Semi-Supervised Learning of Mixture Models and Bayesian Networks

Fabio Gagliardi Cozman FGCOZMAN@USP.BR
Escola Politécnica, Univ. of São Paulo, Av. Prof. Mello Moraes, 2231 - Cidade Universitária 05508900, São Paulo, SP - Brazil

Ira Cohen IRACOHEN@IFP.UIUC.EDU
The Beckman Institute, 405 N. Mathews Ave., Urbana, IL 61801

Marcelo César Cirelo MARCELO.CIRELO@POLI.USP.BR
Escola Politécnica, Univ. of São Paulo

Abstract

This paper analyzes the performance of semi-supervised learning of mixture models. We show that unlabeled data can lead to an *increase* in classification error even in situations where additional labeled data would *decrease* classification error. This behavior contradicts several empirical results reported in the literature. We present a mathematical analysis of this "degradation" phenomenon and show that it is due to the fact that bias may be adversely affected by unlabeled data. We study the impact of these theoretical results to classifiers based on Bayesian networks: some situations call for structural learning, while others are best handled by relatively simple classifiers.

1. Introduction

Semi-supervised learning has received considerable attention in the machine learning literature due to its potential in reducing the need for expensive labeled data (Seeger, 2001). Applications such as web search, text classification, genetic research and machine vision are examples where cheap unlabeled data can be added to a pool of labeled samples. The literature seems to hold a rather optimistic view, where "unclassified observations should certainly not be discarded" (O'Neill, 1978). Perhaps the most representative summary of recent literature comes from (McCallum & Nigam, 1998), who declare that "by augmenting this small set [of labeled samples] with a large set of unlabeled data and combining the two pools with EM, we can improve our parameter estimates."

Unfortunately, several experiments indicate that unlabeled data are quite often detrimental to the performance of classifiers (Section 3. That is, the more unlabeled data are used, the poorer is the performance of the resulting classifier. We make this statement cautiously, for some readers may find it obvious, while others may find it unbelievable — and some will dismiss it as incorrect. We have occasionally met with such reactions when communicating our findings. Several machine learning researchers argue that numerical errors in EM or similar algorithms are the natural suspects for such performance degradation. Thus we want to stress that our results concern performance degradation *even in the absence of numerical instabilities*. Other researchers even doubt unlabeled data can be of any use when labeled data are available; these researchers are apparently unconvinced by the testimonials quoted in the previous paragraph. On top of such testimonials, we observe that there are several situations where unlabeled data are *provably* useful (Castelli & Cover, 1996). Yet another group of researchers seems unsurprised that unlabeled data can be deleterious in the presence of modeling errors, arguing that after all anything can be expected in the presence of modeling errors. However, any classifier is an inexact model of reality, and yet labeled data can be almost always useful to classification, *even* in situations where unlabeled data lead to performance degradation. We have made extensive tests with semi-supervised learning, only to witness a complex interaction between modeling assumptions and classifier performance. Unlabeled data do requite a delicate craftsmanship, and we suspect that most researchers are unaware of such complexities. With this paper we wish to contribute to a better understanding of semi-supervised learning.

In Section 4 we show that performance degradation from unlabeled data depends on bias. Our main result is Theorem 1, where we show how semi-supervised learning can be viewed as a convex combination of supervised and unsupervised learning, and how to understand performance degradation in semi-supervised learning. We discuss simple examples that illustrate the kinds of behavior one can see when dealing with semi-supervised learning. We finish by discussing the behavior of Bayesian network classifiers learned with labeled and unlabeled data, indicating situations where unlabeled data do have a beneficial effect.

2. Semi-supervised learning

The goal is to classify an incoming vector of observables $\mathbf{X}$. Each instantiation of $\mathbf{X}$ is a *sample*. There exists a *class variable* C; the values of C are the *classes*. To simplify the discussion, we assume that C is a binary variable with values $\{c', c''\}$. We want to build *classifiers* that receive a sample $\mathbf{x}$ and output either c' or c''. We assume 0-1 loss; thus our objective is to minimize the probability of classification errors. If we knew exactly the joint distribution $F(C, \mathbf{X})$, the optimal rule would be to choose class c' when the probability of $\{C = c'\}$ given $\mathbf{x}$ is larger than $1/2$, and to choose class c'' otherwise. This classification rule attains the minimum possible classification error, called the *Bayes error*.

We take that the probabilities of $(C, \mathbf{X})$, or functions of these probabilities, are estimated from data and then "plugged" into the optimal classification rule. We assume that a parametric model $F(C, \mathbf{X}|\theta)$ is adopted. An estimate of θ is denoted by $\hat{\theta}$. If the distribution $F(C, \mathbf{X})$ belongs to the family $F(C, \mathbf{X}|\theta)$, we say the "model is correct"; otherwise we say the "model is incorrect." When the model is correct, the difference between the expected value $E_\theta[\hat{\theta}]$ and θ, $(E_\theta[\hat{\theta}] - \theta)$, is called *estimation bias*. If the estimation bias is zero, the estimator $\hat{\theta}$ is *unbiased*. When the model is incorrect, we use "bias" loosely to mean the difference between $F(C, \mathbf{X})$ and the $F(C, \mathbf{X}|\hat{\theta})$. The classification error for θ is denoted by $\mathbf{e}(\theta)$; the difference between $E[\mathbf{e}(\hat{\theta})]$ and the Bayes error is the *classification bias*.

