Semi-Supervised Learning

Tobias Scheffer
Humboldt-Universität zu Berlin, Germany

INTRODUCTION

For many classification problems, unlabeled training data are inexpensive and readily available, whereas labeling training data imposes costs. Semi-supervised classification algorithms aim at utilizing information contained in unlabeled data in addition to the (few) labeled data.

Semi-supervised (for an example, see Seeger, 2001) has a long tradition in statistics (Cooper & Freeman, 1970); much early work has focused on Bayesian discrimination of Gaussians. The Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is the most popular method for learning generative models from labeled and unlabeled data. Model-based, generative learning algorithms find model parameters (e.g., the parameters of a Gaussian mixture model) that best explain the available labeled and unlabeled data, and they derive the discriminating classification hypothesis from this model.

In discriminative learning, unlabeled data is typically incorporated via the integration of some model assumption into the discriminative framework (Miller & Uyar, 1997; Titterington, Smith, & Makov, 1985). The Transductive Support Vector Machine (Vapnik, 1998; Joachims, 1999) uses unlabeled data to identify a hyperplane that has a large distance not only from the labeled data but also from all unlabeled data. This identification results in a bias toward placing the hyperplane in regions of low density $p(x)$. Recently, studies have covered graph-based approaches that rely on the assumption that neighboring instances are more likely to belong to the same class than remote instances (Blum & Chawla, 2001).

A distinct approach to utilizing unlabeled data has been proposed by de Sa (1994), Yarowsky (1995) and Blum and Mitchell (1998). When the available attributes can be split into *independent* and *compatible* subsets, then *multi-view learning* algorithms can be employed. Multi-view algorithms, such as co-training (Blum & Mitchell, 1998) and co-EM (Nigam & Ghani, 2000), learn two independent hypotheses, which bootstrap by providing each other with labels for the unlabeled data.

An analysis of why training two independent hypotheses that provide each other with conjectured class labels for unlabeled data might be better than EM-like self-training has been provided by Dasgupta, Littman, and McAllester (2001) and has been simplified by Abney (2002). The disagreement rate of two independent hypotheses is an upper bound on the error rate of either hypothesis. Multi-view algorithms minimize the disagreement rate between the peer hypotheses (a situation that is most apparent for the algorithm of Collins & Singer, 1999) and thereby the error rate.

Semi-supervised learning is related to active learning. *Active learning* algorithms are able to actively query the class labels of unlabeled data. By contrast, semi-supervised algorithms are bound to learn from the given data.

BACKGROUND

Semi-supervised classification algorithms receive both labeled data $D_l = (x_1, y_1),..., (x_{m_l}, y_{m_l})$ and unlabeled data $D_u = x^u_1,..., x^u_{m_u}$ and return a classifier $f: x \rightarrow y$; the unlabeled data is generally assumed to be governed by an underlying distribution $p(x)$, and the labeled data by $p(x,y) = p(y \mid x)\ p(x)$. Typically, the goal is to find a classifierthat minimizes the error rate with respect to $p(x)$.

In the following sections, we distinguish between *model-based* approaches, mixtures of *model-based and discriminative* techniques, and *multi-view learning*. Model-based approaches can directly utilize unlabeled data to estimate $p(x, y)$ more accurately. Discriminative classification techniques need to be augmented with some model-based component to make effective use of unlabeled data. Multi-view learning can be applied

when the attributes can be split into two independent and compatible subsets.

Model-Based Semi-Supervised Classification

Model-based classification algorithms assume that the data be generated by a parametric mixture model $p(x, y \mid \Theta)$and that each mixture component contains only data belonging to a single class. Under this assumption, in principle, only one labeled example per mixture component is required (in addition to unlabeled data) to learn an accurate classifier. Estimating the parameter vector Θ from the data leads to a *generative* model; that is, the model $p(x, y \mid \Theta)$ can be used to draw new labeled data.

In the context of classification, the main purpose of the model is *discrimination.* Given the model parameter, the corresponding classifier is $f_\Theta(x)= \arg\max_y p(x, y \mid \Theta)$. For instance, when $p(x, y \mid \Theta)$ is a mixture of Gaussians with equal covariance matrices, then the discriminator $f_\Theta(x)$ is a linear function; in the general Gaussian case, $f_\Theta(x)$ is a second-order polynomial. The Expectation Maximization (EM) algorithm (Dempster et al., 1977) provides a general framework for semi-supervised model-based learning — that is, for finding model parameters Θ. Semi-supervised learning with EM is sketched in Table 1; after initializing the model by learning from the labeled data, it iterates two steps. In the E-step, the algorithm calculates the class probabilities for the unlabeled data based on the current model. In the M-step, the algorithm estimates a new set of model parameters from the labeled and the originally unlabeled data for which probabilistic labels have been estimated in the E-step.

The EM algorithm, which is a greedy method for maximizing the likelihood $p(D_l, D_u \mid \Theta) = p(D_l \mid \Theta)\, p(D_u \mid \Theta)$ of the data, has three caveats. The first is that no obvious connection exists between the *maximum likelihood* model parameters Θ and the Bayesian discriminator that minimizes the conditional risk given a new instance x. Practical semi-supervised learning algorithms apply some form of regularization to approximate the *maximum a posteriori* rather than the *maximum likelihood* parameters. The second caveat is that the resulting parameters are a *local* but not necessarily the *global* maximum. The third caveat of semi-supervised learning with EM is more subtle: When the assumed parametric model is *correct* — that is, the data has, in fact, been generated by $p(x, y \mid \Theta)$for some Θ— then the idea is arguable that unlabeled data will improve the accuracy of the resulting classifier $f_\Theta(x)$ under fairly reasonable assumptions (Zhang & Oles, 2000; Cozman, Cohen, & Cirelo, 2003). However, as Cozman et al. have pointed out, the situation is different when the model assumption is *incorrect* — that is, no Θ exists such that $p(x, y \mid \Theta)$ equals the true probability $p(x, y)$, which governs the data. In this case, the best approximation to the labeled data — $\Theta_l = \arg\max_\Theta p(D_l \mid \Theta)$ — can be a much better classifier f_{Θ_l}than $f_\Theta(x)$ with $\Theta= \arg\max_\Theta p(D_l, D_u \mid \Theta)$, which approximates the labeled and unlabeled data. In other words, when the model assumption is incorrect, then semi-supervised learning with EM can generally result in poorer classifiers than supervised learning from only the labeled data.

Semi-supervised learning with EM has been employed with many underlying models and for many applications, including mixtures of Gaussians and naïve Bayesian text classification (Nigam, McCallum, Thrun, & Mitchell, 2000).

Table 1. Semi-supervised classification with EM.

Input: labeled data $D_l = (x_1, y_1),..., (x_{m_l}, y_{m_l})$; unlabeled $D_u = x^u_1,..., x^u_{m_u}$.
Initialize model parameters Θ by learning from the labeled data.
Repeat until a local optimum of the likelihood $p(x, y \mid \Theta)$ is reached.
E-step: For all unlabeled data x^u_i and class labels y, calculate
$E(f(x^u_1) = y \mid \Theta)$, the expected probability that y is the class of x^u_1 given Θ; that is, use $p(y \mid x, \Theta)$ to probabilistically label the x^u_i.
M-step: Calculate the maximum likelihood parameters
Θ = arg max $p(D_l, D_u$ estimated class probabilities for $Du)$; that is, learn from the labeled and probabilistically labeled unlabeled data.
Return classifier $p(y \mid x, \Theta)$.