L'incohérence de l'*aire sous la courbe ROC*, que faire à ce propos?

David J. Hand
Imperial College London
d.j.hand@imperial.ac.uk

Résumé : Différents critères sont largement utilisés pour évaluer la performance de règles de classement. L'un d'eux est l'aire sous la courbe ROC (*AUC* : the *area under the curve*). Cette mesure a l'agréable propriété de synthétiser la performance pour tous les seuils de classement possibles. Malheureusement, au cœur de l'*AUC* se trouve une distribution qui dépend de l'outil de classification dont la performance est évaluée, si bien que les estimations qui utilisent cette mesure sont fondamentalement incohérentes: c'est-à-dire qu'aucune comparaison ne peut être faite quand l'*AUC* est utilisé. Cette incohérence est examinée, ses implications sont présentées, et une alternative cohérente est décrite.

Mots clés : Discrimination, aire sous la courbe ROC, performance de la discrimination, la mesure *H*

Abstract: Various different criteria are in widespread use for evaluating the performance of classification rules. One of these is the area under the ROC curve (the *AUC*). This measure has the attractive property that it summarises performance over all possible values of the classification threshold. Unfortunately, at the heart of the *AUC* lies a distribution which depends on the classifier being evaluated, so that evaluations using this measure are fundamentally incoherent: like is not being compared with like when the *AUC* is used. This incoherence is explored, its implications noted, and a coherent alternative is described

Keywords: Classification, area under the curve, ROC curve, classifier performance, *H* measure

1. Introduction

La Classification supervisée intervient dans de nombreuses activités : le diagnostic médical, le dépistage épidémiologique, la sélection de bons crédits, la reconnaissance de la parole, la détection de fraudes et d'erreurs, le classement de personnel employé, et dans une foule d'autres applications. Ces problèmes ont tous la même structure : étant donné un ensemble d'objets, dont chacun est connu par son appartenance à une classe et décrit par un ensemble de mesures, construire une règle qui permette d'assigner un nouvel objet à une seule classe sur la base du vecteur de ses mesures. Du fait que ces problèmes sont largement répandus, ils ont donné lieu à des investigations dans plusieurs disciplines différentes (bien qu'imbriquées), incluant les statistiques, la reconnaissance des formes, l'apprentissage machine (*machine learning*), l'exploration de données (*data mining*), et un grand nombre de techniques ont été développées (voir par exemple Hand, 1997; Hastie et al, 2001; et Webb, 2002). L'existence de ces approches variées pose la question de comment choisir entre elles. C'est-à-dire, étant donné un problème de classification, parmi de nombreux outils de classement possibles lequel faut-il adopter?

Il est répondu à cette question en évaluant les outils et en en choisissant un qui semble performant. Pour ce faire bien sûr, un critère d'évaluation adapté est nécessaire – un moyen d'estimer la performance de chaque règle. Malheureusement, la performance comporte plusieurs aspects. A un haut niveau, on devrait s'intéresser à des questions telles que : Avec quelle rapidité peut-on