Statistics and Probability Letters 119 (2016) 91–100

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

An asymptotically optimal kernel combined classifier

Majid Mojirsheibani*, Jiajie Kong

Department of Mathematics, California State University Northridge, CA, 91330, USA

ARTICLE INFO

Article history:
Received 27 March 2016
Received in revised form 10 July 2016
Accepted 19 July 2016
Available online 27 July 2016

Keywords:
Kernel
Hamming distance
Smoothing parameter
Combined classifier

ABSTRACT

A kernel ensemble classifier is developed for accurate classification based on several initial classifiers. A data-driven choice of the smoothing parameter of the kernel is considered and the resulting classifier is shown to be asymptotically optimal. Therefore, the proposed combined classifier asymptotically outperforms each individual classifier.

1. Introduction

This article focuses on methods and procedures that can be used to combined several individual classifiers in order to develop more effective classification rules. Different classifiers are usually constructed based on different theories or different estimation procedures and, depending on the type of distributional assumptions imposed on the data, some classifiers perform better in the sense of having lower misclassification error rates. However, given a number of individual classifiers, it is not always clear as to how to develop a systematic approach or a general framework for choosing the classifier with the smallest error rate; this is particularly true in nonparametric situations. Furthermore, since the exact distribution of the data is virtually always unknown, and since each classifier may have certain unique merits, it would be logical to somehow combine the individual classifiers in such a way that the resulting combined classifier would be superior. Combined or ensemble classification results in the literature may be divided into two main types: (i) those that start with a large number of *base* classifiers, which are usually homogeneous in nature (such as a large number of tree classifiers), which will then be combined into a final classification rule. (ii) Those that combine a number of classifiers that have been constructed using different theories or estimation procedures, which is the main focus of this paper. Breiman's (1995, 2001) popular Random Forests classifier falls under (i) Other results that fall under (i) include Biau et al. (2008) and Lin and Jeon (2006). On the other hand, the work of LeBlanc and Tibshirani (1996), Mojirsheibani (1999), and Mojirsheibani and Montazeri (2015) fall primarily under (ii) For more on the taxonomy of ensemble classifiers one may refer to, for example, Zhang and Duin (2011).

There are many intuitively appealing methods for combining classifiers in the literature; these methods may be put into two categories themselves: *weighting methods* and *meta-learning*. Weighting methods include the majority voting approach used, for example, by Breiman (1996, 2001) for tree classification, and by Xu et al. (1992) in handwriting recognition. Averaging and weighted-averaging of estimated class conditional probabilities (that are produced by each individual classifier) have also been considered by some authors; see, for example, Xu et al. (1992), Breiman (1995), and LeBlanc and Tibshirani (1996). The work of Adler et al. (2011) on combined classification for paired data, and De Bock et al. (2010) on a

* Corresponding author.
E-mail addresses: majid.mojirsheibani@csun.edu (M. Mojirsheibani), jiajie.kong.595@my.csun.edu (J. Kong).

http://dx.doi.org/10.1016/j.spl.2016.07.017

generalized additive model based approach to ensembles provide additional new directions. There are also other weighting methods that can be found in Rokach (2009, 2010). Meta-learning methods usually work by using the predicted values of the individual classifiers on the data. Relevant results along these lines include the stacked generalization of Wolpert (1992), Breiman's (1995) stacked method, and Mojirsheibani's (1999) nonlinear combined classifiers. For a detailed account of meta-learning methods one may refer to Rokach (2009). Combined or ensemble estimation has also been employed in regression and model selection problems in the literature. See, for example, the work of Yang (2000, 2004) and van der Laan et al. (2007). In a more recent article, Biau et al. (2016) proposed a combined regression estimator which can, asymptotically, outperform each of the individual regression estimators in the L_2 sense.

The rest of the paper is organized as follows. In Section 2 we discuss and study kernel combined classifiers; our key contributions appear in Theorem 2. In Section 2.2 we carry out some numerical studies that confirm our theoretical findings. All proofs are deferred to Appendix.

