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

construire la règle ? A quelle vitesse cette règle fournit-elle les classements ? Peut-on traiter de grands ensembles de données et/ou de très nombreuses variables caractéristiques ? Gère-t-elle bien les données manquantes ? Et ainsi de suite… A un niveau plus immédiat, on sera généralement intéressé à la capacité de la règle à assigner de nouveaux objets à la bonne classe : l'exactitude du classement. Cet article explore comment mesurer l'exactitude du classement. En particulier, il montre qu'une mesure très largement utilisée, l'aire sous la courbe ROC (*AUC*), a une imperfection fondamentale faisant partie intégrante de sa définition. La conséquence en est des choix sous-optimaux de règle de classement dans des domaines variés, avec des implications malheureuses de mauvais classements, mauvais diagnostics, mauvaises identifications. La nature de cette imperfection fondamentale est explorée et une alternative rationnelle est présentée. Une présentation plus détaillée sur le plan mathématique est donnée dans Hand (2009).

Par commodité, dans ce papier nous supposons qu'il y a juste deux classes, dénommées classe 0 et classe 1.

2. AUC

Une règle de classification consiste en une fonction reliant le vecteur de variables descriptives de chaque objet à son score, *s*, et grâce à un seuil de classement, *t*, tel que les objets ayant un score $s>t$ sont assignés à la classe 1, et sinon à la classe 0. Malheureusement les règles de classification sont rarement parfaites : la complexité des problèmes réels, et le nombre limité de données réelles, impliquent que les règles de classement affectent certains objets à une classe incorrecte. Les différentes règles, correspondant à différentes fonctions et différents seuils de classement *t*, affectent de manière erronée des objets variés et en nombre divers, et le but de l'estimation de la performance est de mesurer l'importance des mauvais classements.

Pour estimer la performance d'une règle, celle-ci est appliquée à un ensemble d'objets dont la classe d'appartenance est connue. Pour des raisons bien compréhensibles, cet ensemble ne doit pas être le même que celui utilisé pour construire la règle (c'est-à-dire pour estimer ses paramètres), puisqu'une évaluation de performance basée sur cet ensemble mènerait probablement à une estimation optimiste de la future performance. .
On peut traiter ce problème en utilisant un « ensemble test » d'objets, ou en pratiquant des méthodes sophistiquées de ré-échantillonnage, méthodes *leave-one-out* ou *bootstrap* – pour une discussion voir par exemple Hand (1997). Ici, on suppose qu'une telle méthode a été appliquée, et nous ne la discuterons pas plus avant.

En appliquant la règle à un ensemble d'objets utilisés pour l'évaluation, on obtient un ensemble de scores pour les objets de la classe 0 et un autre pour les objets de la classe 1. Ainsi nous avons des distributions empiriques de scores pour chacune des classes. Dans ce qui suit nous pouvons supposer que l'ensemble consacré à cette évaluation est suffisamment grand pour que nous puissions ignorer les fluctuations d'échantillonnage et traiter ces deux distributions de scores, $f_0(s)$ et $f_1(s)$ comme les distributions réelles sur ces deux classes. Si on voulait construire des intervalles de confiance, ou faire des tests de significativité ou des tests d'hypothèse, il faudrait alors prendre en compte que les distributions empiriques sont en fait tout simplement des estimateurs des vraies distributions $f_0(s)$ et $f_1(s)$.

Quand ces distributions de scores sont comparées au seuil t, une proportion $1-F_0(t)$ des objets de la classe 0 sont mal classés, et une proportion $F_1(t)$ des objets de la classe 1 sont mal classés, où $F_0(t)$ et $F_1(t)$ sont les distributions cumulées correspondant à $f_0(s)$ et $f_1(s)$. Chacune de ces proportions est un indicateur de l'efficacité de l'outil. Mais chacun, en lui-même, échoue à saisir la

totalité détaillée de la performance. En particulier en accroissant le seuil *t*, on peut rendre la proportion de mauvais classements de la classe 0 aussi petite que l'on veut, mais c'est au prix de l'augmentation de la proportion de mauvais classements des objets de la classe 1. Une combinaison des deux est requise pour avoir une mesure globale de performance.