We assume throughout that probability models satisfy the conditions adopted by (White, 1982); essentially, parameters belong to compact subsets of Euclidean space, measures have measurable Radon-Nikodym densities and are defined on measurable spaces, and all functions are dominated by integrable functions and differentiable to the first or second order as necessary.

In semi-supervised learning, classifiers are built from a combination of N_l labeled and N_u unlabeled samples. We assume that the samples are independent and ordered so that the first N_l samples are labeled. We consider the following scenario. A sample $(c, \mathbf{x})$ is generated from $p(C, \mathbf{X})$. The value c is then either revealed, and the sample is a *labeled* one; or the value c is hidden, and the sample is an *unlabeled* one. The probability that any sample is labeled, denoted by λ, is fixed, known, and independent of the samples. Thus the same underlying distribution $p(C, \mathbf{X})$ models both labeled and unlabeled data; we do not consider the possibility that labeled and unlabeled samples have different generating mechanisms.

The likelihood of a labeled sample $(c, \mathbf{x})$ is $\lambda p(c, \mathbf{x}|\theta)$; the likelihood of an unlabeled sample $\mathbf{x}$ is $(1 - \lambda)p(\mathbf{x}|\theta)$. The density $p(\mathbf{X}|\theta)$ is a mixture model with *mixing factor* $p(c'|\theta)$ (denoted by η):

$$p(\mathbf{X}|\theta) = \eta p(\mathbf{X}|c', \theta) + (1 - \eta)p(\mathbf{X}|c'', \theta) \,. \qquad (1)$$

We assume throughout that mixtures (1) are identifiable: distinct values of θ determine distinct distributions (permutations of the mixture components are allowed).

The distribution $p(C, \mathbf{X}|\theta)$ can be decomposed either as $p(C|\mathbf{X}, \theta)\, p(\mathbf{X}|\theta)$ or as $p(\mathbf{X}|C, \theta)\, p(C|\theta)$. A parametric model where both $p(\mathbf{X}|C, \theta)$ and $p(C|\theta)$ depend explicitly on θ is referred to as a *generative model*. We adopt the *maximum likelihood* method for estimation of parameters in generative models.

A strategy that departs from the generative scheme is to focus only on $p(C|\mathbf{X}, \theta)$ and to take the marginal $p(\mathbf{X})$ to be independent of θ. Such a strategy produces a *diagnostic model* (for example, logistic regression (Zhang & Oles, 2000)). Attempts to maximize log-likelihood with respect to θ in diagnostic models are not be affected by unlabeled data; in the narrow sense of diagnostic models defined above, maximum likelihood cannot process unlabeled data for *any* given dataset (see (Zhang & Oles, 2000) for a discussion). In this paper we adopt maximum likelihood estimators and generative models; we do not discuss different strategies, such as co-training (Blum & Mitchell, 1998) or active learning (Zhang & Oles, 2000), that can be the object of future work.

3. Do unlabeled data improve or degrade classification performance?

It would perhaps be reasonable to expect an average improvement in classification performance for any increase in the number of samples (labeled or unlabeled): the more are processed, the smaller the variance of estimates, and the smaller the classification error. In Section 4 we show how the limits of this informal argument.

As we have mentioned in Section 1, there are several positive reports in the literature concerning unlabeled data. Investigations in the seventies are quite optimistic (Cooper & Freeman, 1970; Jr., 1973; O'Neill, 1978). More recently, there has been plenty of applied work with semi-supervised learning,[1] with some notable successes. There have also been workshops on semi-supervised learning at NIPS1998, NIPS1999, NIPS2000 and IJCAI2001. Overall, these publications and meetings advance an optimistic view of semi-supervised learning, where unlabeled data can be profitably used whenever available.

Perhaps more important (at least for the sceptical reader)

[1] Relevant references: (Baluja, 1998; Bruce, 2001; Collins & Singer, 2000; Comité et al., 1999; Goldman & Zhou, 2000; McCallum & Nigam, 1998; Miller & Uyar, 1996; Nigam et al., 2000; Shahshahani & Landgrebe, 1994b).

are positive theoretical results concerning unlabeled data. (Castelli & Cover, 1996) and (Ratsaby & Venkatesh, 1995) use unlabeled samples to estimate decision regions (by estimating $p(C, \mathbf{X})$), and labeled samples are used to determine the labels of each region (Ratsaby and Venkatesh refer to this procedure as "Algorithm M"). Castelli and Cover basically prove that Algorithm M is asymptotically optimal under various assumptions, and that, asymptotically, labeled data contribute exponentially faster than unlabeled data to the reduction of classification error. These authors make the critical assumption that $p(C, \mathbf{X})$ belongs to the family of models $p(C, \mathbf{X}|\theta)$ (the "model is correct").

However, a more detailed analysis of current empirical results does reveal some puzzling aspects of unlabeled data.[2] We have reviewed descriptions of performance degradation in the literature in (Cozman & Cohen, 2002); here we just mention the relevant references. Four results are particularly interesting: (Shahshahani & Landgrebe, 1994b) and (Baluja, 1998) describe degradation in image understanding, while (Nigam et al., 2000) report on degradation in text classification and (Bruce, 2001) describe degradation in Baysian network classifiers. Shahshahani and Landgrebe speculate that degradation may be due to deviations from modeling assumptions, such as outliers and "samples of unknown classes" — they even suggest that unlabeled samples should be used only when the labeled data alone produce a poor classifier. Nigam et al suggest several sources of difficulties: numerical problems in the EM algorithm, mismatches between the natural clusters in feature space and the actual labels.

