A Strategy on Selecting Performance Metrics for Classifier Evaluation

Yangguang Liu, Ningbo Institute of Technology, Zhejiang University, Ningbo, China
Yangming Zhou, Ningbo Institute of Technology, Zhejiang University, Ningbo, China
Shiting Wen, Ningbo Institute of Technology, Zhejiang University, Ningbo, China
Chaogang Tang, China University of Mining and Technology, Xuzhou, China

ABSTRACT

The evaluation of classifiers’ performances plays a critical role in construction and selection of classification model. Although many performance metrics have been proposed in machine learning community, no general guidelines are available among practitioners regarding which metric to be selected for evaluating a classifier’s performance. In this paper, we attempt to provide practitioners with a strategy on selecting performance metrics for classifier evaluation. Firstly, the authors investigate seven widely used performance metrics, namely classification accuracy, F-measure, kappa statistic, root mean square error, mean absolute error, the area under the receiver operating curve, and the area under the precision-recall curve. Secondly, the authors resort to using Pearson linear correlation and Spearman rank correlation to analyses the potential relationship among these seven metrics. Experimental results show that these commonly used metrics can be divided into three groups, and all metrics within a given group are highly correlated but less correlated with metrics from different groups.

Keywords	Classifiers, Classifiers’ Performances, Correlation, Machine Learning Community, Performance Metrics

1. INTRODUCTION

The correct selection of performance metrics is one of the most key issues in evaluating classifiers’ performances. A number of performance metrics have been proposed in different application scenarios. For example, accuracy is typically used to measure the percentage of correctly classified test instances. It is so far the primary metric for assessing classifier performance (Ben et al.2007) and (Huang et al.2005); precision and recall metrics are widely applied in information retrieval (Baeza-Yates 1999); medical decision making community prefers the area under the receiver operating characteristic (ROC) curves (i.e., AUC) (Lasko et al. 2005). It is a very common situation, where a classifier performs well on one performance metric but badly on others.

DOI: 10.4018/IJMCMC.2014100102
For example, boosted trees and SVM classifiers achieve good performances on classification accuracy, while they yield poor performances on root mean square error (Caruana et al. 2004).

In general, a widely accepted consensus is to choose performance metrics depending on the practical requirements of specific applications. For example, neural networks typically optimize squared error, and thus the metric of root mean square error can better reflect the actual performance of a classifier than other metrics. However, in some cases, specific criteria are unknown in advance, and practitioners tend to select several measures from widely adopted ones, such as classification accuracy, kappa statistic, F-measure and AUC, for evaluating a new classifier (Sokolova et al. 2006), and (Sokolova et al. 2009). Additionally, most metrics are derived by calculating the confusion matrix of the classifier. It could be reasonable to think that some of such performance metrics are closely related, which may cause redundancy on measuring the performance of classifiers. On the other hand, it is difficult for practitioners to reach concrete conclusion when two metrics provide conflicting results.

This study focuses on providing a strategy on selecting appropriate performance metrics for classifiers by using Pearson linear correlation and Spearman rank correlation to analyses the potential relationship among seven widely used performance metrics, namely accuracy, F-measure, kappa statistic, root mean squared error (i.e., RMSE), mean absolute error (i.e., MAE), AUC, and area under the precision recall (PR) curve (i.e., AUPRC). We first briefly describe these performance metrics. By definition, we sketch out their characteristic features by confusion matrix and preliminarily classify them into three groups, namely threshold metrics, rank metrics, and probability metrics. Then, we use correlation analysis to measure the correlations of these metrics. The experimental results show that metrics from the same group are closely correlated but less correlated with metrics from different groups. Additionally, we compare the correlation changes caused by the size and class distribution of the datasets, which are the main factors affecting measured values.

The main contributions made in this work are summarized as follows. First, we divide these seven performance metrics into three groups by analyzing their definitions. Experimental results confirm that metrics inside the same group have high correlation, and metrics from different groups have low correlation. Second, we also provide practitioners with the following strategies on selecting performance metrics for evaluating a classifier’s performance based on experiment results. For balanced training data sets, one should select multiple metrics to evaluate the classifier, and at least one metric is selected in each group. For imbalanced training data sets, a classifier is not necessary to achieve the optimal performance on all groups of metrics; instead, as long as the classifier meets the performance requirement of an application measured by certain group(s) of metrics, we recommend adopting it regardless of its less satisfactory performance on other groups of metrics.

Compared with existing work, our work concentrates on investigating the relationships among some special popular performance metrics. By the definitions and the experiments, a clear taxonomy of these metrics has been given. It is necessary to note that we are not attempting to give specific performance metrics for the practical applications. Instead, we resort to correlation analysis to discover the potential relationships among several most commonly used performance metrics, and provide practitioners with a more profound understanding of the performance metrics and the strategy on selecting performance metrics for classifier evaluation.

The outline of this paper is as follows. Section II briefly reviews some related work. Seven common performance metrics used in practice are described in Section III. Section IV introduces the correlation analysis methods and explains the detail of the experimental setup. The experimental results and some discussions are provided in Section V. Finally, Section VI gives our conclusions and suggestions for future research.
Mobile Computing and Commerce Framework
www.igi-global.com/chapter/mobile-computing-commerce-framework/26484?camid=4v1a

Content Personalization for Mobile Interfaces
www.igi-global.com/chapter/content-personalization-mobile-interfaces/17062?camid=4v1a