Journal of Mathematical Psychology 57 (2013) 237–248

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Optimal receiver operating characteristic manifolds

Yun-Jhong Wu[a], Chin-Tsang Chiang[b,*]

[a] Department of Statistics, University of Michigan, United States

[b] Department of Mathematics and Institute of Applied Mathematical Sciences, National Taiwan University, Taiwan

HIGHLIGHTS

- The general ROC space was established to assess the performance of a marker with a family of classifiers.
- The conditions are addressed to ensure the well-behaved hypervolume under the optimal ROC manifold (HUM).
- The equality for HUM and the correctness probability for general multi-classification is demonstrated.
- A simple and easily implemented estimation approach is proposed.
- More precise bounds in terms of the HUMs in partial classifications are provided.

ARTICLE INFO

Article history:
Received 14 September 2012
Available online 6 September 2013

Keywords:
Gaussian process
Hypervolume
Manifold
Optimal classification
Receiver operating characteristic
U-estimator

ABSTRACT

To evaluate the overall discrimination capacity of a marker for multi-class classification tasks, the performance function is a natural assessment tool and fully provides the essential ingredients in receiver operating characteristic (ROC) analysis. The optimal ROC manifolds supply a geometric characterization of the magnitude of separation among multiple classes. It has been shown that the hypervolume under the optimal ROC manifold (HUM) is a well-defined and meaningful accuracy measure only in suitable ROC subspaces. In this article, we provided a rigorous proof for the equality of HUM and its alternative form, the correctness probability, which is directly related to an explicit U-estimator. In addition, extensive simulations are conducted to investigate the finite sample properties of the proposed estimators and the related inference procedures. Further, a rule of thumb is given in application to assess for the HUM. Conclusively, our theoretical framework allows more sophisticated modeling on the performance of markers and helps practitioners examine the optimality of applied classification procedures.

1. Introduction

The past decade has seen the rapid development of multi-classification in various areas of science. For instance, distinguishing species in biology, image recognition in electronic engineering, and pricing strategy in business can be formulated as multi-classification problems. Recently, biomedical researchers have shown an increased interest in accurately determining types of specific diseases or staging cancers, provided that markers contain only limited information regarding the true types. Despite well-established statistical methods for binary classification problems, such as for distinguishing between diseased and non-diseased patients, it is still questionable whether the existing methodology can appropriately help working scientists to compare performances of various markers, and, if possible, find an optimal marker based on some rational criteria.

A typical task of multi-classification is mainly based on data of the type $(G, \mathbf{Y})$ and a classifier $\widehat{G}$, where $G \in \{1, \ldots, K\}$ represents the true class, $\mathbf{Y} \in \mathcal{Y}$ denotes a univariate or multivariate marker, and $\widehat{G}$ is a random function from $\mathcal{Y}$ to $\{1, \ldots, K\}$. The performance probabilities $p_{jk}(\widehat{G}) = P(\widehat{G}(\mathbf{Y}) = j|G = k), j, k \in \{1, \ldots, K\}$, of $(\widehat{G}, \mathbf{Y})$ are commonly used to assess the considered classification procedure. For the sake of numerical stability in estimation, $p_{jk}(\widehat{G})$'s are frequently applied and more preferred than the prediction probabilities $P(G = k|\widehat{G}(\mathbf{Y}) = j)$'s. Furthermore, we can show that there is little connection between the optimization in terms of prediction and performance probabilities although these two are equivalent in binary classification. Since an assessment measure based on performance probabilities is of the form $f(\widehat{G}, \mathbf{Y})$, the performance of certain classifiers only represents partial information on the discrimination capacity of markers. Thus, evaluation of markers with respect to only a part of classifiers could be too naive to be used to make a fair comparison among markers. Indeed, a rational assessment index of each marker should be a function only of $\mathbf{Y}$ and then is unchanging with chosen classifiers. One of the most

* Corresponding author.
E-mail address: chiangct@ntu.edu.tw (C.-T. Chiang).

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.08.001

practical ways to adopt for this reason is to address the marker of interest with respect to the overall performance of all classifiers. To achieve this research aim, receiver operating characteristic (ROC) analysis, a technique initially only for binary classification tasks, has been extended to multi-classification in recent years.

Meaning of optimality in classification could be various, but some optimal criteria are shown to be equivalent in the sense of overall performance. A seminal work on optimal ROC manifolds was promulgated by Scurfield (1998) via maximizing $\sum_{k=1}^{K} P(\widehat{G} = k, G = k)$. The author constructed ROC manifolds based on performance probabilities of optimal deterministic classifiers with the maximal sum of true probabilities and derived that HUM equals correctness probability for the ternary classification procedures; the optimal classifiers can be represented as combinations of linear classifiers in the decision space spanned by log-likelihood ratio scores. Since K^2 performance probabilities are available to describe a K-classification procedure, Edwards, Metz, and Kupinski (2004) explained that it would be insufficient to describe the complete information of a classifier only based on true probabilities $p_{kk}(\widehat{G})$'s for $K \geq 3$; they further suggested to maximize the expected utility (or minimize the expected cost), and this criterion can be formulated in a manner of linear classifiers in the decision space. For ternary classification, He and Frey (2006) indicated that the utility classifiers, which means classification based on the criterion of maximizing specified utility, have maximal sum of true probabilities under the setting of equal error utilities. As an alternative approach, Schubert, Thorsen, and Oxley (2011) utilized Minkowski's functionals to determine the optimal classifier. Roughly speaking, the functional is to define an optimal classifier as that with the minimal misclassification rates under constraints on ratios among these probabilities. The criterion is essentially analogue to maximizing the expected utility under some conditions, although their illustrated examples do not actually achieve the optimality in the sense of performance probabilities and it is difficult to give a feasible formula of their defined optimal classifier. Due to the difficulty in visualization for performance in multi-classification, a vital issue arises to define an appropriate summary index for the performance of a marker. Naturally, HUM is a direct extension of the area under the ROC curve (AUC) and was employed in many foregoing works. In binary classification, the induced optimal ROC curve certainly separates the ROC space into two regions. However, optimal ROC manifolds may be unable to enclose a bounded set and, hence, the well-definition of HUM might be thrown in doubt. Besides, Edwards, Metz, and Nishikawa (2005) used some examples to explain that both of the resulting HUMs from near-perfect and non-informative markers are near zero; these authors further concluded that HUM is not a suitable summary index for performance of a marker.

