Pattern Recognition 60 (2016) 770–777

Contents lists available at ScienceDirect

Pattern Recognition

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

ROC representation for the discriminability of multi-classification markers

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

[a] Department of Statistics, University of Michigan, United States
[b] Institute of Applied Mathematical Sciences, National Taiwan University, Taiwan

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
9 April 2016
Accepted 26 June 2016
Available online 1 July 2016

Keywords:
Discriminability
Hypervolume
Manifold
Optimal classification
Receiver operating characteristic
Utility

ABSTRACT

In this paper, the receiver operating characteristic (ROC) representation and its accuracy measures are well-defined and meaningful assessments for the discriminability of multi-classification markers are shown. Given a set of classifiers $\mathcal{C}$, a parameterized system can be used to characterize the corresponding optimal ROC manifold. A connection with the decision set further leads to a better understanding of some geometric features of optimal ROC manifolds and preserves the simplicity in computing the hypervolume under the ROC manifold (HUM). In addition, it motivates us to address the necessary and sufficient conditions for the existence of the HUM. To sum up, this work provides working scientists with an extension of the two-class ROC analysis to the multi-classification ROC analysis in a theoretically sound manner.

1. Introduction

Receiver operating characteristic (ROC) analysis, which was originally developed for radar signal detection, is a technique initially created for assessing the performance of binary classification markers and has been extended to multi-classification (see [1–3]). In application, a marker is generally referred to as a traceable substance whose presence indicates the existence of some state such as a particular disease condition. As the ROC curve for binary classification, the ROC manifold is a natural extension to display the trade-off between the correct classification probabilities and misclassification probabilities. However, the definition of ROC manifold can be stated in a more mathematically rigorous manner. We address the concern about the existence of the hypervolume under the ROC manifold (HUM), which is an analog of the area under the ROC curve (AUC). A significant research finding by [4] also indicated that the existence of the HUM is still in doubt. To clarify the problem and demonstrate its importance in application, a theoretical unification of ROC manifolds will be established in this paper.

Typically,a multi-classification task is mainly based on data of the type $(G, \mathbf{Y})$ and a classifier $\hat{G}$, where a multi-categorical response G stands for the true class with K possible values $\mathcal{K} = \{1, \ldots, K\}$, $\mathbf{Y} \in \mathcal{Y}$ denotes a univariate or multivariate marker value,and $\hat{G}$ is a random function from $\mathcal{Y}$ to $\mathcal{K}$. An extension of ROC analysis to multi-classification has been initially developed in sequential classification procedures,which have excited interest for practical and theoretical simplicity. These algorithms simplify multi-classification tasks to a series of binary classifications as the form $G=k$ versus $G \in \{k+1, \ldots, K\}$ by order $k = 1, \ldots, K$. The first systematic study of a ternary classification problem can be traced back to the paper of [5]. For a univariate marker value $\mathbf{Y}$, Scurfield constructed the ROC manifold for ternary classification to visualize a set generated by $\{p_{1\sigma(1)}(\hat{G}, \mathbf{Y}), p_{2\sigma(2)}(\hat{G}, \mathbf{Y}), p_{3\sigma(3)}(\hat{G}, \mathbf{Y})\} \in [0, 1]^3$, where σ is a permutation function on $\{1, 2, 3\}$ and $p_{jk}(\hat{G}, \mathbf{Y})$ represents the conditional probability $P(\hat{G} = j | G = k)$, which we will call the performance probability hereinafter. To accommodate a multivariate marker, Mossman [1] also developed a classification rule by utilizing a mapping between each G and $\mathbf{Y}$. Although such classification actually offers a perspective to extend traditional ROC analysis for multi-classification, this work is generally not optimal in terms of performance probabilities and is of limited applicability. In practice, a monotone likelihood ratio (MLR) condition should be satisfied (cf. [6]) to ensure the optimality of commonly used sequential procedures. By applying a multinomial logistic regression model, Li and Fine [7] extended the foregoing approach to address a multi-categorical response. Based on the ROC manifold generated by the correct classification probabilities, Zhang and Li [8] further employed a general semi-parametric model of [9] to seek an optimal composite marker. As we shall indicate in this paper, an optimal ROC manifold enjoys some geometric characteristics such as the regularity (see

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

http://dx.doi.org/10.1016/j.patcog.2016.06.024
0031-3203/

[10]) and smoothness. Under some suitable conditions, the equality between the corresponding HUM and the correctness probability (CP) can also be found in [1–3,6,11]. However, non-optimality of ROC manifolds might lead to the lack of explanation for such a particular summary assessment.