Intrigued by such results, we have conducted extensive tests with simulated problems, and have observed the same pattern of "degradation." The interested reader can again consult (Cozman & Cohen, 2002). Here we present a different test, now with real data. Consider the Adult database that is available in the UCI repository. Figure 1 shows the result of learning a Naive Bayes classifier using different combinations of labeled and unlabeled datasets for the Adult classification problem (using the training and testing datasets available in the UCI repository). We see that adding unlabeled data can improve classification when the labeled data set is small (30 labeled data), but degrade performance as the labeled data set becomes larger.

Both (Shahshahani & Landgrebe, 1994a) and (Nigam, 2001) are rather explicit in stating that unlabeled data can degrade performance, but rather vague in explaining how to analyze the phenomenon. There are several possibilities: numerical errors, mismatches between the distribution of labeled and unlabeled data, incorrect modeling assumptions. Are unlabeled samples harmful only because of numerical instabilities? Is performance degradation caused by increases in variance, or bias, or both? Can performance degradation occur in the absence of bias; that is, when modeling assumptions are correct? Do we need specific types of models, or very complex structures, to produce performance degradation?

[2] The workshop at IJCAI2001 witnessed a great deal of discussion on whether unlabeled data are really useful, as communicated to us by George Forman.

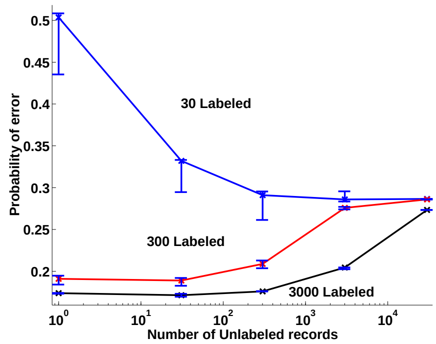

Figure 1. Naive Bayes classifiers generated from the Adult database (bars cover 30 to 70 percentiles).

We propose to study the asymptotic behavior of exact maximum likelihood estimators in semi-supervised learning, adopting the position that this is the best strategy to answer the questions in the last paragraph. The asymptotic results obtained in the next section allows us to analyze semi-supervised learning without resorting to numerical methods, and to obtain insights that are not clouded by the uncertainties of numerical optimization. We do not deny that numerical problems can happen in practice (see (McLachlan & Basford, 1988, Section 3.2) and (Corduneanu & Jaakkola, 2002)), but we are interested in more fundamental phenomena. The examples in the next section show that performance degradation with unlabeled data would occur even if numerical problems were somehow removed.

4. Asymptotics of semi-supervised learning

In this section we discuss the asymptotic behavior of semi-supervised learning. We assume throughout that expectations such as $E[\log p(C, \mathbf{X})]$ and $E[\log p(C, \mathbf{X}|\theta)]$ exist for every θ, and each function attains a maximum at some value of θ in an open neighborhood in the parameter space.

The basic result comes from application of results in (Berk, 1966), (Huber, 1967), and particularly in (White, 1982)). To state the result, a Gaussian density with mean μ and variance σ^2 is denoted by $N(\mu, \sigma^2)$, and the following matrices are defined (matrices are formed by running through the indices i and j): $A_Y(\theta) = E\left[\partial^2 \log p(Y|\theta) / \partial\theta_i\theta_j\right]$, $B_Y(\theta) = E[(\partial \log p(Y|\theta) / \partial\theta_i)(\partial \log p(Y|\theta) / \partial\theta_j)]$. The result we need is as follows. Consider a parametric model $F(Y|\theta)$ satisfying assumptions we have made, and a sequence of maximum likelihood estimates $\hat{\theta}_N$, obtained by maximization of $\sum_{i=1}^{N} \log p(y_i|\theta)$, with an increasing number of independent samples N, all identically distributed according to $F(Y)$. Then $\hat{\theta}_N \to \theta^*$ as $N \to \infty$ for θ in an open neighborhood of θ^*, where θ^* maximizes $E[\log p(Y|\theta)]$. If θ^* is interior to the parameter space, θ^* is a regular point of $A_Y(\theta)$ and $B_Y(\theta^*)$ is non-singular, then $\sqrt{N}\left(\hat{\theta}_N - \theta^*\right) \sim N(0, C(\theta^*))$, where $C_Y(\theta) = A_Y(\theta)^{-1} B_Y(\theta) A_Y(\theta)^{-1}$. This result does not require the distribution $F(Y)$ to belong to the family $F(Y|\theta)$.

