1}^{K} g_k(\mathbf{x}_i, \mathbf{z}_i) \log g_k(\mathbf{x}_i, \mathbf{z}_i) \ , \end{aligned} \qquad (6)$$

where the constant terms in the log-likelihood (2) and log-prior (4) have been droppped.

For a labeled example, $g_k(\mathbf{x}_i, \mathbf{z}_i) = z_{ik}$, and for an unlabeled example, $g_k(\mathbf{x}_i, \mathbf{z}_i) = f_k(\mathbf{x}_i)$. Hence, in semi-supervised learning, $H_{\mathrm{emp}}(Y|\mathcal{L}_n)$ is only affected by the value of $f_k(\mathbf{x})$ on unlabeled examples. More generally, this part of the criterion is only influenced by the predicted distribution of probability masses within the subset of possible labels. In this sense, the role of $H_{\mathrm{emp}}(Y|\mathcal{L}_n)$ is orthogonal to the one of likelihood.

Entropy regularization biases models toward less ambiguity. The posterior probabilities are drived to $\{0, 1\}$, i.e. toward over-confidence. Hence, C should be regarded as a decision-oriented criterion. Let $D(P\|Q)$ denote the Kullback-Leibler divergence between distributions P and Q. Up to an irrelevant constant, C can be written as follows:

$$C(\boldsymbol{\theta}, \lambda; \mathcal{L}_n) = L(\boldsymbol{\theta}; \mathcal{L}_n) + \lambda \sum_{i=1}^{n} D(\mathbf{g}(\mathbf{x_i}, \mathbf{z_i})\|\boldsymbol{\pi}) \ , \qquad (7)$$

where $\boldsymbol{\pi} = (1/K \dots 1/K \dots 1/K)^T$ is the uniform distribution. This writing shows that C pushes probability masses away from uniformity. This behavior is sensible for the $\{0, 1\}$-loss function, for which the decision boundary is defined by $\arg\max_k f_k(\mathbf{x})$, but other reference distributions could be considered for other loss functions.

Finally, note that, in the experimental section below, we added a constraint $E_\Theta[S(\Theta)] = c'$ to build the prior, in order to encode smoothness assumptions on the posterior probabilities. This constraint simply appears as a third additional term in the criterion C (6) with its corresponding Lagrange multiplier ν. It is important to formalize such a smoothness assumption, which may take different forms, in order to prevent the estimate of posterior probabilities $f_k(\mathbf{x}_i)$ to be driven to $\{0, 1\}$ by the minimization of entropy.

3. Related Work

3.1. Self-Training

Self-training (Nigam & Ghani, 2000) is an iterative process, where a learner imputes the labels of examples which have been classified with confidence in the previous step. Amini and Gallinari (2002) analyze this technique and shown that it is equivalent to a version of the classification EM algorithm. The classification EM algorithm (Celeux & Govaert, 1992) minimizes the likelihood deprived of the entropy of the partition. In the context of conditional likelihood with labeled and unlabeled examples only, the criterion is

$$\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} z_{ik} f_k(\mathbf{x}_i)\right) + \sum_{k=1}^{K} f_k(\mathbf{x}_i) \log f_k(\mathbf{x}_i) \ ,$$

which is recognized as an instance of the criterion (6) with $\lambda = 1$.

Self-confident logistic regression (Grandvalet, 2002) is another algorithm optimizing the criterion for $\lambda = 1$. Using smaller λ values is expected to have two benefits: first, the importance of unlabeled examples can be controlled, in the spirit of the EM-λ (Nigam et al., 2000), and second, slowly increasing λ defines a deterministic annealing scheme (Rose et al., 1990) which should help the optimization process to avoid poor local minima of the criterion.

3.2. Minimum entropy methods

Minimum entropy regularizers have already been used in other contexts to encode learnability priors (see e.g. Brand (1999)). In a sense, H_{emp} can be seen as a poor's man way to generalize this approach to continuous input spaces.

3.3. Input-Dependent Regularization

Our criterion differs from input-dependent regularization (Seeger, 2002; Szummer & Jaakkola, 2003) in that it is expressed only in terms of $P(Y|X,Z)$ and does not involve $P(X)$. However, we stress that for unlabeled data, the regularizer agrees with the complete likelihood provided $P(X)$ is small near the decision surface. Indeed, whereas a generative model would maximize $\log P(X)$ on the unlabeled data, the minimum entropy criterion minimizes the conditional entropy on the same points. In addition, when the model is regularized (e.g. with weight decay), the conditional entropy is prevented be too small close to the decision surface. This will favor putting the decision surface in a low density area.

