SEMI-SUPERVISED LEARNING FOR BAYESIAN PATTERN CLASSIFICATION

Julian L. Center, Jr.
Creative Research Corp.
385 High Plain Rd., Andover, MA 01810, USA
jcenter@world.std.com

2 May 2005

Abstract

A Bayesian pattern classification algorithm assesses a situation and determines the probability that a particular class label c applies. The algorithm makes this assessment based on an observed attribute vector $\mathbf{x}$. To tune the algorithm to a specific operational environment, we must have a set of labeled training data, consisting of attribute-label pairs that sample the conditional probability of the class label given the attribute vector, $p(c|\mathbf{x})$. We may also have access to unlabeled data, consisting of a collection of attribute vectors whose distribution is representative of $p(\mathbf{x})$.

In many practical situations, labeled training data is scarce and difficult or expensive to obtain, but unlabeled training data is cheap and abundant. In the past few years, several researchers have studied the problem of combining both labeled and unlabeled data to achieve what is called semi-supervised learning.

It is natural to assume that unlabeled data can be helpful in building the classification algorithm because we feel intuitively that data points clustered together should all get the same class label. This intuition can be transferred into the design of a Bayesian classification algorithm by the proper choice of a family of probability models.

In this paper, we present different probability model families that correspond to several of the semi-supervised learning methods discussed in the literature. All of these models fit into a common framework. Some are based on mixture models; others utilize Gaussian process models, and some use a combination of the two. We compare the performance of these different model types on a simple, but illuminating example. We discuss their relative merits and make general recommendations concerning the applicability of the different model types.

Key Words: Semi-Supervised Learning, Labeled and Unlabeled Data