The breakthrough results we have achieved (Wu & Chiang, 2011) are initially based on a theoretical formulation in terms of utility, which concisely describes current results regarding multi-class ROC analysis. The groundwork leads to a better understanding of some geometric characteristics of optimal ROC manifolds and gives a base to establish the asymptotic process of empirical optimal ROC manifolds. We also address the sufficient and necessary conditions to ensure the well-behaved HUM; one can clearly interpret some peculiar numerical and algebraic results occurring in the foregoing works and further advocate practitioners to create a handy summary assessment. Instead of the setting in ternary classification, we borrowed a tool in graph theory to confirm the validation of the equality between HUM and correctness probability for any K-class optimal procedures. Thus, by using this explicit and meaningful probability expression, a U-estimation for HUM then becomes applicable for more general classification procedures. In considering practical implications, we proposed an estimation approach for HUM with related inference procedures through some widely used models on the relationship between G and $\mathbf{Y}$ through a prospective or retrospective perspective. Furthermore, an empirical rule based on partial-classification HUMs is proposed to assist practitioners in evaluating the discriminability. Although our work focuses on continuous markers, most of these results are comfortably adapted to evaluation of discrete or mixture markers and could serve as a basis for more sophisticated statistical methods.

On the whole, based on the properties of optimal ROC manifolds we have established, we provide an estimation and inference procedure for the discriminability of multi-classification markers. The properties enable us to draw pointwise and functional inference for optimal ROC manifolds in Section 2. Section 3 is devoted to estimation and model-based inference procedures for HUM. Numerical experiments and an application to empirical data in Section 4 illustrate the practicality of our developed methodology. Finally, Section 6 summarizes the findings in this study and makes some remarks for future research.

2. Optimal ROC manifolds

Some researchers have worked on construction of ROC manifolds; however, without optimality in classifiers, the so-called ROC manifold could be an arbitrary subset of a projection of the performance set

$$\phi(\mathcal{C}) = \{\phi(\widehat{G}) : \widehat{G} \in \mathcal{C}\},$$

where $\phi(\widehat{G})$ is the performance of $\widehat{G}$ defined as $(p_{11}(\widehat{G}), p_{12}(\widehat{G}), \dots, p_{1K}(\widehat{G}), p_{21}(\widehat{G}), \dots, p_{KK}(\widehat{G}))^\top$ in $\mathcal{R}$ and $\mathcal{C}$, the set of all possible deterministic and randomized classifiers, in the ROC space

$$\mathcal{R} = \left\{ \mathbf{p} = (p_{11}, p_{12}, \dots, p_{1K}, p_{21}, \dots, p_{KK})^\top : \sum_{j=1}^{K} p_{jk} = 1 \ \forall k = 1, \dots, K \right\}$$

rather than a manifold in the context of geometry. Therefore, few features of the ROC manifold sets could be pinpointed, and estimation of ROC manifold sets and related summary measures might lead to a more complicated situation. We hence introduce optimal ROC manifolds for multi-classification as an extension of optimal ROC curves for binary classification. For K-classification tasks, there are K redundant coordinates in $\mathcal{R}$. Practically, not all K^2 performance probabilities are of interest. We can further consider an ROC subspace $\mathcal{R}_S$ where S denotes the set of coordinates of concern. In the sequel, sets or operators with the subscripted S denotes that they are restricted in the ROC subspace $\mathcal{R}_S$.

Indeed, the performance set $\phi(\mathcal{C})$ is a convex and compact set and hence can be completely characterized through investigating its boundary set $\partial\phi(\mathcal{C})$ (Wu & Chiang, 2011); these features also hold in arbitrary $\mathcal{R}_S$. Moreover, $\partial\phi(\mathcal{C})$ is also able to be regarded as an operator with $\mathcal{Y} \mapsto \phi(\mathcal{C})$ only depending on $\mathcal{Y}$. The admissibility of a classifier G (i.e. there does not exist another classifier G' such that $p_{jj}(G) \leq P_{jj}(G')$, $p_{jk}(G) \geq P_{jk}(G')$ for $j \neq k$ and at least one inequality is strict) is equivalent to that it satisfies the maximizing-utility criterion, which means maximizing $\sum_{j,k} u_{jk} p_{jk}(G)$ for some specified utility values u_{jk} satisfying $u_{jj} \geq 0$ and $u_{jk} \leq 0$ for $j \neq k$. Through the optimality in the above sense, several results are obtained: first, the performance of each admissible classifier is located in $\partial\phi(\mathcal{C})$; this justifies using the optimal ROC manifold

$$M_S = \{\phi_S(\widehat{G}) : \widehat{G} \text{ is admissible in } S.\},$$

where S denotes the set of correct classification or misclassification rates of interest, as a measure of discriminability of markers. Second, the maximizing-utility criterion gives a natural parametric system to describe M_S. Hence, M_S can be treated as actually a function of $\mathbf{u} \in \mathcal{U}$ on an interesting set $\mathcal{U}$ of utility values, denoted by