Consider now semi-supervised learning. Here the samples are realizations of $(C, \mathbf{X})$ with probability λ, and of $\mathbf{X}$ with probability $(1-\lambda)$. Denote by $\tilde{C}$ a random variable that assumes the same values of C plus the "unlabeled" value u. We have $p(\tilde{C} \neq u) = \lambda$. The actually observed samples are realizations of $(\tilde{C}, \mathbf{X})$, and we obtain $\tilde{p}(\tilde{C} = c, \mathbf{X}) = (\lambda p(C = c, \mathbf{X}))^{I_{\{\tilde{C} \neq u\}}(c)} ((1-\lambda) p(\mathbf{X}))^{I_{\{\tilde{C} = u\}}(c)}$, where $p(\mathbf{X})$ is a mixture density obtained from $p(C, \mathbf{X})$ (Expression (1)) and $I_A(Z)$ is the indicator function (1 if $Z \in A$; Accordingly, the parametric model adopted for $(\tilde{C}, \mathbf{X})$ is: $\tilde{p}(\tilde{C} = c, \mathbf{X}|\theta) = (\lambda p(C = c, \mathbf{X}|\theta))^{I_{\{\tilde{C} \neq u\}}(c)} ((1-\lambda) p(\mathbf{X}|\theta))^{I_{\{\tilde{C} = u\}}(c)}$. Using these definitions, we obtain our main technical result:

Theorem 1 *Consider supervised learning where samples are randomly labeled with probability λ. Adopting previous assumptions, the value of θ^* (the limiting value of maximum likelihood estimates) is:*

$$\arg\max_{\theta} \lambda E[\log p(C, \mathbf{X}|\theta)] + (1-\lambda) E[\log p(\mathbf{X}|\theta)], \quad (2)$$

where the expectations are with respect to $p(C, \mathbf{X})$. □

Proof. The value θ^* maximizes $E[\log \tilde{p}(\tilde{C}, \mathbf{X}|\theta)]$ (expectation with respect to $\tilde{p}(\tilde{C}, \mathbf{X})$), and $E[\log p(\tilde{C}, \mathbf{X}|\theta)]$ is equal to $E[I_{\{\tilde{C} \neq u\}}(\tilde{C}) (\log \lambda + \log p(C, \mathbf{X}|\theta)) + I_{\{\tilde{C} = u\}}(\tilde{C}) (\log(1-\lambda) + \log p(\mathbf{X}|\theta))]$; thus the expected value is equal to $\lambda \log \lambda + (1-\lambda) \log(1-\lambda) + E[I_{\{\tilde{C} \neq u\}}(\tilde{C}) \log p(C, \mathbf{X}|\theta)] + E[I_{\{\tilde{C} = u\}}(\tilde{C}) \log p(\mathbf{X}|\theta)]$. The first two terms of this expression are irrelevant to maximization with respect to θ. The last two terms are equal to $\lambda E[\log p(C, \mathbf{X}|\theta) \, | \tilde{C} \neq u] + (1-\lambda) E[\log p(\mathbf{X}|\theta) \, | \tilde{C} = u]$. As we have $\tilde{p}(\tilde{C}, \mathbf{X}|\tilde{C} \neq u) = p(C, \mathbf{X})$ and $\tilde{p}(\mathbf{X}|\tilde{C} = u) = p(\mathbf{X})$ the last expression is equal to $\lambda E[\log p(C, \mathbf{X}|\theta)] + (1-\lambda) E[\log p(\mathbf{X}|\theta)]$, where the last two expectations are now with respect to $p(C, \mathbf{X})$. Thus we obtain Expression (2). □

Expression (2) indicates that the objective function in semi-supervised learning can be viewed asymptotically as a "convex" combination objective functions for supervised learning ($E[\log p(C, \mathbf{X}|\theta)]$) and for unsupervised learning ($E[\log p(\mathbf{X}|\theta)]$). Denote by θ^*_λ the value of θ that maximizes Expression (2) for a given λ; use θ^*_l for $\theta^*(1)$ and θ^*_u for $\theta^*(0)$.[3] With a few additional assumptions on the modeling densities, Theorem 1 and the implicit function theorem can be used to prove that θ^*_λ is a continuous function of λ (Cozman & Cohen, 2003). This shows that the "path" followed by the solution is a continuous one, as also assumed by (Corduneanu & Jaakkola, 2002) in their discussion of numerical methods for semi-supervised learning.

The asymptotic variance in estimating θ under the conditions of Theorem 1 can also be obtained using results in (White, 1982). The asymptotic variance is ABA, where $A = \left(\lambda A_{(C,\mathbf{X})}(\theta^*) + (1-\lambda) A_{\mathbf{X}}(\theta^*)\right)^{-1}$ and $B = \left(\lambda B_{(C,\mathbf{X})}(\theta^*) + (1-\lambda) B_{\mathbf{X}}(\theta^*)\right)$. It can be seen that this asymptotic covariance matrix is positive definite, so asymptotically an increase in N (the number of labeled and unlabeled samples), leads to a reduction in the variance of $\hat{\theta}$.

