Neurocomputing 193 (2016) 81–91

Contents lists available at ScienceDirect

Neurocomputing

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

A classification performance measure considering the degree of classification difficulty

Xiaoli Zhang [a,b], Xiongfei Li [a,b,*], Yuncong Feng [a,b]

[a] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
[b] College of Computer Science and Technology, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 5 August 2015
Received in revised form
17 December 2015
Accepted 3 February 2016
Communicated by Deng Cai
Available online 11 February 2016

Keywords:
Classification models
Performance evaluation
Classification difficulty
ROC curves
AUC

ABSTRACT

In the field of classification, classification difficulty of instances is one of vital factors that influence the performance of classifiers, however it has been totally neglected. In this paper, a new performance measure for classification algorithms based on Receiver Operator Characteristic (ROC) curves is proposed with the ability of incorporating the information of difficulty. First, a new ROC curve with the information on classification difficulty is defined, which is abbreviated as diROC curve. The curve is constructed in a two-dimensional graph, on which weighted true positive rate is plotted on *Y*-axis and weighted false positive rate is plotted on *X*-axis. The weights of true positive rates are proportional to classification difficulty index, while those of false positive rates are inversely proportional to classification difficulty index. Then, the Area Under diROC Curves, or simply diAUC, is defined to represent the performance of classifiers quantitatively. We test the diROC curves and diAUC on real-word datasets, the experimental results suggest that they are insensitive to changes in class distribution, and superior to traditional ROC curves and AUC in terms of discrimination.

1. Introduction

Receiver Operator Characteristic (ROC) analysis was initially developed in the field of signal detection as a means of judging whether a blip on the screen represents an enemy, a friend, or noise [1,2]. Then its use was broadened in the 1970s to the biomedical field to interpret medical test results [3–5]. Now, ROC curves have become one of the widely used tools in the performance assessment, and parameter optimization of classifiers [6]. At first, ROC curves are specialized for two-class (or binary) classifiers. But now, many methods have been invented to modify ROC curves to make them hold for validating multi-class classifiers [7–9]. In imprecise environments, ROC curves are particularly useful, because they provide the means for comparing algorithms over a range of operating conditions. However, there is not always dominating relationship between two ROC curves under any operation condition, so the area under the ROC curve (AUC) is invented as a summary of ROC curve. And now it has become a standard measure in this field, since it is invariant to operating conditions [10].

As class imbalance exerts a serious impact on the performance of classification models, it has been extensively studied recently [11,12]. Liu et al. proposed two novel undersampling based algorithms which are free of the deficiency that many majority class examples are ignored [13]. In addition, class imbalance should be valued in the evaluation of classifiers as well. Fortunately, another interest of ROC curves is that they are insensitive to changes in class distribution, which makes ROC curves more competent in some fields where class imbalance is frequently observed [14,15].

Recently, some scholars noted that instances should be dealt with differently. Guyon et al. proposed to group instances into two categories based on typical vs. informative [16]. Li et al. argued that instances should be grouped into three categories: typical, informative, and noisy [17]. Meler et al. grouped instances into easy ones and hard ones [18]. After these operations are brought into learning algorithm, the dataset size can be reduced without affecting the learning performance. When analyzing the performance of classifiers, researchers also argue that it is necessary to treat instance differently. Turney summarized the cost of misclassification errors in inductive concept learning [19]. McDonald proposed a Mean Subjective Utility (MSU) score to measure a classifiers performance with respect to a given data sample under a known cost structure [20]. Here we would like to demonstrate that the performance analysis of classifiers needs to take the classification difficulty of each instance into account. Assume that there are two classifiers c_1 and c_2 for medical diagnosis, and both of their classification accuracy are 70%. However, if the cases

* Corresponding author at: Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China.

E-mail addresses: xiaolizhang1987@gmail.com (X. Zhang), lxf@jlu.edu.cn (X. Li), fengyc14@mails.jlu.edu.cn (Y. Feng).

http://dx.doi.org/10.1016/j.neucom.2016.02.001
0925-2312/

that misdiagnosed by c_1 are easy and the cases that misdiagnosed by c_2 are difficult, c_1 is not a reliable model. We think c_1 does not obey the natural principle that easy instances are correctly classified with a larger probability comparing to difficult ones. In effect, many instances are very difficult to be classified both for human being and automatic algorithms because they contain confusing features and it may be very difficult to make correct classification in the current feature space. It is acceptable to misclassify these difficult instances. Hence, it is reasonable that c_1 is given a more severe punishment and receives a lower rating than c_2 from performance evaluation measures. Unfortunately, though the issue is very important in the assessment of the performance of classifiers, the existing ROC analysis fails in reflecting this feature. In this paper, we invent a reliable measure for predicting the performance of classifiers which incorporates classification difficulty. The proposed method is an upgrade version of ROC curve and AUC, which are denoted as diROC curve and diAUC, respectively.

The main features of diROC analysis are summarized as follows: (1) the diROC curve is the trace of weighted false positive rate and weighted true positive on a bi-dimensional space. A diROC curve closer to the point (0, 1) implies a better performance of classification; (2) diAUC is within the interval of [0, 1]; (3) diAUC is also invariant to operating conditions; (4) as diROC curve takes classification difficulty into account, it provides comparably high discriminating evaluation.