Indeed, ROC analysis gives an illuminating insight into the assessment for the discriminability of markers. It is rational to adopt the performance probabilities $p_{jk}(\widehat{G}, \mathbf{Y}) = P(\widehat{G} = j|G = k)$, $j, k = 1, \ldots, K$, to assess the considered classification procedures. For any $\widehat{G}$ in a family of classifiers $\mathcal{C}$, its performance function $\mathbf{p}(\widehat{G}, \mathbf{Y}) = (p_{11}(\widehat{G}, \mathbf{Y}), \ldots, p_{1K}(\widehat{G}, \mathbf{Y}), \ldots, p_{K1}(\widehat{G}, \mathbf{Y}), \ldots, p_{KK}(\widehat{G}, \mathbf{Y}))^\top$ can be naturally plotted in a general ROC set:

$$\mathcal{R} = \left\{ (\xi_{11}, \ldots, \xi_{1K}, \ldots, \xi_{K1}, \ldots, \xi_{KK})^\top \in [0, 1]^{K^2} : \sum_{j=1}^{K} \xi_{jk} = 1, 0 \le \xi_{jk} \le 1 < \right\}. \tag{1.1}$$

Following from the above definition, $\mathcal{R}$ is a unit cube in $\mathbb{R}^{K^2}$ space that contains a $K(K-1)$-dimensional hyperplane and sufficiently represents all possible performance functions. In application, practitioners often focus on partial information of $\mathcal{R}$ such as a subset corresponding to the correct classification probabilities $\{p_{kk}: k \in \mathcal{K}\}$ or the misclassification probabilities $\{p_{jk}: j, k \in \mathcal{K} \text{ with } j \neq k\}$. For such purpose, the symbol S is used to denote the performance probabilities of interest, which generates a smaller ROC set $\mathcal{R}_S$ in $\mathcal{R}$, and the considered operators or sets restricted to $\mathcal{R}_S$ are subscripted by S. Thus, a partial performance can be determined as a projection from $\mathcal{R}$ onto $\mathcal{R}_S$. As it is well known in binary classification, the corresponding performance probabilities of a set of classifiers in $\mathcal{R}_S$ with $S = \{p_{11}, p_{12}\}$ might not necessarily form a ROC curve but can be plotted as a representation for the discriminability of classification procedures. It is noted that such a representation is not straightforward for arbitrary K-classification tasks. Starting with the concept of a proper assessment for the discriminability of markers, the ROC representation is brought up in this work and is introduced in the next section.

2. ROC representation

The performance of a classification procedure can be represented by a function of performance probabilities $\phi(\mathbf{p}(\widehat{G}, \mathbf{Y}))$ for different choices of functions ϕ such as the performance function and the expected utility defined in Sections 2.1 and 2.2. To assess the discriminability of $\mathbf{Y}$ in the sense of "fairness", the probability-based performance assessment should be a function only of markers and invariant with respect to chosen classifiers. With an argument slightly different from [12], this assessment is also shown to be equivalent to the ROC representation.

2.1. Performance sets

To simplify notations, the performance probability $p_{jk}(\widehat{G}, \mathbf{Y})$ of a given $\mathbf{Y}$ is denoted by $p_{jk}(\widehat{G})$, $j, k \in \mathcal{K}$. The performance function $\phi(\cdot)$ is the conditional probabilities acting on the classifier $\widehat{G}$ denoted by

$$\phi(\widehat{G}) = \left(p_{11}(\widehat{G}), \ldots, p_{1K}(\widehat{G}), \ldots, p_{K1}(\widehat{G}), \ldots, p_{KK}(\widehat{G})\right)^\top, \tag{2.1}$$

in which it still depends on $\widehat{G}$. As for the construction of a proper accuracy measure for the discriminability of $\mathbf{Y}$, it is reasonable to