Une façon bien connue de représenter cette situation se fait via une courbe ROC (*receiver operating characteristic*). C'est un graphique où $F_1(t)$ est représenté sur l'axe vertical et $1-F_0(t)$ sur l'axe horizontal (toutefois occasionnellement d'autres représentations axiales sont utilisées). Un outil donné va permettre de tracer une courbe du point $(0,0)$ au point $(1,1)$ quand *t* décroît. Une telle courbe représente la performance de l'outil, pour les deux taux de mauvais classements, pour n'importe quel choix du seuil *t*. Il s'en suit immédiatement que par définition la courbe ROC est monotone non décroissante de $(0,0)$ à $(1,1)$. En outre, une séparation parfaite entre les deux distributions, telle que la classification parfaite entre les deux classes pourrait être atteinte pour un seuil approprié *t*, correspond à une courbe ROC qui passe par le point $(0,1)$. En général, pour n'importe quelle valeur de $1-F_0(t)$ sur l'axe horizontal, une valeur plus élevée de $F_1(t)$ correspond à une meilleure règle de classification. Cela signifie que les meilleures règles ont des courbes ROC proches du coin supérieur gauche $(0,1)$ du carré ROC. Cette observation a conduit à un résumé synthétique de la performance de la règle de classification : l'aire en dessous de la courbe, *AUC*. Plus élevé est l'indicateur *AUC*, meilleure est la règle. Les courbes ROC, leurs propriétés et applications, sont présentées dans leurs grandes lignes par Krzanowski et Hand (2009).

Pour la facilité de l'exposé, dans ce qui suit, nous supposerons que $F_0(t)$ domine stochastiquement $F_1(t)$, que $F_0(t)$ et $F_1(t)$ sont partout différentiables, et que le ratio $f_1(t)/f_0(t)$ est monotone croissant. Ses hypothèses n'imposent aucune contrainte sur le développement théorique, mais elles nous permettent simplement d'ignorer les anomalies. Une discussion complète sur comment traiter les situations quand ces hypothèses ne s'appliquent pas est donnée dans Hand (2009).

Si la proportion d'objets qui appartiennent effectivement à la classe 0 est π, alors on peut combiner $1-F_0(t)$ et $F_1(t)$ grâce à

$$\pi\left(1-F_0(t)\right)+\left(1-\pi\right)F_1(t),$$

pour représenter la proportion globale d'objets mal classés (le taux d'erreur de classement). C'est une mesure très largement utilisée, mais intégrant implicitement une hypothèse – laquelle, à mon avis, est très rarement appropriée : c'est-à-dire que les deux types d'affectation erronée, classer des objets de la classe 0 en classe1, et classer des objets de la classe 1 en classe 0, sont d'égale sévérité. Si on relâche cette hypothèse, on généralise en attribuant un coût de mauvais classement c_0 aux objets de la classe 0, et un coût de mauvais classements c_1 aux objets de la classe 1.
La perte globale dut au mauvaise classement est alors :

$$Q(t;c_0,c_1)=\pi c_0\left(1-F_0(t)\right)+\left(1-\pi\right)c_1F_1(t). \qquad (1)$$

Cette perte est une fonction du couple spécifié des coûts (c_0,c_1), ainsi que du seuil choisi *t*.