Model is correct Suppose first that the family of distributions $F(C, \mathbf{X}|\theta)$ contains the distribution $F(C, \mathbf{X})$; that is, $F(C, \mathbf{X}|\theta_\top) = F(C, \mathbf{X})$ for some $\theta_\top$. When such a condition is satisfied, $\theta^*_l = \theta^*_u = \theta_\top$ given identifiability, and then $\theta^*_\lambda = \theta_\top$ (so maximum likelihood is consistent and bias is zero). Also, we obtain $A(\theta^*_\lambda) = -B(\theta^*_\lambda)$, and then the asymptotic covariance of the maximum likelihood estimator is governed by the inverse of the Fisher information, $\mathbf{I}(\theta_\top)^{-1}$. As we approach an infinitely large number of samples, the classification error should approach the Bayes error. By following a derivation in (Shahshahani & Landgrebe, 1994b) for unbiased estimators, we can argue (approximately) that the expected classification error depends essentially on the variance of $\hat{theta}$. This covariance matrix is asymptotically determined by the *Fisher information* of θ, denoted by $\mathbf{I}(\theta)$. As $\mathbf{I}(\theta)$ is a sum of the information from labeled data and the information from unlabeled data (Zhang & Oles, 2000; Cozman & Cohen, 2003), and because the information from unlabeled data is always positive definite, the conclusion is that *unlabeled data must cause a reduction in classification error when the model is correct.* Similar derivations and conclusions can be found in (Ganesalingam & McLachlan, 1978) and in (Castelli, 1994).

Model is incorrect We now study the scenario that is more relevant to our purposes, where the distribution $F(C, \mathbf{X})$ does not belong to the family of distributions $F(C, \mathbf{X}|\theta)$. In view of Theorem 1, it is perhaps not sur-

[3] We have to handle a diffi culty with $e(\theta^*_u)$: given only unlabeled data, there is no information to decide the labels for decision regions, and the classifi cation error is 1/2 (Castelli, 1994). To simplify the discussion, we assume that, when $\lambda = 0$, an "oracle" will be available to indicate the labels of the decision regions.

prising that unlabeled data can have the deleterious effect discussed in Section 3. Suppose that $\theta_u^* \neq \theta_l^*$ and that $\mathbf{e}(\theta_u^*) > \mathbf{e}(\theta_l^*)$. If we observe a large number of labeled samples, the classification error is approximately $\mathbf{e}(\theta_l^*)$. If we then collect more samples, most of which unlabeled, we eventually reach a point where the classification error approaches $\mathbf{e}(\theta_u^*)$. So, the net result is that we started with classification error close to $\mathbf{e}(\theta_l^*)$, and by adding a great number of unlabeled samples, classification performance degraded. The basic fact here is that estimation and classification bias are affected differently by different values of λ. Hence, a necessary condition for this kind of performance degradation is that $\mathbf{e}(\theta_u^*) \neq \mathbf{e}(\theta_l^*)$; a sufficient condition is that $\mathbf{e}(\theta_u^*) > \mathbf{e}(\theta_l^*)$.

The focus on asymptotics is adequate as we want to eliminate phenomena that can vary from dataset to dataset. If $\mathbf{e}(\theta_l^*)$ is smaller than $\mathbf{e}(\theta_u^*)$, then a large enough labeled dataset can be dwarfed by a much larger unlabeled dataset — the classification error using the whole dataset can be larger than the classification error using only labeled data.

A summary 1) Labeled and unlabeled data contribute to a reduction in variance in semi-supervised learning under maximum likelihood estimation. 2) When the model is correct, the maximum likelihood estimator is unbiased and both labeled and unlabeled data reduce classification error by reducing variance. Unlabeled data alone define the decision regions and labeled data can be used only to label regions, as in Algorithm M. 3) When the model is incorrect, there may be different asymptotic estimation bias for different values of λ; asymptotic classification error may also be different for different values of λ — an increase in the number of unlabeled samples may lead to a larger estimation bias and a larger classification error.

An example: performance degradation with Gaussian data The previous discussion alluded to the possibility that $\mathbf{e}(\theta_u^*) > \mathbf{e}(\theta_l^*)$ when the model is incorrect. To understand how such a phenomenon can occur, consider an example of obvious practical significance. Consider Gaussian observations (X, Y) taken from two classes c' and c''. We know that X and Y are Gaussian variables, and we know their means and variances given the class C. The mean of (X, Y) is $(0, 3/2)$ conditional on $\{C = c'\}$, and $(3/2, 0)$ conditional on $\{C = c''\}$. Variances for X and for Y conditional on C are equal to 1. We do not know, and have to estimate, the mixing factor $\eta = p(C = c')$. The data is sampled from a distribution with mixing factor 3/5.

We want to obtain a Naive-Bayes classifier that can approximate $p(C|X, Y)$. Suppose that X and Y are independent conditional on $\{C = c'\}$ but that X and Y are *dependent* conditional on $\{C = c''\}$ — the correlation $\rho = E[(X - E[X])(Y - E[Y])|C = c'']$ is equal to $4/5$. If we knew the value of ρ, we would obtain an optimal classification boundary on the plane $X \times Y$ (this optimal classification boundary is quadratic). Under the incorrect assumption that $\rho = 0$, the classification boundary is linear: $y = x + 2\log((1 - \hat{\eta})/\hat{\eta})/3$, and consequently it is a decreasing function of $\hat{\eta}$. With labeled data we can easily obtain $\hat{\eta}$ (a sequence of Bernoulli trials); then $\eta_l^* = 3/5$ and the classification boundary is given by $y = x - 0.27031$.

Note that the (linear) boundary obtained with labeled data is not the best possible linear boundary. We can in fact find the best possible linear boundary of the form $y = x + \gamma$. The classification error can be written as a function of γ that has positive second derivative; consequently the function has a single minimum that can be found numerically (the minimizing γ is -0.45786). If we consider the set of lines of the form $y = x + \gamma$, we see that the farther we go from the best line, the larger the classification error. Figure 2 shows the linear boundary obtained with labeled data and the best possible linear boundary. The boundary from labeled data is "above" the best linear boundary.

