Learning from Partially Labeled Data

Martin Szummer

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Learning from data with both labeled training points (x,y pairs) and unlabeled training points (x alone). For the labeled points, supervised learning techniques apply, but they cannot take advantage of the unlabeled points. On the other hand, unsupervised techniques can model the unlabeled data distribution, but do not exploit the labels. Thus, this task falls between traditional supervised and unsupervised learning.

Motivation: Supervised learning performance improves with larger training data sets. Unfortunately, it is often infeasible to obtain labels for large training sets. Assigning labels can require expensive resources such as human labor or laboratory tests. In some cases ground truth labels are impossible to obtain, e.g. if the necessary measurements can no longer be made, or if the labels will be given only in the future. In contrast, unlabeled training data is frequently easy to obtain in large quantities, and can outnumber the amount of labeled data by a large factor. For example, it is expensive to collect image databases of only faces, but it is cheap to collect arbitrary imagery with occasional faces, e.g. by crawling the world wide web, or by pointing a video camera out the window.

There are also developmental motivations for studying the process of learning from partially data. Children acquire language mainly by listening and imitating, with very limited feedback from adults. Human beings also excel at other partially labeled learning tasks, such as visual discrimination with hyperacuity [1].

Previous Work: Learning from partially labeled data is not well understood from the theoretical perspective. Labeled data has been shown to be exponentially more useful than unlabeled data under certain assumptions [2]. Discriminative and generative learning architectures have been shown to take advantage of unlabeled data in different settings [3].

Several practical algorithms for partially labeled data have been proposed. The EM algorithm is the most widely used. EM iterates between estimating model parameters and inferring soft labels for unlabeled points. It assumes a generative model (typically a mixture of Gaussians) and has been applied to text classification [4] and face pose determination [5]. Unfortunately, when the data does not match EM's generative assumptions, the algorithm goes astray, and the information from labeled data is overwhelmed by unlabeled points.

Co-training [4] classifies data that exhibits factorial structure, specifically when its attributes can be partitioned into groups that are individually sufficient for learning but mutually independent. Co-training works by feeding outputs of one learner to be fed as examples for the other and vice-versa. However, in practice the individual learners are noisy, and then co-training easily veers down the wrong path.

Transduction [6] is a technique that attempts to maximize the classification margin on both labeled and unlabeled data, while classifying the labeled data as correctly as possible (Figure 1). This discriminative method imposes fewer restrictions on the data model. However, finding the optimal decision boundary requires solving a mixed integer programming problem that is NP-complete. The maximum entropy discrimination framework [7] also optimizes the margin based on both labeled and unlabeled data. However, instead of setting the margin directly, a distribution over margins is optimized to lie as close as possible to a prior. Likewise, labels of points are optimized to lie as close as possible to another prior; priors for labeled points peak around 0 or 1, whereas priors on unlabeled points are peaked at 0.5.

Approach: I am studying the problem of learning from partially labeled data from a theoretical and a practical perspective.

1. For the theoretical part I am characterizing the mechanisms through which unlabeled data can and cannot be used. For example, discriminative techniques that factor the distribution as $p(x, y|\alpha) = p(x)p(y|x, \alpha)$ cannot

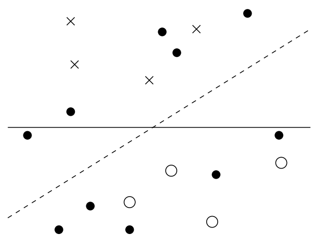

Figure 1: The decision boundary (solid line) based on labeled data (circles, crosses) has a smaller margin on all data than a boundary (dashed line) based on both labeled and unlabeled points (dots).

directly use the unlabeled data, because an improved knowledge of $p(x)$ does not help in estimating the model parameters α. On the other hand, models that assume $p(x, y|\alpha) = p(x|\alpha)p(y|x, \alpha)$ can immediately take advantage of partially labeled data.

At the same time, I am also characterizing the mechanisms through which existing algorithms use partially labeled data. An improved understanding of these algorithms enables predictions on what data sets they will work. I am also proposing new algorithms that better exploit partially labeled data [8].

2. For the practical part, I am implementing the main existing algorithms (maximum relative entropy, EM, transduction); this requires resolving the difficulties described above. I am benchmarking the algorithms on real applications, and describing their strengths and weaknesses. Application areas include computer vision (face detection and image classification), text classification and bioinformatics.

Difficulty: Existing techniques suffer from various limitations, including restrictive probabilistic assumptions, inability to deal with noisy data, being overwhelmed by unlabeled data, and computational intractability. Moreover, the properties of these algorithms are poorly understood.

Impact: Algorithms that learn from partially labeled data will be useful in all fields where unlabeled data is abundant but labels are difficult to obtain.

Future Work: I will address outstanding difficulties resulting from challenging partially labeled data problems. Examples include learning metrics, learning from very unbalanced classes, and different distributions for labeled and unlabeled data.

Research Support: Research at CBCL is supported by ONR, Darpa, NSF, Kodak, Siemens, Daimler, ATR, ATT, Compaq, Honda and CRIEPI.

References:

[1] M. Fahle, S. Edelman and T. Poggio. Fast Perceptual Learning in Visual Hyperacuity. *Vision Research*, vol. 35, pp. 3003–3013, 1995.

[2] V. Castelli and T. Cover. The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Paramter. *IEEE Trans. Information Theory*, vol. 42, pp. 2102–2117, Nov, 1996.

[3] T. Zhang and F. Oles. A Probability Analysis on the Value of Unlabeled Data for Classification Problems. *International Conference on Machine Learning*; June, 2000.

[4] K. Nigam, A. McCallum, S. Thrun, T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, vol. 39(2/3), pp. 103–134, 2000.

[5] S. Baluja. Probabilistic Modeling for Face Orientation Discrimination: Learning from Labeled and Unlabeled Data. *Neural Information Processing Sys.*, 1998, pp. 854–860.

[6] V.N. Vapnik. Statistical Learning Theory. Wiley, 1998.

[7] T. Jaakkola, M. Meila and T. Jebara. Maximum Entropy Discrimination. *Neural Information Processing Sys.*, vol. 12, 1999, pp. 470–476.

[8] M. Szummer and T. Jaakkola. Kernel Expansions with Unlabeled Data. to appear, *Neural Information Processing Sys.*, vol. 13, 2000