Pattern Recognition 45 (2012) 2308–2320

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Optimal mean-square-error calibration of classifier error estimators under Bayesian models

Lori A. Dalton [a,*], Edward R. Dougherty [a,b,c]

[a] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
[b] Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, USA
[c] Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Article history:
Received 20 October 2011
Received in revised form
23 November 2011
Accepted 5 December 2011
Available online 14 December 2011

Keywords:
Bayesian estimation
Classification
Error estimation
Genomics
Minimum mean-square estimation
Small samples

ABSTRACT

A recently proposed Bayesian modeling framework for classification facilitates both the analysis and optimization of error estimation performance. The Bayesian error estimator is then defined to have optimal mean-square error performance, but in many situations closed-form representations are unavailable and approximations may not be feasible. To address this, we present a method to optimally calibrate arbitrary error estimators for minimum mean-square error performance within a supposed Bayesian framework. Assuming a fixed sample size, classification rule and error estimation rule, as well as a fixed Bayesian model, the calibration is done by first computing a calibration function that maps error estimates to their optimally calibrated values off-line. Once found, this calibration function may be easily applied to error estimates on the fly whenever the assumptions apply. We demonstrate that calibrated error estimators offer significant improvement in performance relative to classical error estimators under Bayesian models with both linear and non-linear classification rules.

1. Introduction

When a large sample is available, classification and error estimation are easy because the data may be partitioned into training and testing datasets without significantly degrading the quality of the classifier or the accuracy of the error estimator. In biomedicine and other applications restricted to a small-sample setting, classifier error estimation is a critical issue since it is the primary measure of the scientific validity of a designed classifier and small-sample error estimation is problematic. To address the issue, a recently proposed Bayesian framework for classification defines a mathematical foundation for both the analysis and optimization of error estimation schemes. In essence, the Bayesian framework parameterizes the underlying feature-label distribution and assigns prior distributions to these parameters. Given a sample, priors are updated to posterior distributions over the model parameters, which quantifies our knowledge of the feature-label distribution from the prior and sample. Within this framework, Bayesian error estimators are defined to have optimal performance for a fixed sample and designed classifier relative to the posterior distributions. Performance is measured with respect to the mean-square error (MSE), which is the expected squared deviation from the true error, or the root-mean-square (RMS), which is the square root of the MSE. Analytical solutions for Bayesian error estimators have been provided for two cases: discrete distributions with Dirichlet priors and arbitrary classification (henceforth referred to as the discrete model) [1] and Gaussian distributions with normal-inverse-Wishart priors and linear classification (the Gaussian model) [2].

When it is reasonable to assume a Bayesian framework but an analytical or closed-form Bayesian error estimator is not available, it may be approximated using Monte–Carlo methods. For instance, source code for Gaussian models with non-linear classification is available in [3]. That being said, approximating a Bayesian error estimator is much more computationally intensive than classical counting methods and may be infeasible. To address this, we propose a new method of optimally calibrating arbitrary error estimators within Bayesian frameworks. Assuming a fixed sample size, fixed classification and error estimation schemes, and a set of priors for the distribution parameters, this is done in two steps. First, we compute a calibration function mapping error estimates (from the specified error estimation rule) to their calibrated values off-line according to the assumed model. Second, in all future experiments a practitioner may perform classification and error estimation in the usual way, but at the last step use the calibration function as a simple lookup table to calibrate the final error estimate on the fly.

* Corresponding author. Tel.: +1 979 862 8896.
E-mail addresses: ldalton@tamu.edu (L.A. Dalton), edward@ece.tamu.edu (E.R. Dougherty).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.12.003

The calibration function is defined to be the minimum mean-square error (MMSE) estimate of the true error of a classifier designed from the assumed classification scheme, given an observed error estimate. Equivalently, this is the expected true error conditioned on the observed error estimate, where uncertainty in the expectation stems from our uncertainty in both the feature-label distribution and the sample. This is similar to Bayesian MMSE error estimation itself, which is equivalent to the expected true error of a designed classifier conditioned on the entire observed sample, except that the calibrated error estimator conditions on only the observed error estimate. In other words, both error estimators minimize MSE in the same assumed Bayesian model, but the Bayesian error estimator has the benefit of the entire sample, which is an array of n sample points with D features each, and the MMSE calibrated error estimator uses only a single statistic (a lossy function of the observed sample) containing information about the true error. Also, a basic property of both Bayesian and calibrated error estimators is that they are unbiased relative to the true error. However, since the MMSE calibrated error estimate averages true errors over all samples producing the observed error estimate, the sample and classifier are not fixed as they are in Bayesian error estimation, where conditioning is on the sample itself.

Classical error estimation analyses evaluate performance over a sampling distribution for a fixed feature-label distribution, which is typically unknown in practice. There, the joint density between true and estimated errors contains the full information about performance, though only a few cases have been solved analytically. For linear-discriminant analysis (LDA), exact joint distributions have been found for both resubstitution and leave-one-out in fixed univariate Gaussian models, and approximate joint distributions are also available in fixed multivariate models with a common known covariance matrix [4]. In contrast, a Bayesian framework considers performance over an entire family of feature-label distributions, where uncertainty in the true distribution is conditioned precisely on the observed data available in hand. Hence, the calibration scheme proposed here stands out as a method to optimally calibrate classical error estimation schemes, not for a fixed distribution but over all distributions in an uncertainty class, with a higher weight assigned to distributions that are more probable given the observed error estimate.