Pour un couple de coûts, il est rationnel de choisir *t* de façon à minimiser la perte – tout autre choix impliquerait irrationnellement un coût de mauvais classement plus élevé que nécessaire. Appelons ce choix du seuil de classement *T*. Grâce aux hypothèses mathématiques sur $f_0(s)$ et $f_1(s)$, à partir de (1) un petit calcul montre que

$$\pi c_0 f_0(T) = (1-\pi) c_1 f_1(T), \qquad (2)$$

d'où la relation entre le ratio $r = c_0/c_1$ et le seuil de classement *T*:

$$r = \frac{c_0}{c_1} = \frac{(1-\pi)}{\pi} \frac{f_1(T)}{f_0(T)} \square\ R_0(T). \qquad (3)$$

Tout cela est très bien si l'on peut décider quel couple des coûts (c_0, c_1) est approprié. Pourtant, c'est souvent difficile. Par exemple, une situation typique survient quand la règle de classification doit être appliquée dans le futur, dans des circonstances qui ne peuvent être entièrement prévues. Un système de diagnostic médical peut être appliqué à différentes populations, dans différentes cliniques ou même dans différents pays, où les conséquences de deux sortes de mauvais classement changent d'une population à l'autre. Un *credit scoring*, prévoyant quel emprunteur va probablement faire défaut, peut avoir un couple de coûts d'erreur de classement qui se modifie en fonction du contexte économique. Un système de reconnaissance de visage pour la détection de suspects de terrorisme dans les aéroports, peut avoir des coûts d'erreur de classement dépendant de l'intensité de la menace. Et ainsi de suite. Changer de coûts signifie changer de seuil optimal *T*.

Si les coûts sont difficiles à déterminer, on devrait pouvoir envisager des distributions des valeurs futures de (c_0, c_1), et, de là, la distribution de *T* (grâce à la relation (3)).

Dans ce cas, une mesure adaptée de performance est obtenue en intégrant $Q(T(c_0, c_1); c_0, c_1)$ sur la distribution des valeurs de (c_0, c_1), où nous avons explicité la dépendance du T optimal à (c_0, c_1). Si $g(c_0, c_1)$ est la distribution conjointe des deux coûts d'erreur de classement, alors la mesure de la performance est

$$L = \int\int \left\{ \pi c_0 \left(1 - F_0\left(T(c_0, c_1)\right)\right) + (1-\pi) c_1 F_1\left(T(c_0, c_1)\right) \right\} g(c_0, c_1)\, dc_0 dc_1. \qquad (4)$$

Cependant nous avons vu qu'en fait d'après (3) T dépend du ratio $r = c_0/c_1$, si bien qu'on peut réécrire (4) comme suit:

$$L = \int\int \left\{ \pi r c_1 \left(1 - F_0\left(T(r)\right)\right) + (1-\pi) c_1 F_1\left(T(r)\right) \right\} h(r, c_1)\, dr dc_1, \qquad (5)$$

où *h* prend en compte le Jacobien de la transformation de (c_0, c_1) en (r, c_1).

L'équation (5) peut être réécrite :

$$\begin{aligned} L &= \int \left\{ \pi r \left(1 - F_0\left(T(r)\right)\right) + (1-\pi) F_1\left(T(r)\right) \right\} \int c_1 h(r, c_1)\, dc_1 dr \\ &= \int \left\{ \pi r(T)\left(1 - F_0(T)\right) + (1-\pi) F_1(T) \right\} w(T)\, dT \end{aligned}$$

$$= (1-\pi) \int \left\{ f_1(T)\left(1 - F_0(T)\right) + f_0(T) F_1(T) \right\} \frac{w(T)}{f_0(T)} dT \qquad (6)$$

où $w(T)$ inclut $\int c_1 h(r, c_1)\, dc_1$ et le Jacobien de la transformation de *r* en *T*, et où la dernière ligne utilise (3).

Supposons, maintenant, que nous choisissons la fonction $w(T)$ telle que $w(T) = f_0(T)$. Alors (6) devient

$$L = (1-\pi)\left\{1 - 2\int F_0(T)\, f_1(T)\, dt\right\}.$$

On voit aisément que $\int F_0(T)\, f_1(T)\, dt$ est l'aire sous la courbe ROC, l'AUC. Donc la perte calculée par (6) peut s'exprimer comme

$$L = (1-\pi)\{1-2AUC\},$$

et

$$AUC = \frac{1}{2}\left(1-\frac{L}{(1-\pi)}\right).$$

Ceci montre que l'*AUC* est linéairement lié à l'espérance de la perte due aux mauvais classements si le seuil *T e*st choisi aléatoirement selon la distribution $f_0(T)$.

