The relative value of labeled and unlabeled samples in pattern recognition *

Vittorio Castelli
Stanford University

Thomas M. Cover
Stanford University

Abstract

We attempt to discover the role and relative value of labeled and unlabeled samples in reducing the probability of error of the classification of a sample based on the previous observation of labeled and unlabeled data. We assume that the underlying densities belong to a regular family that generates identifiable mixtures.

The unlabeled observations, under the above conditions, carry information about the statistical model and therefore can be effectively used to construct a decision rule. When the training set contains an infinite number of unlabeled samples, the first labeled observation reduces the probability of error to within a factor of two of the Bayes risk. Moreover subsequent labeled samples yield exponential convergence of the probability of classification error to the Bayes risk. We argue that labeled samples are exponentially more valuable than unlabeled samples and identify the exponent as the Bhatthacharyya distance.

Summary

Assume we sample from two populations, $\theta = 1$ and $\theta = 2$, with prior probabilities η and $\overline{\eta} = 1 - \eta$. Let observations from population 1 be distributed according to density $f_1(x)$, with respect to some measure μ, and observations from population 2 according to $f_2(x)$. We observe l independent samples together with their classifications, $\{(X_1, \theta_1), .., (X_l, \theta_l)\}$, where the θ_i are Bernoulli(η) and the X_i are i.i.d. $\sim f_{\theta_i}(x)$, and we observe u unlabeled samples $\{X'_1, .., X'_u\}$. The totality constitutes the training set.

Let X be a sample, similarly drawn, which we wish to classify with minimum probability of error. Let $R(l, u)$ be the probability of error of a given decision rule when the training set is composed of l labeled and u unlabeled samples.

If $f_1(x)$, $f_2(x)$ and η are known, the likelihood ratio test

$$\text{Decide} \quad \hat{\theta}(X) = \begin{cases} 1 & \text{if } \dfrac{\eta f_1(X)}{(1-\eta) f_2(X)} \geq 1 \\ 2 & \text{if } \dfrac{\eta f_1(X)}{(1-\eta) f_2(X)} < 1, \end{cases}$$

minimizes the probability of error (Bayes risk) which is equal to

$$R^* = \Pr\left\{\hat{\theta} \neq \theta\right\} = E_\mu\left[\min\left(\eta f_1(x), (1-\eta) f_2(x)\right)\right].$$

If $f_1(x)$ and $f_2(x)$ belong to a regular family $\mathcal{F}$, the distributions and the prior probabilities may be estimated from the labeled data. If an infinite number of labeled samples is available, the risk is given by $R(\infty, u) = R^*$ for any number u of unlabeled samples.

The distributions and prior probabilities can also be estimated using the unlabeled observations, under the additional hypothesis that the family of mixtures [1] generated by $\mathcal{F}$ is identifiable [2]. For example, $\mathcal{F}$ can be the family of Gaussian distributions with mean μ and variance σ^2. Then any mixture of the form $\eta\varphi(\mu_1, \sigma_1^2) + \overline{\eta}\varphi(\mu_2, \sigma_2^2)$ can be uniquely decomposed into its component densities.

Thus $u = \infty$ yields the information that the underlying distributions are either $(\eta f(x), \overline{\eta} g(x))$ or $(\overline{\eta} g(x), \eta f(x))$, but no information is available on whether $f_1(x) = f(x)$ or $f_1(x) = g(x)$. Thus for $l = 0$ labeled samples,

$$R(0, u) = \frac{1}{2} \qquad \text{for all } u.$$

Labeled data are therefore needed. The first labeled sample helps enormously.

Theorem. *When the training set contains an infinite number of unlabeled samples, the first labeled observation yields a probability of error*

$$R(1, \infty) = 2R^*(1 - R^*)$$

for the classification of a new sample.

The expected probability of classification error is thus reduced to within a factor two of the Bayes risk.

Theorem. *When the number of unlabeled samples is infinite, the risk converges to the Bayes risk exponentially fast, i.e.*

$$R(l, \infty) = R^* + O\left(e^{-l\alpha}\right)$$

where $\alpha = -\log\left(\int 2\sqrt{\eta\overline{\eta}}\sqrt{f_1(x) f_2(x)}\, d\mu(x)\right)$.

Labeled samples can reduce the risk exponentially fast, but unlabeled samples reduce the risk only polynomially fast. Under smoothness conditions on the family $\mathcal{F}$, similar to those that allow efficient estimation of parameters [3], there exists a procedure such that

$$R(l, u) = R^* + O(1/u) + O(e^{-l\alpha}).$$

Roughly speaking, labeled samples are exponentially more valuable than unlabeled samples.

References

[1] Teicher, Henry. "On the mixtures of distributions" *Ann. Math. Statist.* 1960, **32** 244-248.

[2] Teicher, Henry. "Identifiability of finite mixtures" *Ann. Math. Statist.* 1963, **34** 1265-1269.

[3] Lehmann E.L. *Theory of point estimation* 1983, John Wiley and Sons, New York.

[4] Cover T.M., Thomas J.A. *Elements of Information Theory.* 1991, John Wiley and Sons, New York.

[5] Duda R.O., HartP.E. *Pattern Classification and Scene Analysis.* 1973, John Wiley and Sons, New York.

[6] Andrews H.C. *Introduction to Mathematical Techniques in Pattern Recognition.* 1972, John Wiley and Sons, New York.

*This work was partially supported by NFS Grant NCR-8914538-02.
Vittorio Castelli (vittorio@isl.stanford.edu) and Thomas M. Cover (cover@isl.-stanford.edu) are with the Information Systems Laboratory, Department of Electrical Engineering, 4055 Stanford University, Stanford, California 94305.