Pattern Recognition 48 (2015) 3766–3782

Contents lists available at ScienceDirect

Pattern Recognition

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

Discrete optimal Bayesian classification with error-conditioned sequential sampling

Ariana Broumand [a,*], Mohammad Shahrokh Esfahani [a,b], Byung-Jun Yoon [a,b,c], Edward R. Dougherty [a,b]

[a] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[b] Center for Bioinformatics and Genomic Systems Engineering (CBGSE), Texas A&M University, College Station, TX, USA
[c] Hamad bin Khalifa University (HBKU), College of Science and Engineering (CSE), Doha, Qatar

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
3 March 2015
Accepted 30 March 2015
Available online 16 April 2015

Keywords:
Optimal Bayesian classifier
Controlled sampling
Prior knowledge

ABSTRACT

When in possession of prior knowledge concerning the feature-label distribution, in particular, when it is known that the feature-label distribution belongs to an uncertainty class of distributions governed by a prior distribution, this prior knowledge can be used in conjunction with the training data to construct the optimal Bayesian classifier (OBC), whose performance is, on average, optimal among all classifiers relative to the posterior distribution derived from the prior distribution and the data. Typically in classification theory it is assumed that sampling is performed randomly in accordance with the prior probabilities on the classes and this has heretofore been true in the case of OBC. In the present paper we propose to forego random sampling and utilize the prior knowledge and previously collected data to determine which class to sample from at each step of the sampling. Specifically, we choose to sample from the class that leads to the smallest expected classification error with the addition of the new sample point. We demonstrate the superiority of the resulting nonrandom sampling procedure to random sampling on both synthetic data and data generated from known biological pathways.

1. Introduction

In many classification applications one is limited to small samples. For instance, in medicine, where classification may involve diagnosis, prognosis, or treatment option, data can be limited due to specimen availability, cost, or the time necessary to obtain and process specimens (which is related to cost). In classification theory it is generally assumed that sampling is random, meaning that the training data are independent and identically distributed (i.i.d.); indeed, assumption of random sampling is typically made throughout a text on classification. For instance, Devroye et al. declare on page 2 of their text that all sampling is random [1]. The assumption is so pervasive that it may be applied without being mentioned. Duda et al. state: "In typical supervised pattern classification problems, the estimation of the prior probabilities presents no serious difficulties." [2]. Implicit in this statement is that the ratio of the number of data points in a class with respect to the total sample size converges to the class probability, as it does in the case of random sampling according to Bernoulli's law of large numbers. No doubt, random sampling has advantages, but is it most efficient in classifier design, especially when one is constrained to small samples?

The effects of nonrandom sampling owing to correlation in the training data have been examined as far back as the early 1970s using numerical examples [3] and the issue subsequently has been examined by studying the effects on asymptotic error rates in the context of linear discriminant analysis (LDA) [4–6]. With small samples, asymptotic results are not really relevant. More recently, nonrandom sampling has been addressed for finite samples by providing representation of the first- and second-order moments for expected errors arising from nonrandom sampling, again in the framework of LDA [7]. In particular, these results demonstrate that nonrandom sampling can be advantageous depending on the correlation structure within the data.

Here we consider a specific scenario for nonrandom sampling. Given a sample, S_n, consisting of n data points, if another data point is to be selected and a classifier designed from the larger sample, S_{n+1}, would it be better to select the new point in an i.i.d. fashion, which means it could come from either class-conditional distribution, or to predetermine the class from which it is to be chosen based on some class-selection criterion, in which case S_{n+1} would not be a random sample, even if S_n were a random sample?

* Corresponding author.
E-mail addresses: broumand@tamu.edu (A. Broumand), m.shahrokh@tamu.edu (M.S. Esfahani), byoon@qf.org.qa (B.-J. Yoon), edward@ece.tamu.edu (E.R. Dougherty).

http://dx.doi.org/10.1016/j.patcog.2015.03.023
0031-3203/

The answer depends on having a suitable criterion whose application leads to making a beneficial choice as to whether or not to select an i.i.d. data point. By working within the framework of optimal Bayesian classification, we can establish such a criterion and obtain an advantageous nonrandom sampling procedure. In this framework, one has an uncertainty class of possible feature-label distributions and a prior distribution governing the uncertainty class. This allows one to determine the minimum mean-square-error (MMSE) estimate of the error based on the prior distribution and the data [8,9]. An optimal Bayesian classifier (OBC) possesses minimum expected error across the uncertainty class [10,11]. Relative to the sampling procedure, the aim is to select the next data point in such a way as to minimize the expected error of the optimal Bayesian classifier, the critical point being that the Bayesian framework facilitates determination of the expected error, which is impossible in the ordinary purely data-driven setting.

This work focuses on discrete classification. Using simulations, both with synthetic and simulated data from real biological pathways, we demonstrate the effectiveness of the proposed nonrandom sampling paradigm relative to random sampling and also examine some of its properties.

Other methods for nonrandom sampling have been proposed that possess conceptual similarities as well as vital differences with the approach proposed herein. These include online learning and active sampling (learning).