Now consider the computation of η_u^*, the asymptotic estimate with unlabeled data. By Theorem 1, we must obtain:

$$\arg\max_{\eta\in[0,1]} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ((3/5)N([0, 3/2]^T, \mathrm{diag}[1, 1])+$$
$$(2/5)N([3/2, 0]^T, \left[\begin{smallmatrix} 1 & 4/5 \\ 4/5 & 1 \end{smallmatrix}\right]))\times$$
$$\log(\eta N([0, 3/2]^T, \mathrm{diag}[1, 1])+$$
$$(1 - \eta)N([3/2, 0]^T, \mathrm{diag}[1, 1]))dydx.$$

The second derivative of this double integral is always negative (as can be seen interchanging differentiation with integration), so the function is concave and there is a single maximum. We can search for the zero of the derivative of the double integral with respect to η. We obtain this value numerically, $\eta_u^* \approx 0.54495$. Using this estimate, the linear boundary from unlabeled data is $y = x - 0.12019$. This line is "above" the linear boundary from labeled data, and, given the previous discussion, leads to a larger classification error than the boundary from unlabeled data. We have: $\mathbf{e}(\gamma) = 0.06975$; $\mathbf{e}(\theta_l^*) = 0.07356$; $\mathbf{e}(\theta_u^*) = 0.08141$. The boundary obtained from unlabeled data is also shown in Figure 2.

This example suggests the following situation. Suppose we collect a large number N_l of labeled samples from $p(C, X)$, with $\eta = 3/5$ and $\rho = 4/5$. The labeled estimates form a sequence of Bernoulli trials with probability $3/5$, so the estimates quickly approach η_l^* (the variance of $\hat{\eta}$ decreases as $6/(25N_l)$). If we add a very large amount of unlabeled data to our data, $\hat{\eta}$ approaches η_u^* and the classification error increases.

By changing the "true" mixing factor and the correlation ρ, we can produce examples where the best linear boundary is

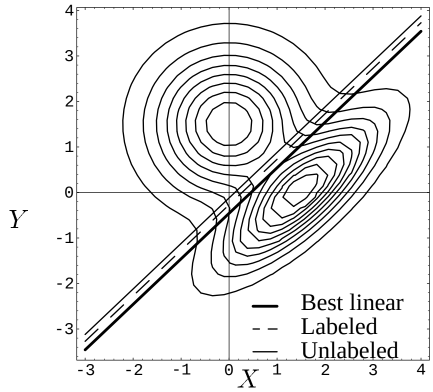

Figure 2. Contour plots of the Gaussian mixture $p(X, Y)$, the best classification boundary of the form $y = x+\gamma$, the linear boundary obtained from labeled data (middle line) and the linear boundary obtained from unlabeled data (upper line).

between the "labeled" and the "unlabeled" boundaries, and examples where the "unlabeled" boundary is between the other two. Different examples of degradation can be easily produced, including examples with univariate models; the interested reader may consult a longer version of this paper (Cozman & Cohen, 2003).

Conclusions The obvious conclusion of the previous results is that unlabeled data can in fact degrade performance even in simple situations. For degradation to occur, modeling errors must be present — unlabeled data are always beneficial in the absence of modeling errors. The most important fact to understand is that *estimation bias depends on the ratio of labeled to unlabeled data*; this is somewhat surprising as bias is usually taken to be a property of the assumed and the "true" models, and not to be dependent on the data. If the performance obtained with the available labeled data is better than the performance with infinitely many unlabeled samples, then at some point the addition of unlabeled data will decrease performance.

5. A brief look at Bayesian network semi-supervised learning

To avoid an excessively negative and theoretical tone, we would like to briefly summarize our experience with semi-supervised learning of Bayesian network classifiers. Bayesian networks form an interesting class of generative classifiers, including Naive-Bayes and TAN classifiers (Friedman et al., 1997).

We have observed that Naive Bayes and TAN classifiers (learned with the EM algorithm) are often plagued by performance degradation, for example in the datasets found in the UCI repository, with TAN classifiers having an edge over Naive Bayes.[4]

At the same time, it is not trivial to find uniformly better ways to handle unlabeled data — it is actually interesting to use our techniques to analyze proposals made in the literature. We briefly discuss here the class of estimators proposed in (Nigam et al., 2000). Nigam et al's estimators maximize a modified log-likelihood of the form $\lambda' L_l(\theta) + (1 - \lambda')L_u(\theta)$ (where L_l is the "likelihood" for labeled data and L_u is the "likelihood" for unlabeled data) while searching for the best possible λ'. There is no reason why this procedure would improve performance, but it may work sometimes: In the Gaussian example in Section 4, if the boundary from labeled data and the boundary from unlabeled data are in different sides of the best linear boundary, then we can find the best linear boundary by changing λ' — we can improve on both supervised and unsupervised learning in such a situation![5] In any case, one cannot expect to find the best possible boundary just by changing λ'; as an example, take the Shuttle dataset from the UCI repository. Using only 100 labeled samples, a Naive Bayes classifier produced classification error of about 18%; with 100 labeled samples and 43400 unlabeled samples, a Naive Bayes learned with the EM algorithm produced classification error of about 70%. The notable fact is that there is a monotonic increase in the classification as we move from fully labeled data ($\lambda' = 1$) to fully unlabeled data ($\lambda' = 0$).