2. The proposed kernel combined classifier

2.1. Methodology

In this article we consider the following standard two-group classification problem. Let $(\mathbf{X}, Y)$ be a random pair, where $\mathbf{X} \in \mathbb{R}^d$ is called the covariate or feature vector and $Y \in \{0, 1\}$, called the class variable, has to be predicted based on $\mathbf{X}$. The aim of classification is to find a function (a classifier) $\psi : \mathbb{R}^d \to \{0, 1\}$ whose misclassification error rate (i.e., the probability of incorrect prediction), $P\{\psi(\mathbf{X}) \neq Y\}$, is as small as possible. The best classifier, also called the Bayes classifier, is given by

$$\psi_B(\mathbf{X}) = \begin{cases} 1 & \text{if } P\{Y = 1|\mathbf{X}\} > \dfrac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $P\{Y = 1|\mathbf{X}\} = E[I\{Y = 1\}|\mathbf{X}]$ and where $I\{C\}$ stands for the indicator function of the set C; see, for example, Devroye et al. (1996, Ch. 2). In practice the distribution of $(\mathbf{X}, Y)$ is always fully or partially unknown and therefore finding the function ψ_B is impossible. Now suppose that we have available n independently and identically distributed (i.i.d.) observations, i.e., the data, $\mathbb{D}_n := \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i) \overset{\text{i.i.d.}}{=} (\mathbf{X}, Y)$, $i = 1, \dots, n$, and let $\widehat{\psi}_{n,1}, \dots, \widehat{\psi}_{n,M}$ be M different individual classifiers based on the data $\mathbb{D}_n$. For example, $\widehat{\psi}_{n,1}$ may be a kernel classifier, $\widehat{\psi}_{n,2}$ a nearest neighbor (NN) classifier, $\widehat{\psi}_{n,3}$ might be something more complicated such as Breiman's (2001) random forests classifier or the support vector machines (SVM) of Boser et al. (1992), etc. Next let

$$\widehat{\boldsymbol{\psi}}_n(\mathbf{X}) = \left(\widehat{\psi}_{n,1}(\mathbf{X}), \dots, \widehat{\psi}_{n,M}(\mathbf{X})\right)'$$

be the vector of M individual data-based classifiers and define the combined classifier ϕ_n^* by

$$\phi_n^*(\widehat{\boldsymbol{\psi}}_n(\mathbf{X})) = \begin{cases} 1 & \text{if } P\left\{Y = 1\middle|\widehat{\boldsymbol{\psi}}_n(\mathbf{X})\right\} > \dfrac{1}{2} \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} 1 & \text{if } E\left[(2Y-1)\middle|\widehat{\boldsymbol{\psi}}_n(\mathbf{X})\right] > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

It is straightforward to show that (2) is theoretically optimal in the sense that its overall misclassification error probability, $P\{\phi_n^*(\widehat{\boldsymbol{\psi}}_n(\mathbf{X})) \neq Y\}$, is less than or equal to that of any other combined classifier. More formally, we have the following elementary result

Theorem 1. *The combined classifier ϕ_n^* in (2) is optimal, i.e.,*

$$P\left\{\phi_n^*(\widehat{\boldsymbol{\psi}}_n(\mathbf{X})) \neq Y\right\} = \inf_{\phi:\{0,1\}^M \to \{0,1\}} P\left\{\phi(\widehat{\boldsymbol{\psi}}_n(\mathbf{X})) \neq Y\right\}.$$

See Appendix for a proof.

Of course, the above result immediately implies that for each individual classifier $\widehat{\psi}_{n,j}$, $j = 1, \dots, M$, we have $P\{\phi_n^*(\widehat{\boldsymbol{\psi}}_n(\mathbf{X})) \neq Y\} \leq P\{\widehat{\psi}_{n,j}(\mathbf{X}) \neq Y\}$. Unfortunately the combined classifier ϕ_n^* in (2) is not available in practice because it depends on unknown conditional probabilities. In what follows we propose a kernel estimator of ϕ_n^*, where the smoothing parameter of the kernel is estimated by a data-driven choice that minimizes the so-called *resubstitution* estimate of the probability of misclassification. We recall that for any classifier ψ_n, constructed based on the data $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, the resubstitution estimator of the misclassification probability of ψ_n is given by $n^{-1}\sum_{i=1}^n I\{\psi_n(\mathbf{X}_i) \neq Y_i\}$. Our proposed approach uses a data-splitting method that works as follows. Start by randomly splitting the data $\mathbb{D}_n := \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ into two parts, $\mathbb{D}_m$ of size m and $\mathbb{D}_\ell$ of size ℓ, where $\ell + m = n$ and $\mathbb{D}_m \cup \mathbb{D}_\ell = \mathbb{D}_n$. Also, let $\widehat{\psi}_{m,1}, \dots, \widehat{\psi}_{m,M}$ be the individual classifiers constructed based on the subsample $\mathbb{D}_m$ only (instead of $\mathbb{D}_n$). Now, consider the