In online learning, sequential measurements are made, one at a time, to improve an uncertain model. In particular, the knowledge gradient (KG) algorithm assumes that one of M alternatives can be measured at each time step, each yielding a random reward with an unknown mean and known variance (corresponding to measurement error) [12]. The aim is to make sequential measurements that will maximize the expected total reward to be collected over a time period, thereby treating the problem as a multi-armed bandit process [13]. To achieve this goal, at every time step one tries to identify the optimal KG policy that allows one to choose a measurement (among the M available alternatives) that is expected to bring the largest improvement. The alternative measurements (or rewards) are typically assumed to be independent Gaussian random variables and prior knowledge concerning the measurements and their correlations can be incorporated into the problem via their joint distribution. Our proposed Bayesian framework for nonrandom sampling utilizes a substantially different approach, in that it puts a prior distribution on an uncertainty class of feature-label distributions. Among the key differences resulting from this Bayesian framework is that the distribution of the reward (cost) is not directly modeled; instead, we estimate the expected cost, which is classification error. Moreover, we do not impose restrictions on the variance of our cost/reward in the case of pursuing each policy.

Active sampling has a long history in machine learning, going back to [14,15]. As discussed in [16], the essence of active sampling algorithms is to control the selection of potential unlabeled training points in the sample space to be labeled and used for further training. A generic active sampling algorithm is described in [17]. While there are conceptual similarities with our work, there are fundamental differences. Our goal is not to search among unlabeled sample points for those for which we wish to generate labels; rather, we generate new sample points from a chosen known label. Moreover, we directly target reduction of classification error. Reducing uncertainty in our class probability distributions is a side effect, not the direct goal. Considering active learning under a Bayesian framework as in [18] does not eliminate the difference because the underlying strategy is to choose sample points to label.

The rest of the paper is organized as follows. In Section 2 the general framework of the discrete classification problem and the optimal Bayesian classifiers is introduced. In Section 3 the proposed sampling algorithm is introduced. Section 4 shows some results of applying the proposed sampling method in the classification problem with synthetic data from a Zipf model. In Section 5 the effect of the proposed method is studied on data generated from pathways. Section 6 concludes the paper.

Throughout this paper, we use bold letters to denote vectors, e.g. $\boldsymbol{p}$ or $\mathbf{U}$. Capital letters are used for random variables; when in bold they denote a random vector. The notation $E_{\pi(\theta)}[\cdot]$ is the expectation with respect to the parameter θ distributed by $\pi(\theta)$.

2. The discrete model and optimal Bayesian classifier

The discrete model consists of b bins and two classes, $y \in \{0,1\}$, with $\{p_i\}_1^b$ and $\{q_i\}_1^b$ being class-conditional probabilities for $i \in \mathcal{X} = \{1,\ldots,b\}$, and c being the prior probability of class 0, i.e. $P(X=i|y=0)=p_i, P(X=i|y=1)=q_i$ for $i=1,\ldots,b$, and $c=P(y=0)$.

A classifier is a function ψ that maps sample points to a class, $\psi:\{1,\ldots,b\}\to\{0,1\}$. The true classification error ε is the probability that a sample point from class y is classified by ψ as belonging to a different class; $\varepsilon = P(\psi(X)\neq y)$. The error can be decomposed as a weighted average of ε^0 and ε^1, the classification errors for classes 0 and 1, respectively, by

$$\varepsilon = c\underbrace{P(\psi(x)=1|y=0)}_{\varepsilon_0} + (1-c)\underbrace{P(\psi(x)=0|y=1)}_{\varepsilon_1}. \quad (1)$$

In the discrete model,

$$\varepsilon^0 = \sum_{i=1}^{b} p_i I_{\psi(i)=1} \quad \text{and} \quad \varepsilon^1 = \sum_{i=1}^{b} q_i I_{\psi(i)=0}, \quad (2)$$

where $I_{(\cdot)}$ is the indicator function, e.g. $I_{\psi(i)=1}=1$ if and only if $\psi(i)=1$.

Let $S_n=\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$ be a sample taken from the feature-label distribution. Let $U_i=u_i$ and $V_i=v_i$ denote the numbers of observed sample points from classes 0 and 1 in bin i, respectively. The *histogram rule* assigns a class label to each bin via majority voting in that bin:

$$\psi_{hist}(X=i)=\begin{cases}1 & \text{if } v_i>u_i\\ 0 & \text{otherwise.}\end{cases}$$

The histogram rule is the plug-in rule for discrete classification and is consistent. Hence, it is well-suited for large-sample applications; however, our interest is with small samples and therefore we utilize the optimal Bayesian classifier (OBC) [10,11].

For the OBC, we assume that the actual model belongs to an *uncertainty class*, Θ, of discrete feature-label distributions parameterized by $\boldsymbol{\theta}=[c,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1]$, where $\boldsymbol{\theta}_0=[P_1,\ldots,P_{b-1}]$ and $\boldsymbol{\theta}_1=[Q_1,\ldots,Q_{b-1}]$. Define $\mathbf{P}=[P_1,\ldots,P_{b-1},P_b]$ and $\mathbf{Q}=[Q_1,\ldots,Q_{b-1},Q_b]$. Because each vector $\mathbf{P}$ and $\mathbf{Q}$ forms a probability mass function (PMF), the last bin probabilities are defined by $P_b=1-\sum_{k=1}^{b-1}P_k$ and $Q_b=1-\sum_{k=1}^{b-1}Q_k$, so they are not free parameters and are dropped from the parameter model. Note that, since in the Bayesian context the PMFs above are random, they are denoted by capital letters.

Table 1 summarizes the main variables in this and the next sections. For a complete description of the details used in our framework, the reader is referred to [10,11].

Prior knowledge pertaining to $\boldsymbol{\theta}$ is in the form of a *prior probability distribution* $\pi(\boldsymbol{\theta})$. The prior is updated by getting new data sample points and obtaining the posterior distribution, $\pi^*(\boldsymbol{\theta})$. To facilitate analytic representations, it is assumed in [8] that $c,\boldsymbol{\theta}_0$,