Utilisant (3), on voit que choisir *T* selon la distribution $w(T) = f_0(T)$ est équivalent à choisir *r* selon la distribution

$$u(r) = \frac{f_0(T)^3}{f_0(T).df_1(T)/dT - f_1(T).df_0(T)/dT},$$

Avec $T = R_0^{-1}(r)$. Ce qui importe ici n'est pas le détail de la distribution $u(r)$ mais le fait qu'elle dépend des distributions empiriques f_0 et f_1. Cette dépendance signifie que différents outils de classification seront évalués en calculant une perte espérée pour différentes distributions de *r*. C'est un nonsense. Le choix de la distribution de probabilité du ratio des coûts de mauvais classements ne peut dépendre de l'outil de classification testé, mais doit être basé sur des propriétés distinctes de celles des outils. Avoir une distribution de *r* dépendante des distributions empiriques équivaudrait par exemple à dire que classer comme bien portant des personnes atteintes de grippe aviaire (contre lequel ils ne seront pas traités) est dix fois plus grave que l'inverse dans le cas où (mettons) on utilise une régression logistique, mais cent fois plus risqué si on utilise un arbre de classification. La gravité relative des mauvais classements ne peut dépendre des outils choisis pour construire la discrimination.

3. Une alternative cohérente à l'AUC

Le problème de l'*AUC* provient de ce que l'espérance de la perte due aux mauvais classements est calculée en utilisant une distribution des coûts d'erreur de classement dépendant des distributions empiriques des scores. Il s'en suit que pour surmonter cette difficulté nous devons sommer en utilisant une distribution indépendante des distributions empiriques des scores. Bien sûr il y a une multitude de choix possibles. Ceux-ci correspondent à différents aspects de la performance de l'outil de classification.

Il n'y a pas de façon d'éliminer cet arbitrage intrinsèque dans la mesure de performance – différents chercheurs peuvent être intéressés par des aspects différents de la performance. Cependant, choisir une distribution indépendante des distributions empiriques signifie que les comparaisons entre outils de classification seront basées sur la même métrique. S'il est compréhensible que des chercheurs différents choisissent des métriques différentes, il est déraisonnable qu'un même chercheur choisisse des métriques différentes pour des outils différents alors que son but est de comparer leurs performances respectives.

Ayant cela à l'esprit, il y a deux types d'approches pour choisir une distribution que nous recommandons d'utiliser en parallèle.

La première approche est basée sur le fait que souvent, alors qu'on peut avoir des difficultés à produire une distribution globale des différentes valeurs du ratio de coûts, on sait

approximativement quelque chose sur ce ratio. Par exemple, dans le cas de la grippe aviaire on peut penser que de mal classer une personne malade (qui donc ne sera pas soignée) est plus grave que l'inverse (où la personne en bonne santé sera soignée alors que ce n'est pas nécessaire). Cela signifierait (la classe 0 étant la classe des personnes saines) que $c_0 < c_1$, si bien que *r* prendrait probablement plus souvent des valeurs inférieures à 1. Cela peut être pris en compte dans le choix de la distribution de *r*.

Par définition, le ratio de coûts considère les deux types de mauvais classements de façon asymétrique. Par commodité, Hand (2009) transforme r en $c = \left(1 + r^{-1}\right)^{-1}$.

c se situe entre 0 et 1, et prend la valeur $1/2$ quand $c_0 = c_1$. Hand (2009) suggère d'utiliser pour *c* une distribution beta:

$$v(c) = beta(c;\alpha,\beta) = c^{\alpha-1}(1-c)^{\beta-1} \Big/ B(1;\alpha,\beta)$$

Si les paramètres α et β sont tous les deux supérieurs à 1, une telle distribution a pour mode $(\alpha-1)/(\alpha+\beta-2)$ ce qui peut être un choix adéquate pour *c* (et le choix de *r* en découle).

