Discrimination-Based Criteria for the Evaluation of Classifiers

Thanh Ha Dang[1,2], Christophe Marsala[1], Bernadette Bouchon-Meunier[1], and Alain Boucher[2]

[1] Université Pierre et Marie Curie - Paris6, CNRS UMR 7606, DAPA, LIP6 8, rue du Capitaine Scott, Paris, F-75015, France
[2] Institut de la Francophonie pour l'Informatique, Equipe MSI, Bât D, 42 rue Ta Quang Buu, Hanoi, Vietnam

Abstract. Evaluating the performance of classifiers is a difficult task in machine learning. Many criteria have been proposed and used in such a process. Each criterion measures some facets of classifiers. However, none is good enough for all cases. In this communication, we justify the use of discrimination measures for evaluating classifiers. The justification is mainly based on a hierarchical model for discrimination measures, which was introduced and used in the induction of decision trees.

1 Introduction

Machine learning techniques become increasingly popular in both academic and industrial domains. In classification problems, the use of such techniques often involves the assessment of how good is a classifier in relation with a dataset. The standard practice is to take a collection of examples in the domain of interest, select randomly a subset of these examples as training set and apply machine learning algorithms to it for obtaining a classifier, which is also named classification model or model for short. Then this one is used to classify the remaining test cases. The performance of classifiers is usually evaluated by the classification results in the test sets.

Most of measures of evaluation are designed with the hypothesis that all examples are equally important and that datasets are distributed in a balanced manner by their classes. In the classical case, each example in the test set is classified in a class. But in the more general ones, probabilistic classification, possibilistic classification and fuzzy classification for instances, several classes may be assigned to an example.

A confusion matrix contains information about actual and predicted classifications done by a classification model. Performance of such model is commonly evaluated using the data in the matrix. Most of measures evaluate the relation between the predicted classes and the actual classes of examples on a test set and do not pay enough attention to the characterization of classification problems. For examples: accuracy, error rate, true/false positive rate, true/false negative rate, sensitivity, specificity, etc. So there is a bias on classification results concerned with the characterization of problems, in particular the distribution of

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 552–563, 2006.

examples as showed by many authors [7, 12]. Sometimes, the classes have very unequal frequency. For instances, in e-commerce: 99% of visitors do not buy anything and only 1% of them buy something; in security, 99.99% of people are not terrorists; etc. The situation is similar with multiple classes. A majority class classifier can have a very high accuracy: 99% in the e-commerce situation, 99.99% in the security problem, but it should be useless.

In the last decade, the ROC curves have been widely used in the machine learning community [2, 6]. An attractive property of the ROC curves is the insensitivity to the class distribution. In many cases, such as rule sets, classifiers produce only a class decision for each example. The use of such a discrete classifier on a test set produces a single confusion matrix. From the matrix, only one point in ROC space is determined. In such cases, classifiers should be converted to generate scores from each example rather than just a class. Moreover, ROC analysis is not convenient for the choice of classifiers. For concluding that a classifier is better than another, the first one should be better than the second one over the whole performance space [11] i.e. it has a higher true positive rate and a lower false positive rate. Several information-based measures [3, 7, 8] which will be recalled in the next section, have been proposed for a similar purpose. They usually try to exclude the effect of prior class probabilities as with the ROC analysis.

In classification process, it can be pointed out two partitions of the test set. The first one is natural: examples are partitioned by their classes. The second one is the partition raised by a classifier. In this paper, we propose to consider the adequation between these two partitions. This allows an evaluation of the discrimination power of classification models with regard to the classes and a comparaison of classifiers. The initial idea was introduced for the induction of decision trees process [13] which helps to select the most adequate attribute for partitioning a learning set. The selected attribute is the most discriminated one with regard to the classes.

The paper is organized as follows. In Section 2, several criteria for classifier evaluation, in particular the information-based criteria are presented and formalized with a common formalism. In Section 3, a hierarchical model for measures of discrimination in inductive learning is recalled. In Section 4, the discrimination-based criteria are introduced based on the hierarchical model. In Section 5, several properties of the proposed criteria are presented. In Section 6, a set of experiments is done to illustrate and validate the use of these criteria. In the last section, a conclusion is done and future work is proposed.

2 Criteria for Classifier Evaluation

There are many criteria of performance for a classification model. We cite here a list of some usual criteria : accuracy, error rate, true/false positive rate, true/false negative rate, sensitivity, specificity, positive/negative predictive value, fallout, precision/recall, F measure, ROC convex hull, area under the ROC curve, calibration error, mean cross-entropy, root mean squared error, expected cost,