Related work by Xu et al. [5] considers the expected true error conditioned on an error estimate when the feature-label distribution is modeled as Gaussian or mixed-Gaussian with fixed means and scalable covariance matrices. There, the Bayes error of a feature-label distribution is assigned a beta prior scaled between 0 and 0.25, indirectly corresponding to a distribution on the scale for the covariances used in the model. Here, we place prior distributions directly on parameters of the feature-label distribution itself, which is a more fundamental state of nature in the problem. Furthermore, the Bayesian framework utilized here is founded on deeper theory, including analytical representations of the MSE performance for arbitrary error estimators conditioned on the sample and the consistency of Bayesian error estimation in both the discrete and Gaussian models [6].

2. Review of Bayesian error estimation

Consider a binary classification problem with class labels 0 and 1, and a sample, S_n, with n sample points in a sample space $\mathcal{X}$. Let n_0 and n_1 denote the number of sample points in class 0 and class 1, respectively. The observed sample is used to train a classifier, $\psi_n : \mathcal{X} \to \{0,1\}$. The true error of ψ_n can be decomposed as,

$$\varepsilon_n = c\varepsilon_n^0 + (1-c)\varepsilon_n^1, \tag{1}$$

where c is the *a priori* probability that a sample point is from class 0, ε^0 is the probability that the classifier mislabels a class 0 point, and ε^1 is the probability that the classifier mislabels a class 1 point.

In practice, the feature-label distribution is unknown, so that the true error must be estimated via error estimation rules. Classical training data error estimation methods, such as cross-validation [7,8] and bootstrap [9,10], are typically counting methods that are "model-free", in the sense that their evaluation does not utilize modeling assumptions. Bolstered error estimation [11] is a smoothed counting method which associates a bolstering kernel with each sample point to spread its mass so that each point contributes to the bolstered error estimate based on its distance from the classifier decision boundary. Bayesian error estimation is distinct because it uses modeling assumptions in a Bayesian framework to quantify the uncertainty in our knowledge of the feature-label distribution parameters. Denoting the parameters of class $y \in \{0,1\}$ by θ_y and the corresponding class-conditional distribution by $f^y_{\theta_y}$, the feature-label distribution is completely characterized by $\theta = [c, \theta_0, \theta_1]$. In particular, the true error of ψ_n can be written as,

$$\varepsilon_n(\theta) = c\varepsilon_n^0(\theta_0) + (1-c)\varepsilon_n^1(\theta_1), \tag{2}$$

where we have explicitly indicated the dependence of ε_n and ε_n^y on the distribution parameters (dependence on the sample/classifier will be suppressed in ε_n, ε_n^y and all other related terms).

To simplify the analysis, we assume c, θ_0 and θ_1 are independent prior to observing the data, and denote their marginal priors by $\pi(\theta_0)$, $\pi(\theta_1)$ and $\pi(c)$. After observing the sample, independence is preserved and the sample is used to update the priors to posteriors, $\pi^*(\theta_0)$, $\pi^*(\theta_1)$ and $\pi^*(c)$. For instance, given a uniform prior on c from 0 to 1, it can be shown that

$$\pi^*(c) = \frac{(n+1)!}{n_0! n_1!} c^{n_0}(1-c)^{n_1}, \tag{3}$$

$$E_{\pi^*}[c] = \frac{n_0+1}{n+2}, \tag{4}$$

where E_{π^*} is shorthand notation for the expectation given the sample, or equivalently, the expectation relative to the posterior uncertainty in the relevant feature-label distribution parameters. A more general class of priors on c is considered in [1]. In addition, we may write the posterior distributions for θ_y by

$$\pi^*(\theta_y) \propto \pi(\theta_y) \prod_{i=1}^{n_y} f^y_{\theta_y}(\mathbf{x}_i^y), \tag{5}$$

where $\mathbf{x}_i^y$ is the ith sample point in class y and the constant of proportionality is found by normalizing the integral of $\pi^*(\theta_y)$ to 1. When the prior density is proper, this follows from Bayes' rule and if $\pi(\theta_y)$ is improper this is taken as a definition, but in all cases it is mandatory that $\pi^*(\theta_y)$ be a proper density.

Under weak regularity assumptions it is well known that the MMSE estimator of a random variable is equivalent to its conditional expectation given the observations. Hence, given an observed sample, S_n, and a fixed classifier, ψ_n, the Bayesian error estimator, defined to be the MMSE estimate of the true error, is the expected true error conditioned on the sample

$$\widehat{\varepsilon}_{\mathrm{MMSE}} = E_{\pi^*}[\varepsilon_n(\theta)] = E_{\pi^*}[c]\widehat{\varepsilon}^0 + (1-E_{\pi^*}[c])\widehat{\varepsilon}^1, \tag{6}$$

where we have used the posterior independence between θ_0, θ_1 and c, and we define $\widehat{\varepsilon}^y = E_{\pi^*}[\varepsilon_n^y(\theta_y)]$. When the prior probabilities are improper, this is called the generalized Bayesian error estimator. Note that this is a training data error estimator, meaning that no sample points are held out for error estimation and the entire sample set is used to update the priors and estimate the true error.