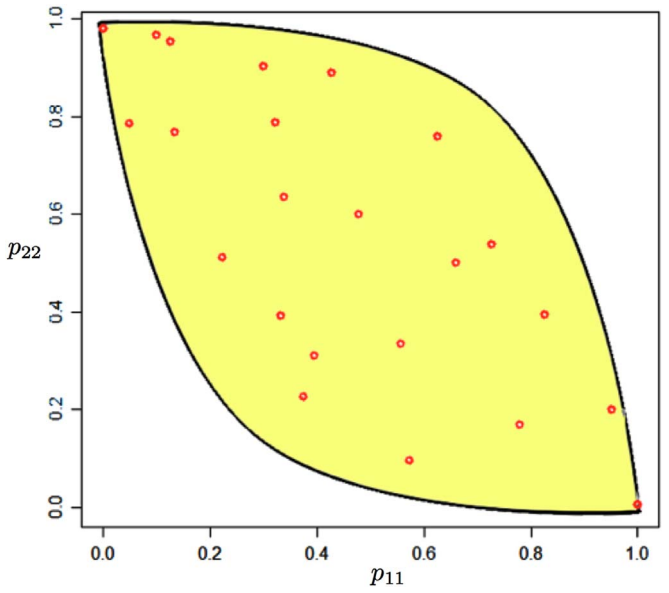

Fig. 1. A notional performance set $\phi_{\{p_{11},p_{22}\}}(\mathcal{C})$ (yellow region and black boundary curve) for a set of binary classification classifiers $\mathcal{C}$ with each red point representing the performance function of a particular classifier in $[0, 1]^2$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

define the performance set of the collection to be

$$\phi(\mathcal{C}) = \{\phi(\widehat{G}): \widehat{G} \in \mathcal{C}\}. \tag{2.2}$$

Here, the set $\mathcal{C}$ consists of a collection of deterministic and random classifiers (see Fig. 1) defined on $\mathcal{Y}$ with outputs in $\mathcal{K}$. Let $\mathbf{P_r}$ denote a probability measure that generates the conditional probabilities p_{jk}'s that define ϕ, $1(\cdot)$ be the indicator function, and W_λ follow a Bernoulli distribution with parameter $\lambda \in [0, 1]$. Throughout this paper, we consider the set $\mathcal{C}$ in a probability sense.

Definition 2.1. The set $\mathcal{C}$ is said to be convex with respect to $\mathbf{P_r}$ and S if $p_{jk}(\widehat{G}_\lambda) \le \lambda p_{jk}(\widehat{G}_1) + (1-\lambda)p_{jk}(\widehat{G}_2)$ for each $\lambda \in [0, 1]$ and all $p_{jk} \in S$ implies $\widehat{G}_\lambda \in \mathcal{C}$ for every $\widehat{G}_1, \widehat{G}_2 \in \mathcal{C}$, where $\widehat{G}_\lambda = \sum_{\ell=1}^{2} 1(W_\lambda = 2 - \ell)\widehat{G}_\ell$.

This definition defines the convexity of a set of classifiers based upon the convexity of the corresponding set of performance vectors. Intrinsically, the set $\phi_S(\mathcal{C})$ contains performance vectors of all existing classifiers in $\mathcal{C}$ with respect to a specified marker value $\mathbf{Y}$ and conveys information about the classification capacity of $\mathbf{Y}$ with respect to $\mathcal{C}$. As a representation of the discrimination ability, $\phi(\mathcal{C})$ should depend only on $\mathbf{Y}$. A remarkable characteristic of the performance set in the following theorem further enables us to identify optimal classification procedures.

Theorem 2.1. *Suppose that $\mathcal{C}$ is a convex set with respect to $\mathbf{P_r}$ and S. Then, the performance set $\phi(\mathcal{C})$ of $\mathbf{Y}$ is convex and compact, and so is $\phi_S(\mathcal{C})$ for any subset of performance probabilities S.*

Proof. See Appendix A.

To characterize the convexity and compactness of $\phi(\mathcal{C})$ (or $\phi_S(\mathcal{C})$), it suffices to portray the boundary set $\partial\phi(\mathcal{C})$ (or $\partial\phi_S(\mathcal{C})$), which will be shown to be related to the optimality of classifiers. In the next subsection, a parameterized system in the decision theory is employed to analyze and compute $\partial\phi(\mathcal{C})$. The properties in Theorem 2.1 further assure the existence of utility classifiers (see Definition 2.3), whose performances fall in $\partial\phi(\mathcal{C})$, and elucidate the optimality of $\partial\phi(\mathcal{C})$. Thus, one can just consider $\partial\phi(\mathcal{C})$ rather than the whole $\phi(\mathcal{C})$ or an arbitrary subset of $\phi(\mathcal{C})$. For an