3.4. Maximal Margin Separators

Maximal margin separators are well founded models which have shown great success in supervised classification. For linearly separable data, they have been shown to be a limiting case of probabilistic hyperplane separators (Tong & Koller, 2000). In the framework of transductive learning, Vapnik (Vapnik, 1998) proposes to broaden the margin definition to unlabeled examples, such as the margin is the smallest Euclidean distance between any (labeled and unlabeled) training point to the classification boundary. The following theorem generalizes (Tong & Koller, 2000) to the margin defined in transductive learning.

Theorem 1 *In the two-class linear separable case, the minimum entropy criterion applied to the regularized logistic regression model converges toward a maximum margin separator (with maximal distance from labeled and unlabeled examples) as the regularization term goes to zero.*

Sketch of proof – Consider the logistic regression model parameterized by $\boldsymbol{\theta} = (\mathbf{w}, \mathbf{b})$: $f_k(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{1+\exp(-\mathbf{w}^T(\mathbf{x}-\mathbf{b}))}$. Let $(\mathbf{w}^*, \mathbf{b}^*)$ be the maximum of $C(\boldsymbol{\theta}, \lambda; \mathcal{L}_n) - \nu\|\mathbf{w}\|^2$.

The first term on the right-hand-side of (6) is only affected by labeled data, and goes to its maximum, zero, when these examples are all correctly classified and that the responses of the model are saturated $(2z_{ik}-1)(2f_k(\mathbf{x}_i)-1) \to 1$. The second term is only affected by unlabeled examples and also goes to zero provided $|2f_k(\mathbf{x}_i)-1| \to 1$. These two objectives can be pursued at the same time in the limit of $\nu \to 0$, where logistic regression may converge toward any arbitrarily hard linear separator as $\|\mathbf{w}\|^2 \to \infty$. Hence, at $(\mathbf{w}^*, \mathbf{b}^*)$, there should be no misclassified examples. Let $m_i = \mathbf{w}^T(\mathbf{x_i} - \mathbf{b})$ denote the *margin* for example i. For labeled samples, the gradient as $(2z_{ik}-1)(2f_k(\mathbf{x}_i)-1) \to 1$ goes exponentially to 0: $\partial C/\partial m_i \to \exp(-m_i)$ if m_i is positive and $\partial C/\partial m_i \to -\exp(m_i)$ if m_i is negative; for unlabeled samples, the gradient as $|2f_k(\mathbf{x}_i)-1| \to 1$ goes exponentially to 0: $\partial C/\partial m_i \to \lambda m_i \exp(-m_i)$ if m_i is positive and $\partial C/\partial m_i \to \lambda m_i \exp(m_i)$ if m_i is negative. Therefore, once the labeled examples are hardly separated, the influence of examples on C decreases exponentially with their distance to the decision boundary. Thus the decision boundary is essentially determined by the examples with smallest margin (whether they are labeled or unlabeled), the so-called support vectors. Furthermore, the cancellation of the contribution of support vectors to the gradient requires that they should all be at the same distance from the decision boundary.

Hence, the minimum entropy solution can closely mimic the semi-supervised SVM (Bennett & Demiriz, 1999), which partially solves the enumeration problem of the original solution proposed by Vapnik. Note however that our criterion is not concave, so that the convergence toward the global maximum cannot be guaranteed. To our knowledge, this apparent fault is shared by all semi-supervised algorithms learning a decision rule and dealing with large samples of unlabeled data in reasonable time. Since generative and diagnosis algorithms consist in imputing labels expliciety or implicitly, avoiding enumeration involves some kind of heuristic process which may fail.

4. Experiments

For simplicity, we focus on the semi-supervised learning task which is the most frequently encountered. The experimental setup is simple in order to avoid artifacts stemming from optimization problems. Our goal is to check to what extent supervised learning can be improved by unlabeled examples, and if minimum entropy can compete with generative models which are usually advocated in this framework.

The minimum entropy criterion is applied to the logistic regression model, and compared to the EM algorithm for mixture models. Logistic regression fitted by maximum likelihood (i.e. ignoring unlabeled data) and logistic regression with all labels known are also computed for reference. The former shows what has been gained (or lost) by trying to benefit from unlabeled data, and the latter shows what has been lost with the missing labels; it thus provides a bound on the ultimate performance that one could possibly achieve. All hyper-parameters (weight-decay for all logistic regression models plus the λ parameter (7) for minimum entropy) are tuned by ten-fold cross-validation.

4.1. Correct joint density model

In the first series of experiments, we consider two-class problems in an input space of size 50. Each class is generated with equal probability from a multivariate normal distribution. Class ω_1 is multivariate normal with mean $(aa \dots a)$ and unit covariance matrix. Class ω_2 is multivariate normal with mean $-(aa \dots a)$ and unit covariance matrix. Parameter a tunes the Bayes error which varies from 1 % to 20 % (1 %, 2.5 %, 5 %, 10 %, 20 %). The learning sets comprise n_l labeled examples, ($n_l = 50, 100, 200$) and n_u unlabeled examples, ($n_u = n_l \times (1, 3, 10, 30, 100)$). Overall, 75 different setups are evaluated, and for each one, 10 different training samples are generated. Generalization performances are estimated on a test set of size $10\,000$.

