Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*

Marina Sokolova[1], Nathalie Japkowicz[2], and Stan Szpakowicz[3]

[1] DIRO, Université de Montréal, Montreal, Canada
sokolovm@iro.umontreal.ca
[2] SITE, University of Ottawa, Ottawa, Canada
nat@site.uottawa.ca
[3] SITE, University of Ottawa, Ottawa, Canada
ICS, Polish Academy of Sciences, Warsaw, Poland
szpak@site.uottawa.ca

Abstract. Different evaluation measures assess different characteristics of machine learning algorithms. The empirical evaluation of algorithms and classifiers is a matter of on-going debate among researchers. Most measures in use today focus on a classifier's ability to identify classes correctly. We note other useful properties, such as failure avoidance or class discrimination, and we suggest measures to evaluate such properties. These measures – Youden's index, likelihood, Discriminant power – are used in medical diagnosis. We show that they are interrelated, and we apply them to a case study from the field of electronic negotiations. We also list other learning problems which may benefit from the application of these measures.

1 Introduction

Supervised Machine Learning (ML) has several ways of evaluating the performance of learning algorithms and the classifiers they produce. Measures of the quality of classification are built from a confusion matrix which records correctly and incorrectly recognized examples for each class. Table 1 presents a confusion matrix for binary classification, where tp are true positive, fp – false positive, fn – false negative, and tn – true negative counts.

Table 1. A confusion matrix for binary classification

Class \ Recognized	as Positive	as Negative
Positive	tp	fn
Negative	fp	tn

This paper argues that the measures commonly used now (accuracy, precision, recall, F-Score and ROC Analysis) do not fully meet the needs of learning problems in which

* We did this work while the first author was at the University of Ottawa. Partial support came from the Natural Sciences and Engineering Research Council of Canada.

A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 1015–1021, 2006.

the classes are *equally important* and where *several algorithms are compared*. Our findings agree with those of [1] who surveys the comparison of algorithms on multiple data sets. His survey, based on the papers published at the International Conferences on ML 2003–2004, notes that algorithms are mainly compared on accuracy.

2 Commonly-accepted Performance Evaluation Measures

The vast majority of ML research focus on the settings where the examples are assumed to be identically and independently distributed (IID). This is the case we focus on in this study. Classification performance without focussing on a class is the most general way of comparing algorithms. It does not favour any particular application. The introduction of a new learning problem inevitably concentrates on its domain but omits a detailed analysis. Thus, the most used empirical measure, *accuracy*, does not distinguish between the number of correct labels of different classes:

$$accuracy = \frac{tp \ + \ tn}{tp \ + \ fp \ + \ fn \ + tn} \tag{1}$$

Conversely, two measures that separately estimate a classifier's performance on different classes are sensitivity and specificity (often employed in biomedical and medical applications, and in studies which involve image and visual data):

$$sensitivity = \frac{tp}{tp \ + \ fn}; specificity = \frac{tn}{fp \ + tn} \tag{2}$$

Focus on one class prevails in text classification, information extraction, natural language processing and bioinformatics, where the number of examples belonging to one class is often substantially lower than the overall number of examples. The experimental setting is as follows: within a set of classes there is a class of special interest (usually *positive*). Other classes are either left as is – multi-class classification – or combined into one – binary classification. The measures of choice calculated on the positive class[1] are:

$$precision = \frac{tp}{tp \ + fp}; recall = \frac{tp}{tp \ + \ fn} = sensitivity \tag{3}$$

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4}$$

All three measures distinguish the correct classification of labels within different classes. They concentrate on one class (positive examples). Recall is a function of its correctly classified examples (true positives) and its misclassified examples (false negatives). Precision is a function of true positives and examples misclassified as positives (false positives). The F-score is evenly balanced when $\beta = 1$. It favours precision when $\beta > 1$, and recall otherwise.

A comprehensive evaluation of classifier performance can be obtained by the ROC:

$$ROC = \frac{P(x|positive)}{P(x|negative)} \tag{5}$$

[1] The same measures can be calculated for a negative class, but they are not reported.