In the proposed measure, how to get the classification difficulty of each instance is a key point. As nearly none of existing dataset provides such information, we need to predict the information based on features of instances. In the paper, classification applications are classified into two catalogues according to whether the true label (i.e. "gold standard") of each instance is known. Different prediction rules are designed for the two catalogues, based on which instances' classification difficulty can be easily achieved.

The newly established analysis method in our paper may spur new research in machine learning and other related fields in which the classification is needed. First of all, the idea of incorporating the classification difficulty can be exerted on other measures, such as precision and accuracy. In addition, we can redesign the existing classification algorithms which are based on optimizing performance evaluation measures [21–23]. Ling and Zhang reported that AUC provides a more discriminating evaluation than accuracy does, so they proposed an AUC-based Bayesian learning algorithm, in which AUC is maximized [21]. Guvenir and Kurtcephe invented a supervised algorithm, which is called ranking instances by maximizing the area under the ROC curve (RIMARC) [23]. Results in this paper imply that we should design learning algorithms by maximizing diAUC.

The rest of the paper is constructed as follows: first we briefly explain the background of this work in Section 2; then we present a variant of ROC curve which is denoted as the diROC curve in Section 3. The area under diROC curve (diAUC) is also defined in this section; Section 4 lists some properties of the diROC curve and diAUC; in Section 5, we test the diROC curve and diAUC experimentally; finally, conclusions are presented in Section 6.

2. Background

2.1. Classification difficulty of instances

In machine learning field, it is unavoidable that some instances are hard to be classified while some not because of their distribution in feature space. Some instances which are close to the boundary of classes may confuse classifiers, and result in misclassification; but other instances can be correctly classified by most classifiers because they are far from boundaries or near the center of the class. A good classifier should be the one which can correctly classify hard instances and avoid misclassifying easy instances. Formally, the difficulty of each instance is defined as follows.

Definition 1 (*Difficulty of an instance*). Given a dataset $D = \{(\chi_1, l_1), \ldots, (\chi_{m+n}, l_{m+n})\}, \chi_i \in \mathbb{R}^k, l \in \{-1, 1\}$, let $f(\chi)$ be the boundary between the two classes. For $\chi_i \in D$, its distance from the boundary is d_i, then the instance χ_i's difficulty ξ_i is defined as

$$\xi_i \propto 1/d_i. \tag{1}$$

Here, class boundary $f(\chi)$ refers to a surface capable of separating the two classes. For any positive instance $\chi_i, f(\chi_i) > 0$; for any negative instance $\chi_j, f(\chi_j) < 0$.

See such a case as shown in Fig. 1: instances are distributed in an $\mathbb{R}^2$ space, positives are labeled as "+" and negatives are labeled as "−". The line represents the true boundary of the two classes.

As the positive instance A is much closer to the boundary than another positive instance B, the classification difficulty of A is higher than that of B. Hence, if the instance A is classified correctly by a classifier, we should render the classifier a big reward, and in contrast, if the instance B is classified correctly by a classifier, we should grant the classifier a relatively small reward. Meanwhile, if a classifier misclassifies the instance A, we should exert a small punishment on it; if a classifier misclassifies the instance B, we should apply a relatively severe punishment to it. Similarly, the two negative instances C and D differ from each other greatly in terms of the distances from the boundary and consequently their classification difficulties are also different. However, all the existing measures, including AUC, treat each instance equally in the process of assessing classifiers.

Example 1. Assume that there is an artificial test set in which 5 positives and 5 negatives are involved. Both c_1 and c_2 are two-class classifiers. The scores of the ten instances distributed by c_1 and c_2 are listed in Table 1. Along with the scores, the difficulty index ξ of each instance is also provided (it is assigned artificially). ξ is within the interval of [0,1] where 0 means the lowest level of difficulty and 1 means the highest level of difficulty.

From Table 1, we note that the scores predicted by c_1 and c_2 are the same except the sixth and ninth instances. According to the definition of the ROC curve, the two classifiers c_1 and c_2 share the same ROC curve (shown in Fig. 2). Consequently, they also receive the same AUC value of 0.76. However, if taking the classification difficulty into account, we deem that c_1 is superior to c_2. Take the sixth instance as an example, its classification difficulty index (ξ) is 0.3. c_1 deems that the instance belongs to positive instance with the probability of 0.45, while c_2 deems that the instance belongs to positive instance with the probability of 0.80. In this case, the classification difficulty is relatively low, and c_2 commits a more

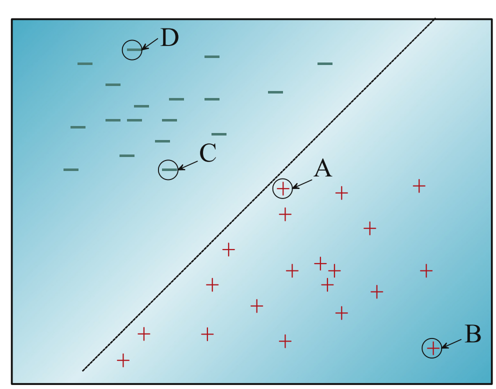

Fig. 1. Instances predicted in $\mathbb{R}^2$ space, each "+" represents a positive and each "−" does a negative.