This benchmark provides a comparison for the algorithms in a situation where unlabeled data are known to convey information. It is favorable to the generative model in two respects. First, the mixture models use the *correct* model that generated data (two Gaussian subpopulations, with identical covariances). The logistic regression model is only *compatible* with the joint distribution, which is a weaker fulfillment than the correctness. Second, the problem of local maxima in the likelihood function is artificially cured for mixture models: the EM estimation algorithm is initialized with the parameters of the true distribution. This initialization advantages mixture models, since it guaranties to pick, among all local maxima of the likelihood, the one which is in the basin of attraction of the optimal value.

As there is no modeling bias, differences in prediction error rates are only due to differences in estimation efficiency. The overall error rates (averaged over all settings) are in favor of minimum entropy logistic regression (14.1 ± 0.3 %). The EM algorithm (15.6 ± 0.3 %) does worse in average than logistic regression (14.9 ± 0.3 %). For reference, the average Bayes recognition rate is 7.7 % and logistic regression reaches 10.4 ± 0.1 % when all examples are labeled. Figure 1 provides more informative summaries than these raw numbers. The plots represent the recognition rates versus Bayes error rate and the n_u/n_l ratio. Each curve reports the results averaged over n_l. The first plot shows that for the generative and the diagnosis models, unlabeled examples are mostly beneficial when the Bayes error is low (the classes don't overlap much). This experimental observation confirms what intuition and asymptotic theory suggest (Castelli & Cover, 1996; O'Neill, 1978), and validates the relevance of the minimum entropy assumption. This graph also illustrates the consequence of the demanding parametrization of generative models. Mixture models are outperformed by the simple logistic regression model when the sample size is low, since their number of parameters grows quadratically (compared to linearly) with the number of input features.

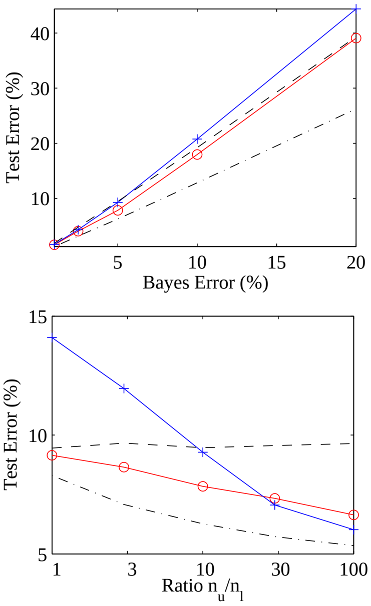

Figure 1. Left: test error *vs.* Bayes error rate for $n_u/n_l = 10$; right: test error *vs.* n_u/n_l ratio for 5 % Bayes error ($a = 0.23$). Average results of minimum entropy logistic regression ($\circ$) and mixture models ($+$). The performance of logistic regression (dashed), and logistic regression with all labels known (dash-dotted) are shown for reference.

The second plot shows that the minimum entropy model takes quickly advantage of unlabeled data when classes are well separated. With $n_u = 3n_l$, the model considerably improves upon the one discarding unlabeled data. At this stage, the generative models do not perform well, as the number of available examples is low compared to the number of parameters in the model, which grows quadratically with the number of variables, while it grows linearly for the diagnosis model. However, for very large sample sizes, with 100 times more unlabeled examples than labeled examples, the generative approach becomes more accurate than the diagnosis approach.

4.2. Misspecified joint density model

In a second series of experiments, we kept everything fixed, except that the class-conditional densities are now slightly corrupted by outliers. For each class, the examples are now generated from a mixture of two Gaussians centered on the same mean: a unit variance component gathers 98 % of examples, while the remaining 2 % are generated from a large variance component, where each variable has a standard deviation of 10. The model of the distribution is not modified in the fitted generative model which is now slightly misspecified. Again, the EM estimation algorithm is advantaged by initializing it with the optimal parameters on the test set.

The results are displayed in the left-hand-side of Figure 2. They should be compared with the right-hand-side of Figure 1. The generative model suffers greatly from the slightly misspecified distribution model and behaves much worse than logistic regression for all sample sizes. The unlabeled examples have first a beneficial effect on test error, but they turn to have a detrimental effect when they overwhelm the number of labeled examples. On the other hand, the diagnosis models behave smoothly as in the previous case, and the minimum entropy criterion improves performance.