Bien sûr, il est tout à fait possible que différents chercheurs fassent des choix de probabilités sur les différentes valeurs de *r* qui soient différenciés, ainsi ils choisissent des distributions béta différentes, et donc obtiennent des mesures de performance différentes pour le même outil discriminant.

La seconde approche résout cette difficulté en proposant un standard universel. En particulier, Hand (2009) suggère que la loi $beta(c;2,2)$ soit utilisée. Elle est symétrique par rapport à $c = 1/2$, et tend vers 0 pour $c = 0$ et $c = 1$.

Si on a connaissance indicative ou mieux assurée sur les valeurs probables de *c* (ou de *r*), alors une distribution béta appropriée peut être utilisée. Chaque chercheur peut alors choisir entre plusieurs règles de classement sur la base de ce qu'ils pensent être les degrés de gravités relatives des coûts de mauvais classements. Cependant pour faciliter les échanges et pour rendre claire pour les autres chercheurs ce qu'est la performance, nous recommandons que l'approche par la loi standard $beta(c;2,2)$ soit toujours présentée.

L'*AUC* prend des valeurs entre 0 et 1, les valeurs élevées correspondant à de bonnes performances. Pour les besoins de l'interprétation nous proposons de faire une simple transformation de la perte moyenne, en la divisant par sa valeur maximum et en lui soustrayant 1, ainsi elle se situera aussi entre 0 et 1, les valeurs élevées indiquant de bonnes performances. Ceci fournit la mesure de performance $H(\alpha,\beta)$.

4. Conclusion

Pour évaluer la performance d'une règle de discrimination, l'*AUC* est une mesure très utilisée. Cet indice apparaît aussi sous d'autres formes – par exemple, en *credit scoring* l'*indice de Gini* qui est une simple transformation de l'*AUC*. C'est une mesure attrayante car tous les chercheurs qui l'appliquent aux mêmes données obtiendront les mêmes résultats. Mais malheureusement, sa définition calcule implicitement la perte moyenne avec une distribution qui dépend de l'outil qu'on est en train d'évaluer. C'est-à-dire que les différents outils de discrimination sont évalués avec différentes métriques. C'est comme si on décidait que Jim est plus grand que Fred parce que Jim pèse 70 000 grammes tandis que Fred pèse 70 kg et que 70 000 > 70. Quand on utilise l'*AUC* pour choisir entre les règles discriminantes, la conséquence est que les comparaisons ne sont pas fiables. Ceci peut conduire à des choix sous optimaux de règles, et il survient des mauvais classements qui

auraient pu être évités. Les conséquences peuvent être très sérieuses dans certaines applications, et couter très cher dans d'autres.

Ce problème est résolu en calculant la perte moyenne selon une distribution qui est la même pour tous les outils de classification. Cette distribution doit refléter ce que pense le chercheur de la probabilité des degrés relatifs de gravité pour divers types de mauvais classements. Quand un chercheur adopte une telle distribution les comparaisons entre outils de classification sont valides.

Bien sûr, différents chercheurs adopteront des distributions différentes. Cela convient parfaitement – différents chercheurs peuvent apprécier différemment la gravité relative des différents types de mauvais classements. Mais cela ne veut pas dire que différents chercheurs évaluant le même outil arriveraient à des conclusions différentes. Pour cette raison, nous proposons qu'une distribution standard universelle soit également utilisée et que ses résultats soient présentés à coté de ceux provenant de n'importe quelle distribution particulière que le chercheur aura choisie. Ceci conduit à la mesure *H*.

Références

Hand D.J. (1997) *Construction and Assessment of Classification Rules*. Chichester: Wiley.

Hand D.J. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. To appear in *Machine Learning*.

Hastie T., Tibshirani R., and Friedman J (2001) *The Elements of Statistical Learning*. New York: Springer.

Krzanowski W.J. and Hand D.J. (2009) *ROC Curves for Continuous Data*. London: Chapman and Hall.

Webb A. (2002) *Statistical Pattern Recognition*, (2nd ed.) Chichester: Wiley.