Still, Nigam et al report beneficial effects from unlabeled data, using Naive Bayes classifiers. One explanation is that Naive Bayes is the "correct model" in text classification. A more plausible explanation is that the classifiers built by Nigam et al contain such a large number of observables the variance of estimators is very large for the number of available labeled samples — the reduction in variance offsets increases in bias. We have consistently observed that problems with very large numbers of features and not so large labeled datasets tend to benefit from unlabeled data. Problems in text classification and image understanding typically fit this pattern, and the best results in the literature are exactly in these applications. This agrees with the empirical findings of (Shahshahani & Landgrebe, 1994b), where unlabeled data are useful as more observables are used in classifiers — while Nigam et al suggest that adding observables can worsen the effect of unlabeled data, the opposite should be expected.

What else can be done with unlabeled data? We have actually a variety of new approaches to semi-supervised learn-

[4]The combination of TAN with EM to handle unlabeled data is described in (Meila, 1999).

[5]Some authors have argued that labeled data should be given more weight (Corduneanu & Jaakkola, 2002), but this example shows that there are no guarantees concerning the supposedly superior effect of labeled data.

ing; due to lack of space, we simply state a few observations here, in an attempt to motivate the reader to further pursue the topic.

First we have noticed that feature selection can have an enormous impact on semi-supervised learning: sometimes the removal of a feature "improves the model" and leads to performance improvements with unlabeled data.

Second, we suggest that performance degradation can be used as a "signal" that modeling assumptions are incorrect; we have used techniques that allow us to monitor the classification error and to detect when degradation is statistically significant — indicating the need for modeling changes.

Third, we observe that the most natural way to go beyond Naive Bayes and TAN classifiers is to look for an arbitrary Bayesian network that can represent the relevant distributions; we have had significant success in this direction. Given the many possible approaches for Bayesian network learning, we just mention two interesting ideas. We have developed an stochastic structure search algorithm (named SSS) that essentially performs Metropolis-Hastings runs in the space of Bayesian networks; we have observed that this method, while demanding huge computational effort, can improve on TAN classifiers (algorithm and application to image understanding are described in paper submitted to CVPR2003). We have also developed an algorithm that combines EM iterations and dependency tests, allowing for some restricted forms of feature selection; we have observed performance comparable to SSS (algorithm and results are described in companion paper submitted to ICML2003).

Fourth, we suggest that, when possible, it should be profitable to consider exchanging all unlabeled data in exchange for a few additional labeled samples. It may be better to use a few hundred actively labeled samples than to process thousands of unlabeled samples.

To illustrate the points made in the last two paragraphs, take again the Shuttle dataset from the UCI repository. With 100 labeled samples, classification error is 18%; with 43500 labeled samples, classification error is 0.07% (on independent test set with 14500 labeled samples). Now, consider obtaining a classifier from 100 labeled samples and 30000 unlabeled samples. Naive Bayes leads to classification error of about 70%, and TAN leads to classification error of about 19% (great gains from using more complex structure). SSS does much better, leading to classification error of only 0.02%. Dependency tests produce a classifier with classification error of XXXX. Finally, we could obtain classification error of just YYYY if we discarded the whole unlabeled dataset, selected 100 additional labeled samples *randomly*, and produced a Naive Bayes classifier.

6. Conclusion

In this paper we have derived and studied the asymptotic behavior of semi-supervised learning based on maximum likelihood estimation (Theorem 1), using results from the theory of robust statistics. We have also presented a detailed analysis of performance degradation from unlabeled data, and explained this phenomenon as a consequence of asymptotic bias effects.

In view of the results presented here, several statements in the literature must be properly ammended. Overly optimistic statements concerning semi-supervised learning must be taken in conditional terms. Also, statements that reduce the difference between labeled and unlabeled data to mere labeling of decision regions are incomplete (in particular, Algorithm M would not enjoy a clean motivation in the presence of modeling errors). As discussed in Section 3, there have been previous statements arguing that modeling errors and numerical problems are the causes of performance degradation. We have focused on the quite broad possibility of modeling errors. The term "modeling error" is conveniently vague to allow for almost any kind of behavior, but it is not free of content in the semi-supervised setting. For modeling errors *must* be present for performance degradation to occur. Comments by Nigam and Shahshahani on the effect of modeling errors in semi-supervised learning, discussed in Section 3, however vague, were in the right direction. One of our contributions is to connect in a very precise way modeling errors to performance degradation. The connection, as we have argued, comes from an understanding of asymptotic bias.

Despite these sobering comments, we note that our techniques can lead to better semi-supervised classifiers in a variety of situations, as argued in Section 5.

We have on purpose not dealt with two types of modeling errors. First, we have avoided the possibility that labeled and unlabeled data are sampled from different distributions. Such a form of selection bias is quite serious and can obviously have a deleterious effect on classification error (McLachlan, 1992, pages 42-43). Second, we have avoided the possibility that more classes are represented in the unlabeled data than in the labeled data, perhaps due to the scarcity of labeled samples ((Nigam et al., 2000) discuss techniques to address this issue). We believe that, by constraining ourselves to simpler modeling errors, we have forcefully indicated that performance degradation must be prevalent in practice.

The list of possible extensions of the current work is long and reflects the richness of the subject. It should be interesting to find necessary and sufficient conditions for a model to suffer performance degradation with unlabeled data. Also, the analysis of bias should be much enlarged, with the addition of finite sample results. Another possible avenue is to look for optimal estimators in the presence of modeling errors (Kharin, 1996). Finally, it would be important to investigate performance degradation in other frameworks, such as support vector machines, co-training, or entropy based solutions (Jaakkola et al., 1999). We conjecture that any approach that incorporates unlabeled data (so as to improve performance when the model is correct) may suffer from performance degradation when the model is incorrect. We note that co-training results in the literature seem to corroborate this hypothesis (Ghani, 2001, Hoovers-255 dataset). If we could in fact find an universally robust semi-supervised learning method, such a method would indeed be a major accomplishment.

Regardless of the approach that is used, semi-supervised learning is affected by modeling assumptions in rather complex ways. The present paper should be helpful as a first step in understanding unlabeled data and their peculiarities in machine learning.

Acknowledgements

This work has received continued and substantial support from HP Labs. We thank Alex Bronstein and Marsha Duro for proposing the research on labeled-unlabeled data and for many suggestions and comments during the course of the work, as their help was critical to the results described here. We thank Tom Huang for substantial support to this research; Moises Goldszmidt for suggesting important improvements; George Forman for telling us about an IJCAI workshop; Kevin Murphy for the freely available BNT system; Marina Meila for useful comments in a preliminary version; Íon Muslea for kind remarks on initial results.

References

Baluja, S. (1998). Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *NIPS*.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Math. Statistics*, 51–58.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT*.

Bruce, R. (2001). Semi-supervised learning using prior probabilities and EM. *IJCAI Workshop on Text Learning*.

Castelli, V. (1994). *The relative value of labeled and unlabeled samples in pattern recognition*. Doctoral dissertation, Stanford University.

Castelli, V., & Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters*, *16*, 105–111.

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.

Collins, M., & Singer, Y. (2000). Unupervised models for named entity classification. *ICML* (pp. 327–334).

Comité, F. D., Denis, F., Gilleron, R., & Letouzey, F. (1999). Positive and unlabeled examples help learning. *Int. Conf. on Algorithmic Learning Theory* (pp. 219–230). Springer-Verlag.

Cooper, D. B., & Freeman, J. H. (1970). On the asymptotic improvement in the outcome of supervised learning provided by additional nonsupervised learning. *IEEE Trans. on Computers*, *C-19*, 1055–1063.

Corduneanu, A., & Jaakkola, T. (2002). Continuations methods for mixing heterogeneous sources. *UAI* (pp. 111–118). San Francisco, California: Morgan-Kaufmann.

Cozman, F. G., & Cohen, I. (2002). Unlabeled data can degrade classification performance of generative classifiers. *FLAIRS* (pp. 327–331). Pensacola, Florida.

Cozman, F. G., & Cohen, I. (2003). The Effect of Modeling Errors in Semi-Supervised Learning of Mixture Models: How Unlabeled Data Can Degrade Performance of Generative Classifiers, at `http://www.poli.usp.br/p/fabio.cozman/Publications/lul.ps.gz`.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Ganesalingam, S., & McLachlan, G. J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, *65*.

Meila, M. (1999). *Learning with mixtures of trees*. Doctoral dissertation, MIT.

Ghani, R. (2001). Combining labeled and unlabeled data for text classification with a large number of categories. *IEEE Int. Conf. on Data Mining*.

Goldman, S., & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. *Int. Joint Conf. on Machine Learning*.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Fifth Berkeley Symposium in Mathematical Statistics and Probability*, 221–233.

Jaakkola, T. S., Meila, M., & Jebara, T. (1999). Maximum entropy discrimination. *NIPS 12*.

Hosmer Jr., D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, *29*, 761–770.

Kharin, Y. (1996). *Robustness in statistical pattern recognition*. Kluver Academic Publishers.

McCallum, A., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. *ICML* (pp. 359–367).

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: John Wiley and Sons Inc.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: inference and applications to clustering*. New York: Marcel Dekker Inc.

Miller, D. J., & Uyar, H. S. (1996). A mixture of experts classifier with learning based on both labelled and unlabelled data. In *NIPS* 571–577.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–144.

Nigam, K. P. (2001). *Using unlabeled data to improve text classification* (Technical Report CMU-CS-01-126). School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

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

Ratsaby, J., & Venkatesh, S. S. (1995). Learning from a mixture of labeled and unlabeled examples with parametric side information. *COLT* (pp. 412–417).

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*, 195–239.

Seeger, M. (2001). *Learning with labeled and unlabeled data* (Technical Report). Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, United Kingdom.

Shahshahani, B. M., & Landgrebe, D. A. (1994a). *Classification of multi-spectral data by joint supervised-unsupervised learning* (Technical Report TR-EE 94-1). School of Electrical Engineering, Purdue University, West Lafayette, Indiana.

Shahshahani, B. M., & Landgrebe, D. A. (1994b). The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Trans. on Geoscience and Remote Sensing*, *32*, 1087–1095.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Zhang, T., & Oles, F. (2000). A probability analysis on the value of unlabeled data for classification problems. *Int. Joint Conf. on Machine Learning* (pp. 1191–1198).