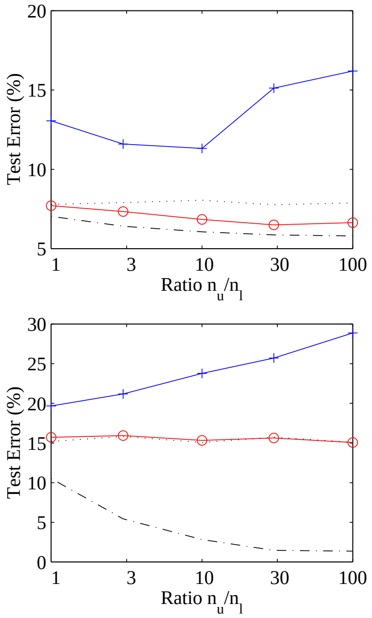

Figure 2. Test error *vs.* n_u/n_l ratio for $a = 0.23$. Average results of minimum entropy logistic regression ($\circ$) and mixture models ($+$). The performance of logistic regression (dotted), and logistic regression with all labels known (dash-dotted) are shown for reference. Left: second experiment with outliers; right: third experiment with uninformative unlabeled examples.

In the last series of experiments, we investigate the robustness with respect to the minimum entropy assumption, by testing it on distributions where unlabeled examples are not informative, and where a low density of $P(X)$ does not indicate a boundary region. The examples are drawn from two Gaussian clusters like in the first series of experiment, but the label is now independent of this clustering: an example $\mathbf{x}$ belongs to class ω_1 if $x_2 > x_1$ and belongs to class ω_2 otherwise. The Bayes decision boundary is now located in the middle of each cluster. The mixture model is unchanged. It is now far from the model used to generate data. We continue to favor the EM estimation algorithm by initializing it with the optimal parameters on the test set. The right-hand-side plot of Figure 1 shows that this favorable initialization does not prevent the model to be fooled by unlabeled data: its test error steadily increases with the amount of unlabeled data. On the other hand, the diagnosis models behave well, and the minimum entropy algorithm is not distracted by the two clusters; its performance is identical to the one of training with labeled data only, which can be regarded as the ultimate performance in this situation.

5. Discussion

In this paper, we proposed a minimum entropy regularizer to allow the application of supervised classification techniques in the context of partial labels (a generalization of the semi-supervised setting in which the target is only given up to a subset of the classes). In this paper, we proposed a minimum entropy regularizer to handle partial labels with supervised classification techniques. This regularizer introduces an induction bias which is motivated by the theoretical results showing that unlabeled examples are mostly beneficial when classes have small overlap. Our approach encompasses self-learning as a particular case. It was also shown to approach the solution of semi-supervised SVM in another limiting case.

The criterion promotes classifiers with high confidence on the unlabeled examples. The solution is biased regarding posterior probabilities, but the estimation of the decision surface can benefit from unlabeled examples.

Our experiments suggest that the minimum entropy regularization may be a serious contender to generative models in semi-supervised learning. It compared favorably to these models in three situations: for small sample sizes where the generative model cannot completely benefit from the knowledge of the correct joint model; for all sample sizes when the joint distribution was even very slightly misspecified; for all sample sizes also when the unlabeled examples turn out to be non-informative regarding class probabilities.

References

Ambroise, C., Denœux, T., Govaert, G., & Smets, P. (2001). Learning from an imprecise teacher: probabilistic and evidential approaches. *10th International Symposium on Applied Stochastic Models and Data Analysis* (pp. 101–105).

Amini, M. R., & Gallinari, P. (2002). Semi-supervised logistic regression. *15th European Conference on Artificial Intelligence* (pp. 390–394). IOS Press.

Bennett, K. P., & Demiriz, A. (1999). Semi-supervised support vector machines. *Advances in Neural Information Processing Systems 11* (pp. 368–374). MIT Press.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer. 2nd edition.

Brand, M. (1999). Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, *11*, 1155–1182.

Castelli, V., & Cover, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. on Information Theory*, *42*, 2102–2117.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*, 315–332.

Grandvalet, Y. (2002). Logistic regression for partial labels. *9th Information Processing and Management of Uncertainty* (pp. 1935–1941).

Jin, R., & Ghahramani, Z. (2003). Learning with multiple labels. *Advances in Neural Information Processing Systems 15*. MIT Press.

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley.

Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. *9th Int. Conf. on Information and Knowledge Management* (pp. 86–93).

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*, 135–167.

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

Rose, K., Gurewitz, E., & Fox, G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, *11*, 589–594.

Seeger, M. (2002). *Learning with labeled and unlabeled data* (Technical Report). Institute for Adaptive and Neural Computation, University of Edinburgh.

Szummer, M., & Jaakkola, T. S. (2003). Information regularization with partially labeled data. *Advances in Neural Information Processing Systems 15*. MIT Press.

Tong, S., & Koller, D. (2000). Restricted bayes optimal classifiers. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)* (pp. 658–664).

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley.