Université de Montréal

Distributions d'auto-amorçage exactes ponctuelles des courbes ROC et des courbes de coûts

Par
David Gadoury

Département de mathématiques et de statistiques
Faculté des arts et des sciences

Mémoire présenté à la faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en mathématiques/actuariat

Avril 2009

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé :

Distributions d'auto-amorçage exactes ponctuelles des courbes ROC et des courbes de coûts

présenté par :

David Gadoury

a été évalué par un jury composé des personnes suivantes :

Jean-François Angers
président-rapporteur

Charles Dugas
directeur de recherche

David Haziza
membre du jury

Résumé

La classification binaire est une tâche à multiples applications. Le choix du classifieur est donc crucial et les courbes ROC sont une des méthodes de prédilection afin d'y parvenir. Les courbes de coûts ont aussi été introduites récemment comme alternative, ou complément, aux courbes ROC. Il est important pour ces deux types de courbes, de calculer des intervalles de confiance afin de déterminer l'efficacité de l'évaluation de la performance des classifieurs. Calculer les intervalles de confiance pour la différence de performance entre deux classifieurs permet de déterminer si un classifieur performe significativement mieux qu'un autre.

Une procédure simple afin d'obtenir des tels intervalles de confiance est de rééchantillonner les données à l'aide de l'auto-amorçage. Dans ce mémoire, nous obtenons la distribution d'auto-amorçage exacte ponctuelle des courbes ROC et des courbes de coûts et utilisons ces distributions afin d'en trouver les intervalles de confiance. Deux approches sont considérées : l'auto-amorçage stratifié qui rééchantillonne indépendamment les instances positives et négatives, ainsi que l'auto-amorçage complet qui rééchantillonne les données globalement. Dans les deux cas, l'auto-amorçage est calculé à l'aide du moyennage vertical et par seuil.

La performance des intervalles de confiance est mesurée en terme de précision de couverture. Les simulations présentent d'excellents résultats.

Mots clés : Fonction d'efficacité de l'observateur. Auto-amorçage. Probabilité de couverture. Sélection de modèle. Courbe de coûts.

Abstract

Binary classification is a task with numerous applications. The choice of the classifier is therefore crucial. ROC curves are one of the main methods in that matter. Cost curves have also been introduced as an alternative, or complement, to ROC curves. It is of importance to both types of curves, that we be able to compute confidence intervals so that reliability of a classifier's performance can be assessed. Computing confidence intervals for the difference in performance of two classifiers allows us to determine if a classifier performs significantly better then another.

A simple procedure to obtain such intervals is to perform bootstrap resampling of the test set. In this master, we derive the pointwise exact bootstrap distributions of ROC and cost curves. We use these distributions in order to find their confidence intervals. Two approaches are presented: stratified bootstrap, which resamples the negative and positive instances independently, and the full bootstrap, which resamples the test set as a whole. In both cases, the bootstrap is computed using vertical and threshold averaging.

Performance of the confidence intervals is measured in term of coverage accuracy. Simulations show excellent results.

Keywords: Receiver operating characteristics. Bootstrap. Coverage probabilities. Model selection. Cost curves.

Table des matières

Table des figures

Table des tableaux

Liste des sigles et abréviations

AUC	Aire sous la courbe ROC
AUCIV	AUC à instances variables
FN	Faux négatifs
FP	Faux positifs
ROC	Fonction d'efficacité de l'observateur
ROCIV	Courbe ROC à instances variables
TN	Vrais négatifs
TP	Vrais positifs

Remerciements

Je voudrais, en premier lieu, remercier mon directeur de recherche, Charles Dugas, sans qui ce mémoire n'aurait jamais vu le jour. Je tiens à le remercier pour son soutient qui fut tout autant physique, psychologique et financier. Je le remercie pour ses conseils, ses suggestions, et sa compréhension. Je le remercie aussi de l'aide qu'il m'a fournie pour l'obtention d'une charge de cours qui, en complément avec la bourse qu'il m'octroyait, m'a permis de soutenir un niveau de vie plus qu'acceptable pour un étudiant.

Je voudrais également remercier mes collègues étudiants qui ont passés bien des soirées à étudier dans les locaux de la bibliothèque et qui m'ont si souvent rappelé que j'avais besoin d'une pause. Un merci particulier a Luis qui fut mon partenaire d'études dans la majorité de mes cours de deuxième cycle.

Je remercie aussi mes colocataires et ma copine, Annie, avec qui j'ai partagé mes inquiétudes et mes soucis concernant mon avenir tout au long de ces deux années. Je n'y serais pas arriver sans leur support moral.

Je voudrais finalement remercier mes parents pour leur soutient moral et financier. Ils ont toujours été là pour moi, même lorsque mes choix ne leur plaisaient pas et je leur suis éternellement reconnaissant.

CHAPITRE 1

Introduction

La classification binaire est un procédé qui consiste à choisir l'état d'un objet, dit l'instance, entre deux options, dites les classes. Il s'agit d'une méthode présente dans plusieurs milieux. Quelques exemples de tels milieux seraient : un diagnostique dans le domaine médical, un système de détection de fraude bancaire ou fiscale, l'identification de pièces défectueuses dans une chaîne de montage, etc. Pour traiter ce genre de cas, la communauté d'apprentissage machine (*machine learning*) a comme approche habituelle de développer un modèle de prédiction en utilisant l'information disponible sur un jeu de données, dit le jeu d'entraînement (*training set*) dont les vraies classifications sont connues. Généralement, le modèle traitera une instance et en tirera un score numérique lié à la probabilité qu'elle soit positive (une tumeur cancéreuse, une transaction frauduleuse, une pièce défectueuse).

Lorsque plusieurs modèles de prédiction sont créés, l'un d'entre eux devra être choisi afin d'être éventuellement appliqué sur le terrain. Ce procédé est connu sous le nom de sélection de modèle. Généralement, le choix se fait en comparant la performance des prédictions de chaque modèle sur un jeu de données test n'ayant pas servi à entraîner les modèles, c'est-à-dire à en estimer les paramètres. Il existe plusieurs méthodes afin de comparer la performance prédictive de modèles, les courbes ROC sont l'une d'entre elles et c'est donc à cette étape dans le processus d'implantation d'un modèle qu'elles interviennent. Les courbes de coûts sont une autre méthode de comparaison de performances considérant l'effet du coût sur les types d'erreurs. Ce mémoire traitera donc de ces deux derniers éléments.

Bien que les courbes ROC soient fréquemment utilisées, entre autre dans le domaine de l'apprentissage machine, et que les courbes de coûts soient une nouvelle alternative très intéressante, il n'existe pas de méthode reconnue permettant de construire des intervalles de confiance. Il est bien connu que l'intervalle de confiance est un outil indispensable afin de valider l'efficacité d'un test. Pour répondre à cette lacune, nous proposons une approche d'auto-amorçage (*bootstrap*) exacte afin de déterminer les distributions de ces deux types de courbes. Nous pourrons ainsi générer de tels intervalles de confiance. Les tests effectués sur les distributions obtenues à l'aide de l'approche d'auto-amorçage sont concluants bien qu'ayant quelques faiblesses lorsque la situation est extrême, c'est-à-dire lorsque nous considérons des cas où le nombre d'instances positives ou celui d'instance négatives est trop près de 0.

La suite de ce mémoire comportera quatre parties. Les deux premières exposeront la théorie pour les courbes ROC et les courbes de coûts, respectivement. Les suivantes présenteront notre théorie sur les distributions d'auto-amorçage exactes, ainsi que les expériences que nous avons effectuées afin de valider la précision des intervalles de confiance tirés de ces distributions.

CHAPITRE 2

Courbes ROC

2.1 Définition et historique

Le terme ROC signifie « fonction d'efficacité de l'observateur » (*Receiver Operating Characteristics*). Une courbe ROC est un outil de visualisation, d'organisation et de sélection de classifieurs basé sur leur performance, c'est-à-dire leur capacité à départager les différents types d'instances. Une courbe ROC est une représentation graphique de la *sensibilité* (*sensitivity*), c'est-à-dire la capacité de détecter des instances positives, par rapport à 1 – *spécificité*, c'est-à-dire 1 – la capacité de détecter des instances négatives, pour un classifieur binaire.

Les courbes ROC furent utilisées pour la première fois pendant la Seconde Guerre mondiale [Green et Swets, 1966]. Suite aux événements de Pearl Harbor en 1941, l'armée américaine a initié les recherches sur les courbes ROC dans le but de prédire plus adéquatement si ce que détectait leurs radars étaient bien des avions japonais.

Après la guerre, elles furent appliquées directement à la théorie de détection de signaux (*signal detection theory*), principalement afin de représenter la relation entre le ratio de succès (*hit rate*) et le ratio de fausse alerte d'un classifieur [Egan, 1975; Swets et al., 2000].

Le spectre d'applications des courbes ROC a été élargi en touchant à la psychologie vers 1950, puis à la médecine en général. Dans ces domaines, on utilise la théorie ROC afin de visualiser et d'analyser le comportement de systèmes de

diagnostique [Swets, 1988] ou simplement de tester ces mêmes diagnostiques [Zhou et al., 2002].

Plus récemment, la théorie ROC s'est avérée un outil très performant dans les domaines de l'apprentissage machine et du forage de données (*data mining*). L'approche standard est de développer un modèle de prédiction en utilisant l'information disponible sur un jeu de données pour lesquelles la vraie classification est déjà connue [Fawcett 2006b]. La première application de la théorie ROC dans ces domaines fut une démonstration de l'efficacité de la courbe ROC afin de comparer et d'évaluer différents algorithmes de classification [Spackman, 1989]. Cette application est celle qui nous intéressera le plus tout au long de ce mémoire.

2.2 Matrice de confusion

Considérons le scénario suivant : soit un classifieur binaire C et un événement E. Par exemple, dans le domaine médical, E pourrait être l'état de santé d'un patient et C pourrait être un test de dépistage afin de déterminer si le patient E est atteint d'un cancer (classe positive) ou non (classe négative). Dans ce scénario (ainsi que pour tout autre cas de classifieur binaire), nous nous retrouvons avec quatre conclusions possibles :

1) Le patient est atteint du cancer et le test le dépiste : un vrai positif (TP, le T provient de l'anglais « *true* » qui signifie vrai)
2) Le patient est en bonne santé et le test ne détecte pas de cancer : un vrai négatif (TN).
3) Le patient est atteint du cancer mais le test ne l'indique pas : un faux négatif (FN).
4) Le patient est en bonne santé mais le test dépiste un cancer : un faux positif (FP).

On peut représenter ces situations par ce qu'on appelle une matrice de confusion (aussi appelée un tableau de contingence). La figure 2.1 présente un exemple d'une telle matrice.

Si on appliquait le classifieur à un jeu de données (dans notre exemple plusieurs patients), on pourrait remplir la matrice de confusion en inscrivant le nombre d'occurrences de chaque conclusion dans la case appropriée.

		Événement E réel (État de santé du patient)	
		Positif (atteint du cancer)	Négatif (en bonne santé)
Classification hypothétique (Résultat du test de dépistage)	+ (cancer dépisté)	Vrais positifs	Faux positifs
	- (Déclaration de santé)	Faux négatifs	Vrais négatifs

Fig. 2.1 Matrice de confusion représentant les quatre scénarios possibles avec entre parenthèses leur interprétation pour l'exemple médical du cancer.

Idéalement, un classifieur parfait nous donnerait 0 dans les deux cases d'erreurs (faux positifs et faux négatifs). Il est rare d'avoir un tel classifieur. En général, si on tente de diminuer la quantité de faux négatifs, le nombre de faux positifs augmentera et vice versa. Pour cette raison, la matrice de confusion d'un classificateur sera obtenue pour un certain seuil d'erreur préalablement choisi. Simplement en regardant la matrice de confusion, il est possible d'avoir une petite idée de la performance d'un classifieur. Il est toutefois possible de tirer plus d'informations de cette matrice. En particulier, on

peut calculer la spécificité et la *sensibilité* du classifieur à partir de la matrice de confusion. Selon Fawcett (2004), voici les formules intéressantes (N représentant la quantité d'instances réellement négatives et P la quantité d'instances réellement positives, TP, FP, TN et FN sont tels que définis précédemment) :

$$spécificité = \frac{TN}{FP + TN} = 1 - \frac{FP}{N}, \qquad exactitude = \frac{TP + TN}{P + N},$$

$$précision = \frac{TP}{TP + FP}, \qquad mesure\ F = \frac{2TP}{TP + FP + P}.$$

On notera donc que 1 - *spécificité* est en fait le ratio de faux positifs (FP/N) et que la *sensibilité* est le ratio de vrais positifs (TP/P).

2.3 L'espace ROC

Nous définissons l'espace ROC comme étant le carré du plan cartésien de sommets (0;0), (0;1), (1;0) et (1;1). En effet, comme les ratios de vrais positifs et de faux positifs sont des valeurs non négatives inférieures ou égales à 1, tout point d'une courbe ROC se trouvera donc à l'intérieur de cet espace.

La figure 2.2 ci-dessous représente un tel espace ROC. Lorsqu'on appliquera un classifieur à un jeu de données, on obtiendra donc une matrice de confusion de laquelle on tirera les ratios de vrais et de faux positifs. Cela résultera donc en un point sur l'espace ROC, à partir duquel on pourra évaluer la performance du classifieur. Puisque le ratio de vrais positifs représente la probabilité d'avoir classé correctement une valeur positive et le ratio de faux positifs celle d'avoir mal classé une valeur négative, on espère que la première valeur sera beaucoup plus grande que la deuxième. Idéalement, un point ROC se trouvera donc le plus près possible du coin supérieur gauche.

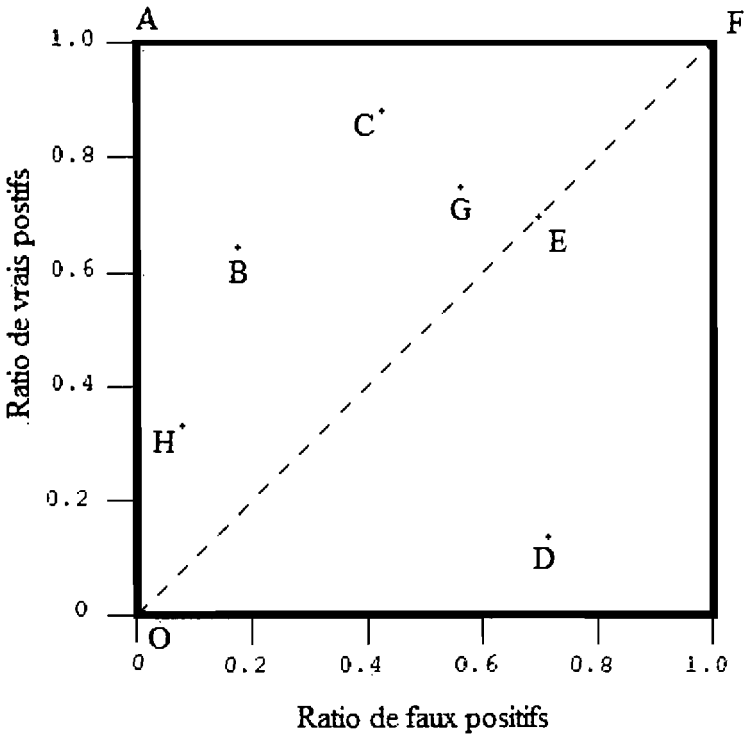

Fig. 2.2 Espace ROC avec 9 classifieurs

Plus explicitement, le point A (0;1) de la figure 2.2 représente donc un scénario parfait où le classifieur détecte la totalité des positifs et ne se trompe jamais. De même, le point O (0;0) représenterait un classifieur qui déclare tout événement comme étant négatif. Il ne détectera donc aucun positif, mais ne déclarera jamais de négatif comme étant positif. Quant au point F (1;1), il représente l'inverse, soit un classifieur qui déclare tout événement comme étant positif.

Il est intéressant de s'attarder aussi à la droite bissectrice (y = x). En supposant un échantillon de taille infinie, un classifieur aléatoire, c'est-à-dire un classifieur qui classerait un événement comme étant positif avec probabilité p, peu importe les spécificités de ce dernier, résulterait en un point de cette droite (le point $(p;p)$ pour être exact). Pour un échantillon de taille finie, le classifieur aléatoire donnerait un point sur l'espace ROC qui converge en probabilité vers la droite bissectrice.

Normalement, un classifieur devrait résulter en un point supérieur ou égal à la droite bissectrice. De plus, tout classifieur résultant en un point qui se situe au nord-ouest d'un autre lui est préférable. C'est le cas du point B qui est préférable au point D.

Par contre, dans le cas des points B et C de la figure 2.2, il est impossible de trancher entre les deux puisque le point C a un meilleur ratio de vrais positifs, mais il se trompe également plus souvent en générant plus de faux positifs. Dans un cas comme celui-ci, si on veut définir le meilleur classifieur, il faudra calculer les coûts reliés aux différents types d'erreurs. On touche ici à une faiblesse de l'analyse dans l'espace ROC puisque ces dernières ne peuvent pas régler facilement ce problème. Il peut toutefois être résolu par l'utilisation des courbes de coûts qui seront traitées au chapitre 3.

Par définition, un classifieur qui détecte moins de positifs (plus près du point (0;0)), sera dit conservateur. À l'inverse, un classifieur plus près du point (1;1) sera dit libéral. Dans la majorité des domaines, les instances négatives sont beaucoup plus fréquentes que les instances positives et, pour cette raison, on priorisera une valeur conservatrice plutôt qu'une valeur libérale [Fawcett, 2004]. Il est intéressant de noter qu'un classifieur qui performe mieux en zone conservatrice qu'en zone libéral signifie qu'il est meilleur pour identifier des événements qui sont clairement positifs que des événements clairement négatifs.

2.4 La courbe ROC

Dans la section précédente, il a été question de résultats provenant de classifieurs binaires discrets (qui produisent un résultat du type vrai/faux pour chaque événement). Ces derniers nous donnaient des points de l'espace ROC. Généralement, un tel résultat

est peu utile puisqu'un test peut rarement déterminer avec certitude la classe d'un événement. Pour cette raison, on utilisera plutôt un classifieur probabiliste.

Dans ce cas, le classifieur probabiliste génèrera un score à partir de l'événement. Sans perte de généralité, on peut supposer qu'un score plus élevé signifie une plus grande probabilité que l'événement soit réellement positif. Le score peut prendre une valeur réelle quelconque, voire négative, mais, dans la plupart des cas, il s'agit d'une valeur positive et bornée. Certains classifieurs fournissent en sortie une estimation de la probabilité conditionnelle. À ce moment, le score se trouve évidemment compris entre 0 et 1. Le fait que le score représente une probabilité conditionnelle est une propriété qui peut être souhaitable pour certaines applications, par exemple si on souhaite calculer un coût espéré en multipliant la probabilité conditionnelle par le coût de l'événement. Pour cette raison, plusieurs techniques dites de « calage (*calibration*) » [Zadrozny et Elkan, 2001] ont été suggérées pour obtenir la fonction monotone qui permet d'obtenir la probabilité conditionnelle en fonction du score.

Afin d'illustrer les différents aspects traités dans ce mémoire, supposons que nous avons un jeu de données composé de 20 instances positives et de 40 instances négatives. Supposons que notre classifieur assigne les scores de l'ensemble {0,20; 0,25; 0,28; 0,30; 0,34; 0,39; 0,47; 0,56; 0,58; 0,62; 0,63; 0,64; 0,65; 0,70; 0,75; 0,77; 0,80; 0,88; 0,90; 0,95} aux instances positives et les scores de l'ensemble {0,01; 0,02; 0,03; 0,04; 0,05; 0,06; 0,07; 0,08; 0,09; 0,10; 0,11; 0,12; 0,13; 0,14; 0,15; 0,16; 0,17; 0,18; 0,19; 0,21; 0,22; 0,23; 0,24; 0,29; 0,31; 0,32; 0,33; 0,35; 0,36; 0,37; 0,38; 0,40; 0,41; 0,49; 0,51; 0,53; 0,55; 0,69; 0,78; 0,80} aux instances négatives.

Dans un tel scénario, afin d'obtenir un point de la courbe ROC, il faudra choisir un score minimal à partir duquel le classifieur classifiera une instance comme étant

positive. Ce score sera défini comme étant le seuil. Tout événement provoquant un score supérieur à ce seuil sera donc déclaré positif par le classifieur et vice versa. Ainsi, pour chaque seuil choisi (variant entre l'infini négatif et l'infini positif ou entre le minimum et la maximum des scores), on obtiendra un point différent de l'espace ROC. Si on organise les scores en ordre croissant et qu'on choisit deux seuils différents, les deux seuils pourront être situés entre les deux mêmes scores, ou séparés d'un ou de plusieurs scores. Dans le premier cas, il n'y a aucune instance entre les deux seuils et les ratios de vrais et de faux positifs seront les mêmes. Cela nous donnera donc deux points identiques dans l'espace ROC. Si par contre les seuils sont séparés de scores, alors tous les événements déclarés positifs par le seuil le plus élevé le seront aussi par le seuil le plus faible, mais ce dernier déclarera aussi d'autres instances comme étant positives. Ici, le seuil le plus faible aura donc nécessairement des ratios de vrais et de faux positifs supérieurs ou égaux à ceux du seuil le plus élevé. Donc le point ROC du seuil le plus faible sera plus libéral que celui du grand seuil. Il sera cependant généralement impossible de trancher (à partir de la courbe ROC exclusivement) lequel des deux seuils est le plus performant. Si le seuil est supérieur au score maximal, toutes les instances seront déclarées négatives et on obtiendra le point (0;0). Inversement, si le seuil est inférieur au score minimal, on aura le point ROC (1;1). En reliant tous les points obtenus lorsqu'on fait varier le seuil, on obtiendra une courbe allant de (0;0) à (1;1). C'est cette courbe qu'on appelle la courbe ROC.

Si on considère notre exemple numérique et la figure 2.2 vu précédemment, un seuil de 1 ne classifiera aucune instance comme étant positive, et provoquera le point O de l'espace ROC. À l'inverse, un seuil de 0 classifierait toutes les instances comme étant positives et nous obtiendrons le point F. Un seuil de 0,48 classifierait

adéquatement 13 des 20 instances positives et se tromperait pour 7 des 40 instances négatives. Ce seuil génèrerait donc un ratio de faux positifs de 0,175 et de vrais positifs de 0,65, c'est-à-dire le point B de la figure 2.2. De la même manière, un seuil de 0,66 génèrerait le point H et un seuil de 0,26 le point C. Nous pourrions donc tracer une courbe ROC passant par les points O, H, B,C puis F.

De manière générale, à chaque seuil est associée une différente matrice de confusion. La courbe ROC est donc la représentation graphique de ces matrices de confusion. Elle illustre le ratio de vrais positifs en fonction du ratio de faux positifs. Puisque les ratios de vrais et de faux négatifs valent 1 – *ratio de faux positifs* et 1 – *ratio de vrais positifs* respectivement, ces informations sont redondantes lorsqu'on a la courbe ROC et on choisit simplement de les ignorer.

La courbe ROC est intéressante car elle estime si un classifieur produit un bon score relatif, c'est-à-dire que le test n'a pas à donner un score en particulier ou à spécifier la vraie probabilité, il suffit que les différences entre les scores soient assez marquées pour que le classifieur tranche correctement entre une instance positive et une négative.

La courbe ROC est indépendante du rapport entre le nombre d'instances positives et le nombre d'instances négatives. En effet, si on étudie la matrice de confusion, on peut remarquer que les ratios de vrais positifs et de faux positifs sont tous deux des mesures influencées strictement par leur colonne respective. Ce qui veut dire que ces deux valeurs sont indépendantes du rapport entre les positifs et les négatifs. D'autres statistiques, dont l'exactitude et la précision, reposent sur les deux colonnes et ne jouissent donc pas de cette propriété d'indépendance (cette propriété à toutefois été remise en question récemment [Fawcett et Flach, 2005; Webb et Ting, 2005]). Fawcett

et Provost (1997) font remarquer dans leur article que cette propriété est un atout important puisque le niveau de symétrie varie beaucoup dans la majorité des domaines.

La courbe ROC sera également indépendante des coûts engendrés par une erreur dans la classification des instances, dits les coûts d'erreurs de classification. Cette dernière propriété peut être à la fois un avantage et un inconvénient selon ce qui nous intéresse. Dans le chapitre trois, nous verrons un type de courbe qui pourra tenir compte des coûts liés aux erreurs.

Il y a bien sûr plusieurs manières d'obtenir une courbe ROC. L'algorithme 2.1 de l'annexe B [Fawcett, 2006a] permet d'en obtenir une. L'idée générale est que, lorsque les données sont ordonnées de manière décroissante, si un seuil déclare une instance comme étant positive, tout seuil inférieur la déclarera aussi positive. Ainsi, avec ce code, on pourra générer la courbe ROC dans un temps de l'ordre de $O(n \ln(n))$.

2.5 Courbe ROC admissible (proper ROC Curve)

Mueller et Zhang (2006) donnent une définition théorique d'une courbe ROC admissible. Selon eux, l'usage de la courbe ROC est trop large et ils la restreignent à ce qu'ils nomment la courbe ROC admissible (*admissible or proper ROC curve*). Voici la définition qu'ils en donnent :

« Une courbe ROC admissible est une courbe croissante, concave et continue presque partout définie sur l'espace carré cartésien [0,1] X [0,1] débutant au point (0;0) et terminant au point (1;1). »

On remarquera que la concavité de la courbe est la différence entre la courbe ROC admissible et la courbe ROC définie jusqu'à présent. L'idée provient de la notion que, théoriquement, la courbe ROC ne devrait idéalement jamais passer sous la barre y = x.

2.6 Aire sous la courbe

La section 2.2 sur les matrices de confusion a relevé que, dans une situation de classifieur binaire, il y avait toujours deux types d'erreurs possibles (un faux positif et un faux négatif). Il n'est donc pas difficile de se convaincre qu'en résumant une statistique sous la forme d'un scalaire, on perdra de l'information en favorisant un type d'erreur par rapport à l'autre. La courbe ROC, étant bidimensionnel, compense ce problème. Par contre, la théorie ROC peut poser des problèmes lorsqu'on veut comparer deux classifieurs.

Nous avons vu à la section 2.3 qu'un point ROC est supérieur à un autre s'il est situé à sont nord-ouest de celui-ci. Il est cependant impossible de trancher si les deux points sont situés autrement (c'est-à-dire que l'un des points est au nord-est de l'autre). De la même façon, une courbe ROC située au nord-ouest d'une autre indiquera un classifieur plus performant. Par contre, lorsqu'une courbe ROC en croise une autre, on se retrouve de nouveau dans l'incapacité de trancher entre les deux classifieurs. L'avantage de l'approche ROC est qu'on conserve la totalité de l'information, c'est-à-dire que dans un scénario avec un croisement de courbes ROC, on peut déduire que, pour un ratio de faux positifs fixé, la courbe la plus élevée est associée au meilleur classifieur. On accepte donc que l'un des classifieurs soit plus performant sous certaines conditions, mais moins performant autrement. Cette distinction ne peut être établie

lorsque la mesure de performance utilisée est une statistique scalaire. Par exemple, si on utilise une statistique scalaire afin de comparer deux classifieurs, la comparaison donnera un résultat de la forme, plus petit ou égal ou plus grand ou égal. Nous n'aurons donc jamais de scénario où un des classifieurs est parfois meilleur parfois moins bon que l'autre.

Bien qu'il soit intéressant de garder toute l'information de la matrice de confusion, il est possible qu'on veuille tout de même trancher entre deux classifieurs. Dans ce cas, l'approche graphique ROC peut devenir inefficace. Une façon d'utiliser l'information ROC afin de trancher malgré tout est de la ramener sous forme scalaire. Une telle solution est de prendre l'aire sous la courbe ROC [Bradley, 1997; Hanley et McNeil, 1982]. Comme un classifieur est plus performant lorsque sa courbe ROC est située au nord-ouest, une aire sous la courbe ROC, dorénavant notée AUC (*Area Under the ROC Curve*), indiquera une meilleure performance si elle est plus grande. Il faut noter qu'étant donné la définition de l'espace ROC, l'AUC prendra une valeur scalaire entre 0,5 et 1. En fait, l'aire pourrait être inférieure à 0,5, mais cela signifierait que le classifieur est moins performant que de classer aléatoirement. Dans un cas comme celui-là, on peut simplement considérer le classifieur comme erroné ou inutile.

L'AUC correspond à la probabilité qu'un classifieur assigne un score plus élevé à une instance positive choisie aléatoirement que celui assigné à une instance négative choisie de la même manière. Cela revient à faire le test de rang de Wilcoxon [Hanley et McNeil, 1982].

2.7 Limites de la courbe ROC

Selon Webb et Ting (2005), la littérature suggère que l'analyse ROC peut être utile pour évaluer la performance espérée d'un classifieur même lorsque les proportions respectives de chacune des deux classes (instances positives et négatives) varient. Voici quelques citations (en traduction libre) à l'appui :

« L'analyse ROC est la seule mesure disponible qui n'est pas influencée par un biais de décision et par les probabilités a priori » [Swets, 1988]

« Les courbes ROC décrivent le comportement prédictif d'un classifieur indépendamment des distributions des classes et des coûts liés aux erreurs. La performance de la classification est donc indépendante de ces facteurs. » [Provost, Fawcett & Kohavi, 1998]

« L'hypothèse-clé de l'analyse ROC est que les ratios de vrais et de faux positifs décrivent la performance d'un model indépendamment de la distribution des classes. » [Flach et Wu, 2003]

Webb et Ting (2005) ajoutent cependant un bémol; il y a certain cas où cette utilisation de l'analyse ROC est inappropriée.

En effet, bien qu'il soit raisonnable de croire que les ratios de vrais et de faux positifs sont indépendants du coûts liés aux erreurs, il n'en est pas nécessairement de même pour l'hypothèse faite dans la citation de Flach et Wu (2003). Webb et Ting (2005) argumentent que cette dernière n'est vraie que si les ratios de vrais et de faux

positifs sont indépendants de la distribution des classes. Une traduction du contre-exemple qu'ils proposent est disponible en annexe A.

La conclusion des études de Webb et Ting (2005) est qu'il n'est pas toujours réaliste de s'attendre à ce que l'analyse ROC prédise précisément la performance d'un modèle lorsque les distributions des classes varient.

Dans un autre ordre d'idées, il a été question précédemment de la robustesse d'une courbe ROC par rapport aux variations dans le coût lié aux erreurs. Cette propriété, bien que considérée comme un avantage dans certaines situations, peut parfois être considérée comme problématique. Par exemple, il n'est pas surprenant, dans un milieu comme la finance, de préférer tester la performance d'un classifieur non pas en fonction de la quantité de succès et d'erreurs, mais en fonction des coûts engendrés par ce classifieur. Fawcett (2006b) propose une solution : la courbe ROC à instance variable (ROCIV). Le concept est généralement le même que pour la courbe ROC classique, mais l'on regarde le ratio du bénéfice total des vrais positifs sur le bénéfice total des positifs.

2.8 Variance et intervalle de confiance de la courbe ROC

Il a été déterminé au cours de ce chapitre que les courbes ROC servaient à évaluer la performance de classifieurs binaires. Cependant, si on regarde simplement une courbe ROC provenant d'un jeu de données, il se peut que le résultat ne soit pas représentatif de la réalité, c'est-à-dire que la courbe ROC pourrait être une mauvaise estimation de la performance d'un classifieur. Afin de vraiment pouvoir évaluer le classifieur à l'aide de la courbe ROC, il faut définir une mesure de variance. Pour y parvenir, il faut trouver une projection unidimensionnelle de la courbe ROC et construire

les intervalles de confiance pour les ratios de faux et/ou de vrais positifs pour différents points fixes de cette projection de la courbe ROC. Il existe plusieurs façons de s'y prendre ([Bradley, 1997], [Provost et al., 1998] et [Fawcett, 2004]). Dans tous les cas, on pourra se questionner sur la validité de la projection choisie par rapport à une autre. Il est possible de définir une mesure de variance à l'aide de bandes de confiance englobant la courbe ROC en entier. Ce mémoire ne traitera toutefois pas de ce sujet et se penchera exclusivement sur les intervalles de confiance ponctuels. Les deux méthodes les plus populaires dans le domaine de l'apprentissage machine [Fawcett, 2004] sont présentées ici.

2.8.1 Le moyennage vertical (vertical averaging)

Dans le cadre du moyennage vertical, plusieurs jeux de données sont générés et on calcule la courbe ROC pour chacun d'eux. Ensuite, on fixe le ratio de faux positifs, puis on prend la moyenne de tous les ratios de vrais positifs correspondants. En faisant cela pour chaque ratio de faux positifs, on obtiendra une nouvelle courbe ROC moyenne.

De la même manière que pour le calcul de la moyenne, on pourra calculer la variance échantillonnale, puis obtenir un intervalle de confiance. Généralement [Fawcett, 2004], puisqu'un événement est soit positif, soit négatif, on estimera la distribution du ratio de vrais positifs par une loi binomiale lorsqu'on voudra construire l'intervalle de confiance.

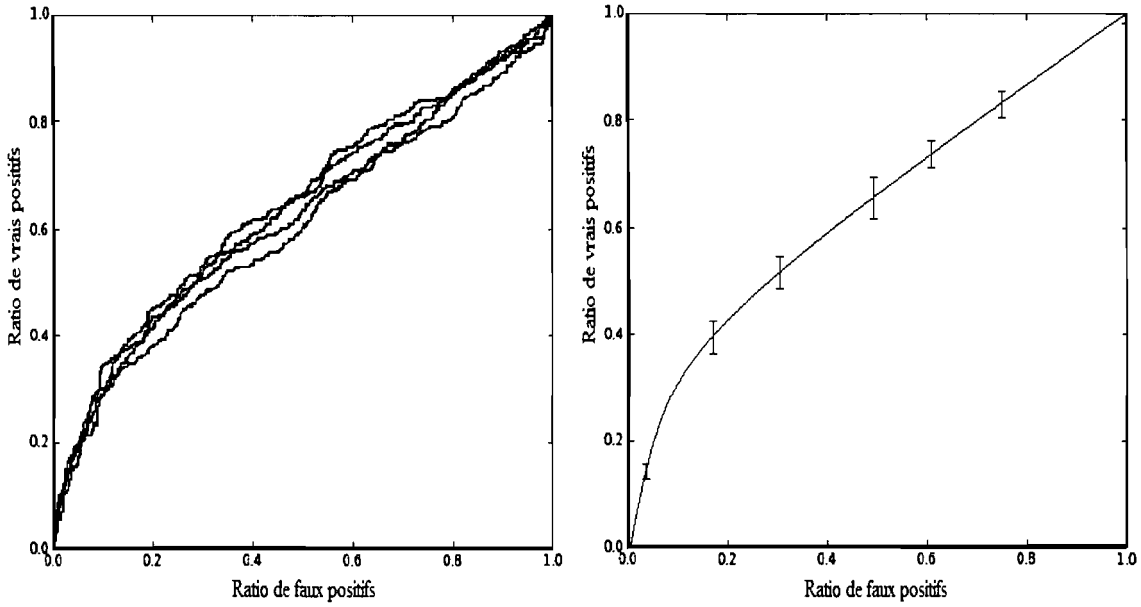

Fig. 2.3 À gauche: 4 courbes ROC provenant d'une même loi.
À droite: Courbe moyenne provenant du moyennage vertical ainsi que l'intervalle de confiance pour quelques points.

Le moyennage vertical est facile à calculer et à visualiser. Malheureusement, Fawcett (2004) fait remarquer que, afin de l'appliquer dans une situation réelle, il faut être en mesure de contrôler le ratio de faux positifs. Comme l'on connaît rarement les résultats d'un test à l'avance, cela est rarement possible. Par exemple, si l'on s'intéresse au nombre de personnes qui répondront à un sondage que l'on envoi par la poste, il est difficile de contrôler le nombre de personnes qui refuseront de répondre à notre sondage.

2.8.2 Le moyennage par seuil (threshold averaging)

Une approche alternative au moyennage vertical est le moyennage par seuil. Le problème principal de la méthode précédente était la nécessité de pouvoir contrôler le ratio de faux positifs. Cette nouvelle méthode règle ce problème. On génère encore plusieurs courbes ROC à partir de la même distribution.

Une fois ces courbes générées, il nous faudra fixer un seuil. Il est important de bien définir ce qu'on veut dire par seuil. Le seuil est le score minimal que le classifieur

déclarera comme provenant d'une instance positive. Tout score supérieur ou égal à cette valeur sera classé comme provenant d'une instance positive et tout score inférieur comme provenant d'une instance négative. Lorsqu'on crée une courbe ROC à partir d'une distribution d'instances positives et négatives, on génère aléatoirement une distribution empirique d'instances positives et une d'instances négatives. La distribution empirique des instances positives donnera, en fonction du seuil, une valeur au nombre de vrais positifs (et donc la coordonnée de l'ordonnée sur le graphique ROC). Pareillement, la distribution empirique des instances négatives donnera, en fonction du seuil, une valeur au nombre de faux positifs (et donc la coordonnée de l'abscisse sur le graphique ROC). En répétant cette expérience pour une nouvelle courbe ROC avec le même seuil, on obtiendra donc deux nouvelles distributions empiriques et le point de l'espace ROC sera donc différent à la fois en abscisse et en ordonnée.

En conséquence, une fois qu'on a généré plusieurs courbes ROC, lorsqu'on les compare par rapport au seuil que l'on a fixé, on obtiendra un point totalement différent sur chacune des courbes. On prendra alors la moyenne (et la variance) séparément par rapport à l'ordonnée et à l'abscisse. On obtiendra donc une mesure de la variance et de la moyenne pour chaque point et on aura notre nouvelle courbe ROC moyenne. Il est intéressant de noter que, lors du calcul de l'intervalle de confiance, nous obtiendrons cette fois des intervalles de confiance bidimensionnels. Fawcett (2004) propose de prendre des intervalles de formes rectangulaires. On peut également calculer des intervalles elliptiques sous l'hypothèse d'indépendance des distributions (voir figure 2.4).

Contrairement au ratio de faux positifs, le seuil peut être fixer arbitrairement par le chercheur, ce qui est un avantage du moyennage par seuil par rapport au moyennage vertical. Fawcett (2004) soulève toutefois un risque. Lorsqu'on étudie les intervalles de

confiance, comme le résultat dépend des scores assignés par le classifieur, il faut s'assurer d'avoir la même échelle et d'être dans la même classe de modèle.

Les deux types de moyennage présentés ont leurs avantages et leurs inconvénients. On pourrait se demander laquelle des deux projections est la plus appropriée. La réponse dépend en fait de ce qui nous intéresse. Le moyennage vertical donne des intervalles unidimensionnels permettant d'ordonner les modèles du meilleur au pire de façon simple, alors que le moyennage par seuil, plus complexe, est lié directement à des applications pratiques.

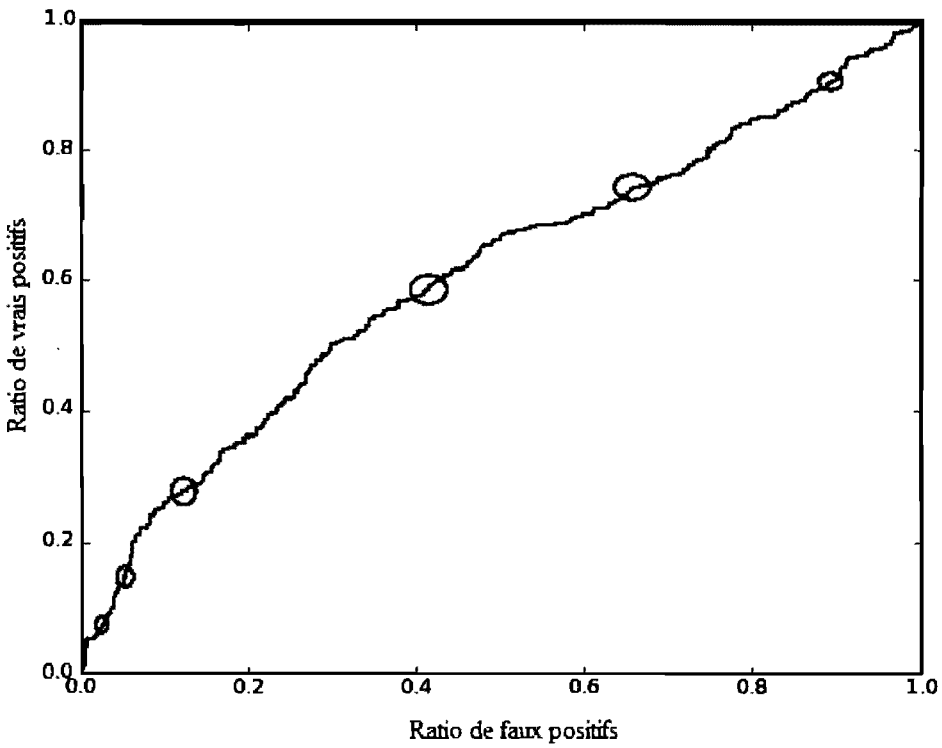

Fig. 2.4 Courbe ROC moyenne provenant du moyennage par seuil avec des intervalles de confiance elliptiques

CHAPITRE 3

Courbes de coûts

Le chapitre précédent présentait un outil efficace de visualisation de la performance d'un classifieur binaire, les courbes ROC. Toutefois, cette analyse par le biais des courbes ROC n'est pas la panacée et nous avons discuté des limites de cette analyse à la section 2.7. À ce sujet, Drummond et Holte (2006) présentent une série de questions auxquelles les courbes ROC n'arrivent pas à répondre à l'aide d'une simple évaluation graphique :

- Quelle est la performance d'un classifieur (son coût espéré) sachant la distribution des classes et le coût des erreurs de classifications?
- Pour quels coûts d'erreurs et distributions des classes est-ce qu'un classifieur performe mieux que le classifieur trivial qui assigne toujours la même classe (positive ou négative) peu importe le score?
- Pour quels coûts d'erreurs et distributions des classes est-ce qu'un classifieur performe mieux qu'un autre?
- Quelle est la différence de performance entre deux classifieurs en tenant compte des coûts?
- Quel est l'intervalle de confiance au niveau alpha pour la performance (l'espérance du coût) pour un classifieur?

Une approche naïve pour déterminer la meilleure courbe serait de choisir le point qui minimise le total des erreurs (faux positifs + faux négatifs). Il faut noter que cette option présuppose des coûts équivalents pour les 2 types d'erreur.

Supposons, par exemple, qu'on cherche à déterminer si une transaction bancaire est frauduleuse ou légale. Dans ce cas, un faux positif représente le fait d'accuser à tort un client de fraude. Cette erreur provoquera du désagrément au client et engendrera un certain coût pour la banque. Parallèlement, un faux négatif représente une fraude que la banque ne détecte pas et engendrera un coût proportionnel à la taille de la transaction. Dans un tel scénario, le coût lié au fait d'avoir un ratio de faux positifs de ¼ ne sera pas le même que celui lié au fait d'avoir un ratio de faux négatifs de ¼. L'objectif de tout organisme étant de minimiser ses coûts, il faudra choisir le ratio de faux positifs qui provoque le moindre coût. Supposons que deux courbes ROC (provenant de deux classifieurs) pour cette situation se croisent, le choix se fera en fonction du ratio optimal. Cela signifie que la courbe ROC en elle-même ne pourra pas nous indiquer directement le classifieur adéquat pour la situation. Nous pourrions appliquer l'approche naïve proposée plus haut. Cependant, comme l'objectif général est de minimiser le coût total espéré et qu'il est peu probable que le coût d'accuser un client à tort soit le même que celui de laisser passer une transaction frauduleuse, cette approche ne sera donc pas valide non plus.

En introduisant les courbes de coûts, on aura un outil qui permettra de minimiser le coût total espéré sans l'hypothèse des coûts égaux qu'impose l'approche naïve. Cet outil permettra, en plus, de répondre aux questions présentées plus haut d'un simple coup d'œil graphique.

3.1 Conditions d'opération (*operating conditions*)

Par conditions d'opération, on entend l'ensemble de valeurs comprenant les probabilités a priori de chacune des classes et les coûts associés à chaque type d'erreur.

Une fois ces conditions fixées, la courbe de coûts devient une simple transformation mathématique de la courbe ROC.

Définissons $p(+)$ et $p(-)$ comme étant les probabilités qu'une instance soit membre de la classe des instances positives et négatives respectivement. Posons aussi $C(+|-)$ et $C(-|+)$, les coûts associés à classifier une instance négative (positive) comme étant positive (négative). Il est raisonnable de supposer que $C(+|-)$ et $C(-|+)$ sont plus grands que $C(+|+)$ et $C(-|-)$ respectivement. Cela signifie qu'il est plus coûteux de se tromper que de bien classifier une instance. Comme on veut trouver le classifieur qui minimisera le coût, il est possible, sans perte de généralité, de soustraire à chaque instance le coût associé à une bonne classification. Ainsi, $C(+|+)$ et $C(-|-)$ vaudront toujours 0. Lorsqu'une condition d'opération est fixée, c'est-à-dire que les probabilités $p(+)$ et $p(-)$ ainsi que les coûts $C(-|+)$ et $C(+|-)$ sont déterminés, on peut utiliser l'équation (3.1) afin de déterminer un scalaire w. Cette valeur est définie comme étant un point d'opération :

$$w = \frac{p(+)C(-|+)}{p(+)C(-|+) + p(-)(C+|-)}. \tag{3.1}$$

La valeur w représente donc un résumé de l'information a des fins d'optimisation des calculs. Il faut noter que, pour différentes combinaisons de $p(+)$, $p(-)$, $C(-|+)$ et $C(+|-)$ générant un même point d'opération w, on obtiendra le même résultat du point de vue de l'optimisation.

3.1.1 Droite d'iso-performance d'un classifieur

Supposons un jeu de données et un classifieur résultant en un point de l'espace ROC, disons (FP_1,TP_1), son coût espéré sera :

$$E[Coût] = FP_i\, p(-)C(+|-) + (1-TP_i)\, p(+)C(-|+). \qquad (3.2)$$

En isolant TP_i on obtient :

$$TP_i = FP_i \frac{p(-)C(+|-)}{p(+)C(-|+)} + \frac{p(+)C(-|+) - E[Coût]}{p(+)C(-|+)}. \qquad (3.3)$$

Puisque les coûts C(-|+) et C(+|-) ainsi que les distributions des classes p(-) et p(+) sont fixés, les classifieurs ayant la même espérance de coûts devront produire des points (FP_i,TP_i) de l'espace ROC qui se trouvent sur la droite décrite par l'équation (3.3) [Hilden & Glasziou, 1996; Fawcett et Provost, 1997]. Cette dernière porte le nom de droite d'iso-performance [Drummond et Holte, 2006]. Il faut noter que la pente de la droite d'iso-performance est nécessairement positive.

Si on relie (FP_1,TP_1) et (FP_2,TP_2), les points de l'espace ROC résultant de deux classifieurs, la droite obtenue aura pour pente :

$$\frac{TP_2 - TP_1}{FP_2 - FP_1}. \qquad (3.4)$$

Pour un point d'opération donné, si la pente (3.4) n'égale pas celle de la droite (3.3), l'un des classifieurs sera plus performant que l'autre. Supposons qu'on trace la droite d'iso-performance d'un classifieur associée à des conditions d'opération données. Tout classifieur générant un point de l'espace ROC situé au nord-ouest de cette droite sera considéré plus performant. Inversement, un classifieur moins performant produira un point au sud-est. La figure 3.1 représente cette situation. Pour le point d'opération générant la pente de la droite pleine du graphique, le classifieur B sera plus performant que le H, mais moins que le C et équivalent à celui qui génère G.

3.1.2 Application à la courbe ROC

Appliquer cette méthode de comparaison sur un grand nombre de points permet de déterminer, pour un point d'opération donné, le ratio de faux positifs le plus performant sur une courbe ROC. La figure 3.2a représente une courbe ROC empirique et une série de droite d'iso-performance ayant une pente de 2 (donc provenant des conditions d'opération).

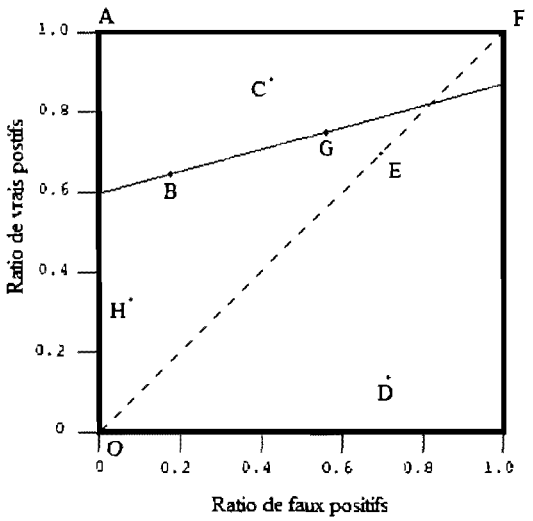

Fig. 3.1 Espace ROC et droite d'iso-performance pour le classifieur B.

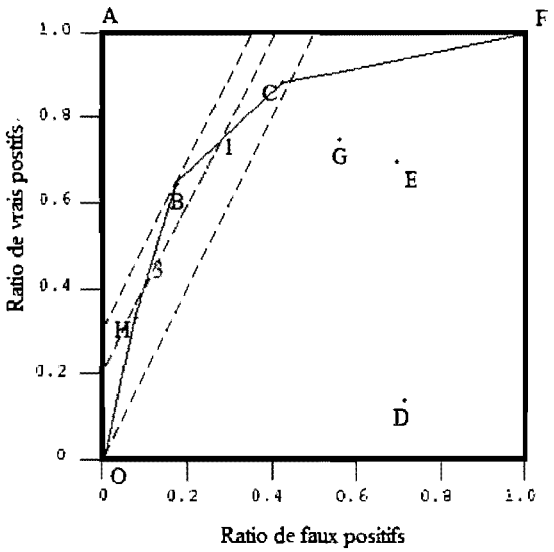

Fig. 3.2a Courbe ROC empirique (ligne pleine) et translation d'une droite d'iso-performance de pente 2. La pente des segments BH et BC sont 3 et 1 respectivement.

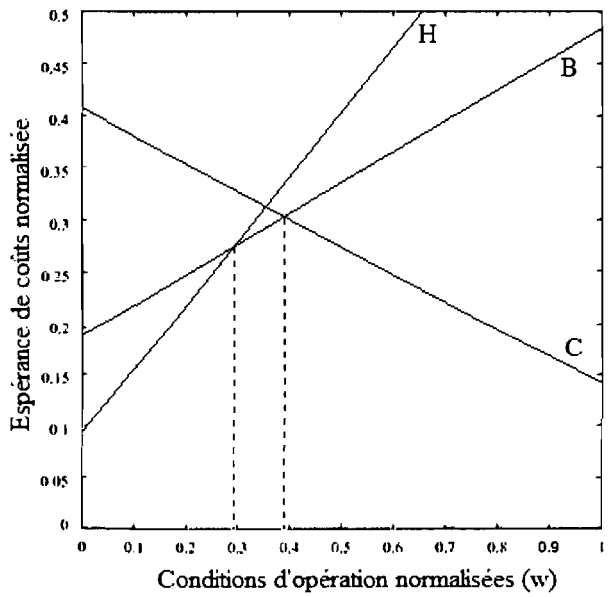

Fig. 3.2b Intervalle des conditions d'opération où le classifieur B est le plus performant (moins coûteux).

Il est facile de se convaincre que le point de la courbe ROC empirique le plus performant est celui dont la droite d'iso-performance est tangente à la courbe. Ce point sera celui où la pente limite à gauche de la courbe ROC est supérieure ou égale à celle de la droite d'iso-performance et où la pente limite à droite de la courbe ROC est inférieure ou égale à celle de la droite. En cas d'égalité des pentes, la conclusion est qu'il existe plusieurs ratios de faux positifs ayant la même performance pour ces conditions d'opération. Si une telle situation n'existe pas, on en conclut que, pour ce point d'opération, l'un des classifieurs triviaux (toutes les instances sont classées positives, ou toutes sont classées négatives) sera plus performant que le classifieur générant la courbe ROC. Si la courbe ROC n'était pas empirique, les points les plus performant pourrait être trouvées de la même façon ou, parfois, en prenant simplement la dérivée de la courbe.

La figure 3.2a permet aussi de remarquer une autre propriété de la courbe ROC empirique. Le point B sera le point le plus performant de la courbe ROC pour toute

combinaison de conditions d'opération dont la droite d'iso-performance est tangente à la courbe ROC au point B. Ici, les pentes des segments précédant et suivant le point B (les pentes 3 et 1 de la figure 3.2a) définissent un intervalle de valeurs qui permet de déterminer l'ensemble de ces combinaisons : toute droite d'iso-performance dont la pente se situe dans cet intervalle nous conduit à choisir le point B de la courbe ROC afin de minimiser le coût espéré.

Bien que nous ne verrons comment obtenir un tel graphique qu'à la section 3.2, la figure 3.2b nous permet d'illustrer ce phénomène plus clairement. Les droites B, C et H de cette figure représentent l'espérance des coûts des classifieurs correspondants. Le classifieur B sera le moins coûteux (donc le plus performant) sur l'intervalle des conditions d'opération entre les droites verticales pointillées. Ce sont ces conditions d'opération qui génèreront les droites d'iso-performances dont la pente est incluse dans l'intervalle de 1 à 3. Dans toutes autres situations, le classifieur C ou H sera plus performant.

Il faut noter que, si une courbe ROC est munie d'une section concave, aucune droite d'iso-performance ne pourra être tangente à cette section. Donc, lorsqu'on regarde la performance de la courbe ROC, le résultat se trouvera nécessairement sur l'enveloppe convexe de la courbe. Ceci implique que le problème soulevé par Mueller et Zhang (2006) sur l'admissibilité d'une courbe ROC (section 2.5 du présent document), ne se posera pas dans une situation où le coût est considéré.

3.2 Une alternative aux courbes ROC

La courbe de coûts est un outil spécifiquement pensé pour être une mesure de performance d'un classifieur qui permet de répondre aux questions posées au début de

ce chapitre d'un simple coup d'œil. Elle est intrinsèquement liée à la courbe ROC et représente en fait la courbe qui minimise le coût espéré selon différentes conditions d'opération.

3.2.1 Coûts d'erreurs équivalents

Cette section considère des scénarios pour lesquels commettre un faux positif est équivalent à commettre un faux négatif. Autrement dit, on suppose C(-|+) = C(+|-). Étant donné qu'on s'intéresse à savoir quel classifieur est le plus performant et non quel est le coût réel de l'application, on peut poser C(-|+) = 1 sans perte de généralité. Nous aurons donc un coût espéré égal au ratio d'erreur. Les conditions d'opération ne seront plus définies que par p(+), soit la probabilité qu'une instance soit de la classe positive. Drummond et Holte (2006) font une mise en garde sur la définition de p(+). Ils relèvent que le pourcentage d'instances positives lors de l'entraînement du classifieur peut être différent de celui obtenu lorsqu'on teste le classifieur en créant sa matrice de confusion et tous deux peuvent être différents de celui qui s'appliquera lorsque le classifieur sera mis en usage réel. Minimiser les coûts nécessite que les conditions d'opération soient définies de sorte que p(+) représente le ratio d'instances positives dans l'application réelle. Malheureusement, cette valeur est inconnue lorsqu'on veut tester un classifieur. Toutes les valeurs possibles de p(+) doivent donc être couvertes afin de pouvoir faire une comparaison.

Une ligne de coûts est donc définie comme étant une droite représentant le coût espéré d'un classifieur sur l'ensemble des conditions d'opération. Donc, puisque p(+) est une probabilité, la courbe prendra des valeurs en abscisse entre 0 et 1. De la même manière, l'ordonnée sera le ratio d'erreur et prendra des valeurs entre 0 et 1. Il faut

remarquer que, si on se trouve aux extrémités de la courbe, alors p(+) = 0 ou p(+) = 1, cela implique que toutes les instances sont négatives ou positives respectivement. Par conséquent, la ligne de coûts prendra la valeur du ratio d'erreur sur les instances négatives (ratio de faux positifs) lorsque p(+) = 0 et sur les instances positives (ratio de faux négatifs) lorsque p(+) = 1. Ces deux points seront reliés par une droite. Il faut remarquer qu'un simple coup d'oeil à une ligne de coûts nous indique donc un point associé à cette droite dans la courbe ROC.

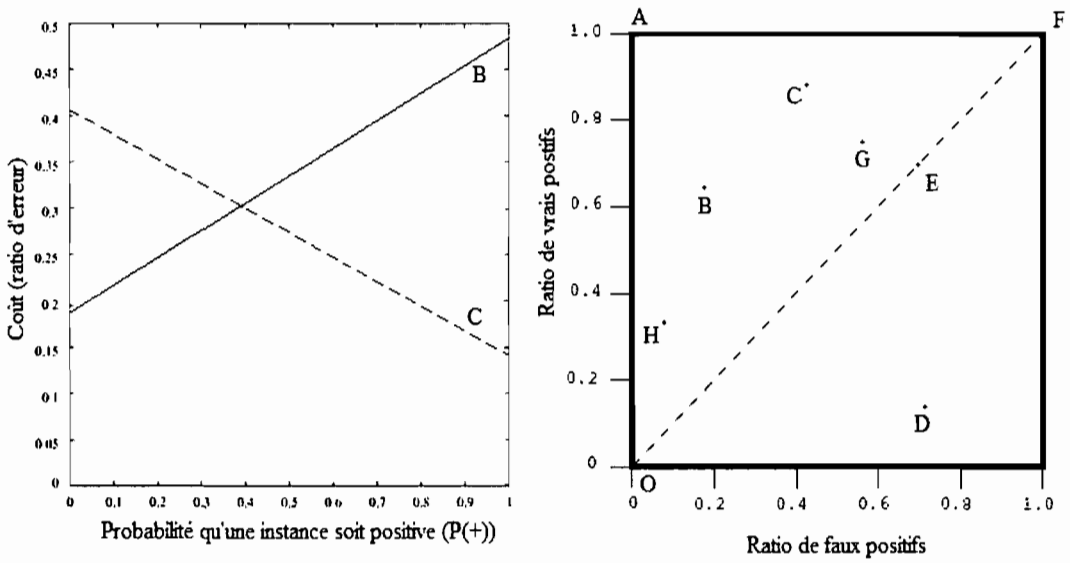

Fig. 3.3 À gauche: deux lignes de coûts dans une situation à coûts d'erreurs équivalents. À droite: le graphique ROC d'où ils ont été obtenus.

La figure 3.3 montre deux lignes de coûts et leurs points ROC associés. Sur les lignes de coûts, on peut remarquer que le classifieur C commet plus d'erreurs sur les instances négatives que sur les instances positives alors que le classifieur B fait le contraire. Sur le graphique ROC, la même information est visible, mais elle requiert de visualiser la droite Y = 1 – X. On y voit aussi que le classifieur B est plus conservateur que le C, il est cependant impossible de dire dans quelles conditions est-ce qu'il sera plus performant que ce dernier. Sur les lignes de coûts, par contre, le croisement des

lignes représente le point ayant le même coût et on peut voir que cette valeur se produit exactement à p(+) = 0,39. On voit donc ici tout l'intérêt de représenter les résultats sous la forme de lignes de coûts. Comme les lignes de coûts et les points de l'espace ROC sont une bijection, toute la théorie s'appliquant à l'une des méthodes a son équivalent pour l'autre. Voici l'équation de la ligne de coûts associée à un point (FP,TP) :

$$Ratio\,d'erreur = (1 - TP - FP)\,p(+) + FP. \tag{3.5}$$

De la même façon, à partir d'une telle ligne, on obtient le point ROC facilement. Le ratio de faux positifs est l'ordonnée à l'origine de la droite et le ratio de vrais positifs est fonction de la pente de la droite et du ratio de faux négatifs :

$$TP = 1 - pente - FP. \tag{3.6}$$

La dualité entre les points ROC et les lignes de coûts implique, entre autres, qu'un classifieur plus performant qu'un autre dans l'espace ROC (situé au nord-ouest), engendrera nécessairement une ligne de coûts inférieure pour toutes conditions d'opération [Preparata et Shamos, 1988]. Drummond et Holte (2006) ajoutent qu'à l'inverse, une droite de l'espace ROC peut être exprimée comme un point sur le graphique des coûts. En effet, chacun des points de l'espace ROC situés sur une même droite donneront des lignes de coûts qui se croiseront en un point. Supposons la droite y = Sx + TP dans l'espace ROC. Le croisement des lignes de coûts se fera donc au point :

$$X = p(+) = \frac{1}{1+S}, \tag{3.7}$$

$$Y = Ratio\,d'erreur = (1 - TP)\,p(+). \tag{3.8}$$

Puisque, dans l'espace ROC, il est généralement question de courbes allant de (0;0) à (1;1) supérieures à la droite y = x, cette bijection est moins utile, mais elle illustre bien la relation entre ces deux espaces. En isolant p(+), puis le ratio d'erreur dans les

équations précédentes, il devient aussi possible de trouver la droite ROC correspondant à un point dans l'espace de coûts.

3.2.2 Génération d'une courbe de coûts

Il a été question jusqu'ici de classifieurs discrets générant un point dans l'espace ROC. La courbe ROC générée par un classifieur probabiliste (avec score) a aussi son équivalent dans l'espace de coûts. La figure 3.4 illustre une telle relation. Puisque la courbe ROC empirique est en fait une série de points de l'espace ROC reliés par des droites, on peut tracer une ligne de coûts pour chacun de ces points. Pour un point d'opération donné, la droite de coûts la plus basse sera associée au point de la courbe ROC (et donc au score) le plus performant. La courbe de coûts est définie par l'ensemble de ces points, c'est-à-dire que pour chaque condition d'opération, le point de la courbe de coûts sera celui situé sur l'enveloppe inférieure de toutes les lignes de coûts. La dualité entre l'espace ROC et l'espace de coûts implique que chaque courbe ROC donnera une courbe de coûts unique. À l'inverse, une courbe de coûts donnera une courbe ROC admissible unique (telle que définie par Mueller et Zhang (2006)). Il y a donc bijection entre l'enveloppe convexe d'une courbe ROC et l'enveloppe inférieure qui forme la courbe de coûts.

Il faut noter que les points (0;0) et (1;1) de l'espace ROC génèrent le triangle (0;0), (0,5;0,5) et (1;0) dans l'espace de coûts. Le point (0,5;0,5) de l'espace de coûts étant le point où les lignes de coûts provenant des points ROC (0;0) et (1;1) se croisent. Puisqu'une courbe ROC passe toujours par ces points ((0;0) et (1;1)), une courbe de coûts sera donc nécessairement incluse dans le triangle (0;0), (0,5;0,5) et (1;0). En d'autres mots, le triangle des valeurs intéressantes ((0;0), (0,5;0,5) et (1;0)) de l'espace

de coûts est la bijection du triangle des valeurs intéressantes ((0;0), (0;1) et (1;1)) dans l'espace ROC.

En faisant tendre le nombre de points de la courbe ROC vers l'infini, on peut appliquer les mêmes résultats à une courbe ROC théorique.

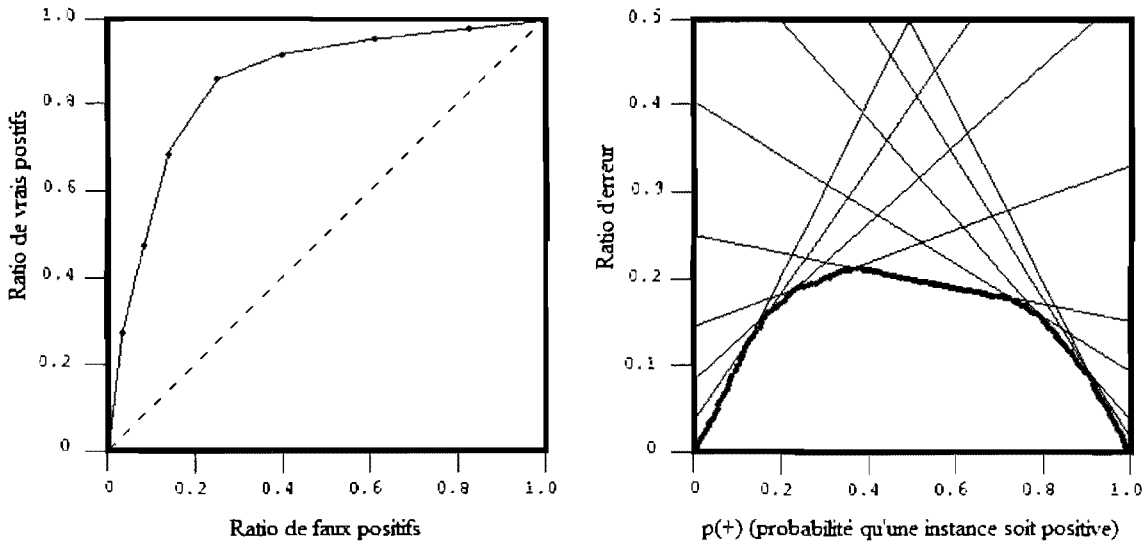

Fig. 3.4 À gauche: courbe ROC empirique générée avec 7 points.
À droite: lignes de coûts associées à chaque point ROC, l'enveloppe inférieure (en gras) est la courbe de coûts associée à la courbe ROC.

3.2.3 Coûts d'erreurs variables

Dans les sections précédentes, il a été supposé que le coût provenant d'un faux positif était le même que celui provenant d'un faux négatif. Dans cette section, cette hypothèse sera relaxée. Encore une fois, puisque l'objectif est de déterminer le meilleur classifieur, il est possible, sans perte de généralité, de translater les coûts de sorte que C(-|-) = C(+|+) = 0. Ainsi, le classifieur parfait aurait un coût espéré nul.

Il a été mentionné à la section 3.2.1 que, pour un point de l'espace ROC (FP_1,TP_1), l'espérance de coût est :

$$E[Coût] = FP_1 p(-)C(+|-) + (1-TP_1)p(+)C(-|+). \qquad (3.9)$$

Sous l'hypothèse C(-|+) = C(+|-), l'espérance du coût avait été normalisée en supposant ces valeurs égales à 1. Sans cette hypothèse, il faudra normaliser en divisant par le coût maximal atteignable. Cela survient lorsque le ratio de faux négatifs et de faux positifs valent tous deux 1. Le coût maximal est donc :

$$\max[coût] = p(-)C(+|-) + p(+)C(-|+). \tag{3.10}$$

Afin de ne pas avoir de division par l'infini, il est nécessaire de supposer que les coûts sont finis. Dans la réalité, cette hypothèse est très réaliste et ne pose aucun problème.

De la même manière, il faudra normaliser les conditions d'opération. Ces dernières deviendront donc :

$$w = \frac{p(+)C(-|+)}{p(+)C(-|+) + p(-)(C+|-)}. \tag{3.11}$$

On reconnaîtra ici l'équation (3.1) définissant l'équation d'un point d'opération. L'espace de coûts représentera donc l'espérance de coûts normalisée en fonction des points d'opération et se trouvera encore une fois dans le carré unitaire. Il faut remarquer les deux points extrêmes $w = 0$ et $w = 1$. Dans le premier cas, soit p(+) = 0 ou alors C(-|+) = 0. Il s'agit donc de cas extrêmes où tous les coûts sont associés aux instances négatives, soit parce que les coûts des instances positives sont nuls, soit parce qu'il n'y a tout simplement pas d'instances positives. Le point $w = 1$ sera l'autre extrême et impliquera, à l'inverse, que p(-) ou que C(+|-) est nul. En substituant w dans l'équation de l'espérance de coût normalisé, on obtient :

$$Norm[E(Coût)] = FP_1(1-w) + (1-TP_1)w. \tag{3.12}$$

Il s'agit d'une équation du premier degré. Le graphique sur l'espace de coûts sera donc une droite reliant les point (0, FP_1) et (1, [1 – TP_1]). La droite aura donc

comme point d'origine le ratio de faux positifs et comme point de terminaison le ratio de faux négatifs. Il s'agit donc de la même droite que pour les lignes de coûts dans le cas C(-|+) = C(+|-). La seule différence se trouve dans les axes qui sont maintenant normalisés.

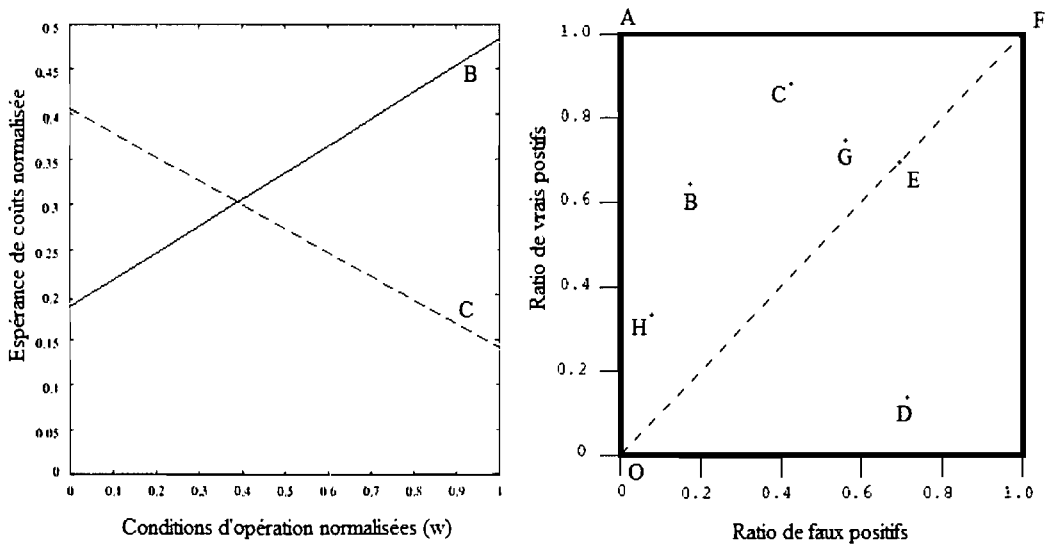

Fig. 3.5 À gauche: deux lignes de coûts.
À droite: le graphique ROC d'où ils ont été obtenus.

La figure 3.5 reprend les mêmes points de l'espace ROC que la figure 3.3 et ces lignes de coûts sous l'hypothèse C(-|+) ≠ C(+|-). On remarque que les deux figures sont identiques, à l'interprétation des axes près. Par conséquent, la formation des courbes de coûts à partir de courbes ROC se fera de la même façon qu'à la section précédente.

Il faut cependant faire attention au fait qu'ainsi définies, les courbes de coûts ne contiennent aucune information sur les valeurs de C(-|+) et C(+|-) et il est donc impossible de conclure sur le coût absolu qui résultera dans l'expérience [Drummond et Holte, 2006]. Il s'agit d'une mesure relative qui sert à comparer des classifieurs sur le plan de leur performance.

3.3 Comparaison de la courbe ROC et de la courbe de coûts

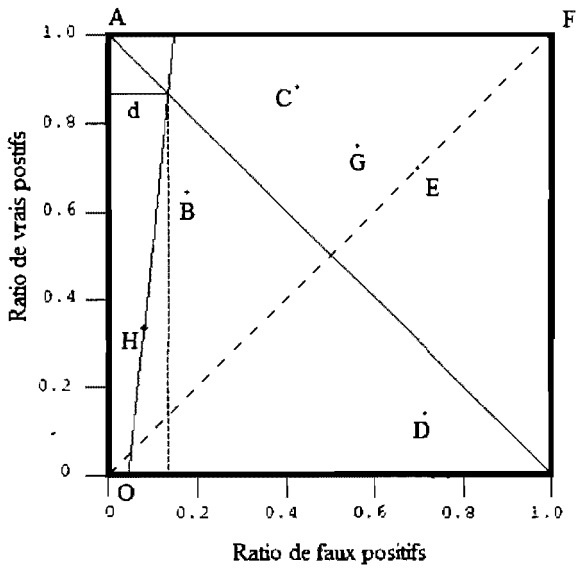

Fig. 3.6 Coût du classifieur H à l'aide du graphique ROC

Puisqu'il est nécessaire de connaître la courbe ROC afin de pouvoir tracer la courbe de coûts, il est pertinent de s'intéresser à la valeur ajoutée de cette dernière. L'avantage principal de la courbe de coûts est qu'elle présente directement un aperçu de l'espérance de coût normalisé. Supposons qu'on cherche à savoir le pourcentage du coût maximal associé à un classifieur simplement à l'aide de la courbe ROC. Drummond et Holte (2006) présentent une méthode permettant d'y arriver (voir la figure 3.6). Il faut tout d'abord tracer la droite d'iso-performance relative au classifieur (H dans notre exemple) et aux conditions d'opération. Ensuite, il faut tracer la diagonale décroissante $y = 1 - x$. Finalement, le segment de droite horizontal (d) reliant l'ordonnée au croisement de la diagonale et de la droite d'iso-performance sera de longueur égale au coût espéré normalisé. Il est donc possible de mesurer cette valeur sur la courbe ROC, mais ce n'est pas aussi rapide et cela deviendra vite fastidieux si on désire connaître la valeur pour de multiples points d'opération ou pour plusieurs points de la courbe ROC.

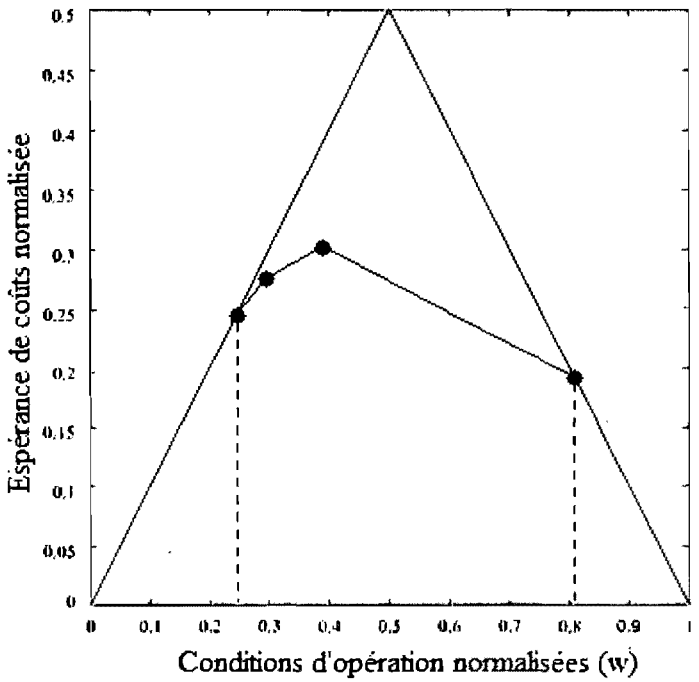

Fig. 3.7 L'intervalle d'opération d'un classifieur par score.

Un autre avantage de la courbe de coûts sur la courbe ROC est sa facilité de comparaison par rapport aux classifieurs triviaux. Pour des conditions d'opération fixées, il faudra, sur la courbe ROC, tracer les droites d'iso-performances partant des classifieurs triviaux. Encore une fois, cela deviendra rapidement fastidieux si on s'intéresse à plusieurs points d'opération. La courbe de coûts illustre ces informations beaucoup plus directement. En effet, tout point se retrouvant à l'intérieur du triangle (0,0), (0.5, 0.5) et (1,0) sera plus performant qu'un classifieur trivial. La figure 3.7 illustre l'intervalle d'opération d'un classifieur à l'aide de la courbe de coûts. Il s'agit de l'ensemble des conditions d'opération sous lesquelles le classifieur est plus performant que les triviaux. Sur une courbe ROC, déterminer la même information nécessiterait de comparer chaque point de la courbe avec toutes les droites d'iso-performances des classifieurs triviaux. Cette caractéristique est importante puisque,

lorsque la courbe ROC est utilisée, il est facile de ne pas réaliser que le classifieur est plus coûteux que d'appliquer bêtement un choix trivial.

Le même phénomène se produit si on veut comparer deux classifieurs entre eux. La figure 3.8 illustre cette situation. Sur la courbe de coûts, on peut déterminer que le classifieur A est au moins aussi performant que le classifieur B pour les conditions d'opération inférieures à 0.5 et inversement pour les conditions supérieures. Au point 0.5, les deux classifieurs sont équivalents. Sur la courbe ROC, la ligne de coûts associée à la condition 0.5 est tangente aux deux courbes ROC. Les points de tangence sont ceux où les deux classifieurs sont aussi performants. Encore une fois, si on s'intéresse au classifieur le plus performant pour l'ensemble des points d'opération, la courbe ROC deviendra rapidement compliquée. Par opposition, sur le graphique des courbes de coûts, on peut trancher par un simple coup d'œil.

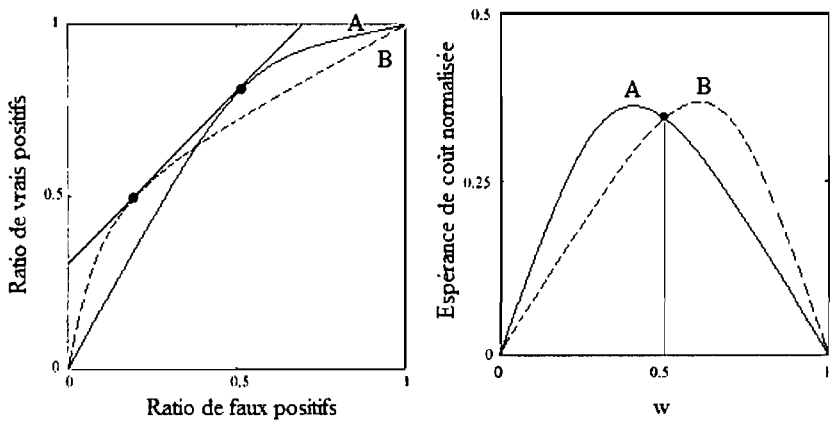

Fig. 3.8 Graphique de comparaison de deux classifieurs (à gauche par la courbe ROC, à droite par la courbe de coûts) Inspiré de Drummond et Holte (2006).

Drummond et Holte (2006) soulèvent que, même dans le cas où un classifieur est supérieur à un autre en tous points, il reste plus compliqué de déterminer, à l'aide d'une courbe ROC, à quel point cette différence est grande. Dans le cas de la courbe de coûts, il suffit de prendre la distance verticale entre les deux courbes. Par contre, si on veut

déterminer cette différence sur une courbe ROC, il faudra tout d'abord trouver les points représentant les coûts minimaux sur chacune des courbes, puis prendre la distance de Manhattan entre eux. Si le nombre de courbes ROC à comparer est élevé, ce processus devient très lourd.

3.4 Moyennage de la courbe de coûts

La section 2.9 traitait du moyennage de la courbe ROC. Il y a été dit qu'il n'existe pas de règles universellement acceptées pour prendre la moyenne entre deux courbes ROC. L'une d'entre elles [Provost et al., 1998] est de prendre la moyenne verticale (on fixe le ratio de faux positifs et on prend la moyenne du ratio de vrais positifs pour chaque cas). Si on trace la courbe de coûts de la courbe résultante au moyennage vertical, la performance obtenue pour une condition d'opération fixée ne sera pas la même que la performance moyenne de deux courbes de coûts initialement calculable. Il en sera de même pour la majorité des méthodes de moyennage. Par contre, si on prend le moyennage par seuil, la courbe de coûts associée au résultat sera exactement celle obtenue si on prend le moyennage vertical des courbes de coûts. Drummond et Holte (2006) expliquent donc qu'il y a une propriété intéressante à choisir cette méthode de moyennage.

Nous croyons toutefois que les deux méthodes de moyennage présentées dans ce mémoire répondent à des questions différentes que pourraient se poser l'usager. Pour cette raison, nous croyons que le choix de la méthode de moyennage devrait plutôt être fait en fonction de l'usage qu'on compte faire du modèle une fois implanté. Le prochain chapitre traitera plus en détail de nos résultats au sujet de l'impact du choix de la

méthode de moyennage sur les intervalles de confiance pour les courbes ROC et de coûts.

3.5 Limitations de la courbe de coûts

Pottmann (2001) démontre que la dualité entre les courbes de coûts et les courbes ROC reste lorsqu'on augmente le nombre de dimensions. Il stipule donc que, pour cette raison, les résultats qui découlent de cette dualité devraient aussi continuer d'exister dans un univers à multiples dimensions.

Par contre, puisque les courbes ROC et les courbes de coûts sont liées par une relation bijective, l'information reste la même dans les deux mondes [Drummond et Holte, 2006]. L'avantage de la courbe de coûts par rapport à la courbe ROC provient donc directement de sa facilité à visualiser graphiquement les résultats. Il faut cependant faire attention : il a été question jusqu'ici de classifieur binaire. Si on augmente le nombre de classes, le nombre de dimensions augmente de façon quadratique et le cadre d'analyse graphique utilisé jusqu'à présent ne s'applique plus. Dans une telle situation, il est donc préférable de projeter les différentes classes afin d'obtenir plusieurs classifications binaires du type : « membre de la classe A versus membre d'une autre classe » [Drummond et Holte, 2006].

CHAPITRE 4

Distribution d'auto-amorçage exacte ponctuelle des courbes ROC

4.1 L'approche d'auto-amorçage

L'auto-amorçage [Efron & Tibshirani, 1993] est une méthode de simulation permettant d'estimer la distribution de statistiques complexes telles les courbes ROC. Le concept est simple. À partir d'un jeu de données initial, un certain nombre d'échantillons sont tirés avec remise. Pour chaque échantillon, la statistique d'intérêt (ici la courbe ROC) est calculée. En prenant la moyenne de ces résultats, on obtient l'estimateur d'auto-amorçage de la statistique. Dans certain cas, il est possible d'obtenir analytiquement la distribution d'auto-amorçage sans avoir à faire de rééchantillonnage. On dira alors qu'il s'agit de la distribution d'auto-amorçage exacte. On peut interpréter la distribution d'auto-amorçage exacte comme étant celle vers laquelle la distribution d'auto-amorçage converge lorsque le nombre d'échantillons d'auto-amorçage tend vers l'infini.

Le calcul de la distribution d'auto-amorçage exacte pose problème dans l'analyse ROC. Cela provient des définitions pour les ratios de vrais et de faux positifs. Rappelons que le ratio de vrais positifs est défini comme le nombre d'instances positives correctement classifiées divisé par le nombre d'instances positives totales de l'échantillon et que le ratio de faux positifs est le nombre d'instances négatives classifiées de manière erronée divisé par le nombre total d'instances négatives dans l'échantillon. Or, si l'une des classes est vide, alors il y a division par zéro. Dans ce cas, la courbe ROC n'est pas définie. Puisque la distribution d'auto-amorçage exacte considère éventuellement toutes les randomisations possibles de rééchantillonnage, les

deux cas mal définis en font partie et la distribution d'auto-amorçage exacte est donc, elle aussi, mal définie. Il est possible de contourner cette problématique à l'aide d'une procédure nommée l'auto-amorçage stratifié. Cette méthode fixe les proportions d'instances positives et négatives des échantillons d'auto-amorçage comme étant égales à celles du jeu de données original. En d'autres termes, les échantillons sont formés à partir de la combinaison de deux échantillons d'auto-amorçage indépendants : l'un tiré des instances positives du jeu original, l'autre des instances négatives. Cette procédure a déjà été utilisée dans le cadre de courbes ROC [Bandos, 2005; Drummond & Holte, 2006]. Il faut noter que, dans le cas d'auto-amorçage complet (classique), si un échantillon composé d'instances positives uniquement est tiré, on doit le rejeter et recommencer la procédure (afin d'éviter une division par zéro). Il en est de même pour un échantillon composé exclusivement d'instances négatives.

L'estimation d'auto-amorçage est obtenue en prenant la moyenne d'une série de courbes ROC. Le chapitre 2 présentait deux méthodes de prise de la moyenne de courbes ROC : le moyennage vertical et le moyennage par seuil. Le choix de la méthode de moyennage découle directement de l'utilisation qu'on compte faire du modèle lors de l'implantation. Pour cette raison, ce mémoire traitera du moyennage vertical et par seuil comme deux problèmes différents plutôt que comme deux méthodes de calcul de la distribution d'une courbe ROC.

L'avantage du moyennage vertical est qu'il engendre un intervalle de confiance unidimensionnel. En contrepartie, le ratio de faux positifs ne peut être mesuré qu'après implantation et utilisation du modèle. En conséquence, cette méthode ne correspondra à aucune application pratique de prédiction. Le moyennage par seuil, à l'inverse, sera un

bon modèle de prédiction, mais engendrera un intervalle de confiance bidimensionnel. Le désavantage est que dans certains cas, on ne peut conclure si un point est meilleur qu'un autre.

Il existe beaucoup de recherches sur le lissage de la courbe ROC empirique et sur l'estimation de sa dispersion. Chacune des approches peut être classée dans l'une des trois catégories suivantes : paramétrique, semi-paramétrique et non paramétrique (empirique). L'approche d'auto-amorçage tombe dans cette dernière catégorie. Bien que le rééchantillonnage d'auto-amorçage soit souvent utilisé afin d'obtenir des mesures de dispersion empirique pour les courbes ROC, les résultats provenant d'auto-amorçage stratifié ne sont présent dans la littérature que pour l'aire sous la courbe ROC [Bandos, 2005]. Au meilleur de nos connaissances, l'auto-amorçage complet n'est, quant à lui, pas traité dans la littérature. Cette section du mémoire spécifiera donc notre apport au domaine des courbes ROC : la distribution d'auto-amorçage exacte stratifiée ponctuelle pour les courbes ROC.

4.2 Approximations

Dans les prochaines sections, plusieurs des variables que nous considérerons auront une distribution d'auto-amorçage exacte binomiale ou qui converge vers une binomiale aux extrémités de la courbe ROC. L'approche la plus commune est d'approximer la distribution binomiale par une distribution gaussienne et d'approximer la variance inconnue en utilisant $\hat{p}$, la proportion observée, plutôt que p, la vraie proportion de la population. Ainsi, au niveau de confiance 1 - α, les bornes de l'intervalle de confiance sont définies par l'équation :

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}. \tag{4.1}$$

où $Z_{1-\alpha/2}$ est le $(1 - \alpha/2)^e$ quantile de la distribution gaussienne et n est la taille de l'échantillon duquel $\hat{p}$ a été obtenu. Ces intervalles sont dits les intervalles de Wald puisqu'ils sont obtenus en inversant le test de normalité de Wald pour les grands échantillons.

Une alternative est d'utiliser l'intervalle de confiance de score. Ce dernier est obtenu similairement en inversant le test de normalité de score qui utilise la variance de l'hypothèse nulle : $p(1 - p)/n$. Dans ce cas, les bornes de l'intervalle de confiance de score sont :

$$\frac{\hat{p} + z^2_{1-\alpha/2}/2n \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + (z_{1-\alpha/2}/2n)^2}}{1 + z^2_{1-\alpha/2}/n}. \tag{4.2}$$

Le principal avantage des intervalles de score est que, contrairement aux intervalles de Wald, ils génèreront strictement des bornes admissibles (c'est-à-dire incluses dans l'intervalle [0;1]). De plus, les intervalles de score seront asymétriques par rapport à $\hat{p}$ et auront comme point milieu $\hat{p} + z^2_{1-\alpha/2}/2n$. Il faut ajouter que $\hat{p}$ fera toujours parti de l'intervalle.

Il est intéressant de remarquer que, pour $\hat{p} \in \{0,1\}$, l'intervalle de Wald sera de taille nulle puisque la variance estimée sera de 0. Ce problème ne se pose pas dans le cas de l'intervalle de score. En pratique, plus la vraie probabilité p est proche de 0 ou de 1, et plus la taille de l'échantillon est petite, plus il y aura d'intervalles de Wald de taille 0. Ceci explique les bris au niveau de la couverture des intervalles de Wald dans les extrémités de la courbe ROC et pour les petits échantillons.

Pour ces raisons, les approximations gaussiennes des résultats de ce chapitre sont faites à partir de la méthode du score.

4.3 Moyennage par seuil

Cette sous-section explicitera nos résultats dans le cas où on considère la prise de la moyenne par le seuil. Elle sera divisée en deux parties. La première traitera le cas où la performance d'un seul modèle est évaluée (*single design*). La comparaison de modèles en utilisant des jeux de données disjoints (*unpaired designs*) utilise cette théorie. La seconde partie comparera des modèles à l'aide d'un même jeu de données (*paired designs*). Un cas hybride, où certaines instances sont utilisées pour évaluer la performance des deux modèles, alors que d'autres ne s'appliquent qu'à l'un des deux, pourrait être obtenu en combinant les résultats des deux scénarios traités dans cette section. Il ne sera pas question de ce scénario ici, mais Metz et al. (1998) traite de ce type de cas.

4.3.1 Jeux de données uniques (*single design*)

Considérons un jeu de données de n instances. Des échantillons d'auto-amorçage stratifiés de taille m (usuellement $m = n$) sont tirés de ce jeu. Soit n^+ et n^-, les nombres d'instances positives et négatives du jeu initial. On définit X^+ et X^-, l'ensemble de tous les échantillons d'auto-amorçage qui peuvent être tirés des n^+ instances positives et des n^- instances négatives respectivement. L'ensemble de tous les échantillons d'auto-amorçage qui peuvent être tirés du jeu initial est donc $\{(x^+,x^-),\ x^+ \in X^+,\ x^- \in X^-\}$.

Dorénavant, toutes valeurs dépendant de x^+ ou de x^- seront accompagnées d'un symbole + ou – en exposant, respectivement.

Puisque nous utilisons la procédure d'auto-amorçage stratifié, les nombres d'instances positives et négatives des échantillons seront fixes et vaudront $m^+ = m \cdot n^+ / n$ et $m^- = m \cdot n^- / n$. Nous faisons l'hypothèse $m^+, m^- \in \mathbb{N}$. Si ce n'était pas le cas, une procédure additionnelle serait requise afin de s'assurer que l'échantillon stratifié reflète précisément les vraies proportions du jeu initial.

Posons n_t^+ comme étant le nombre d'instances, parmi les n^+ positives du jeu de données, qui génère un score plus grand ou égal à t. Définissons M_t^+ comme la variable aléatoire représentant le nombre d'instances positives ayant un score plus grand ou égal à t dans un échantillon aléatoire d'auto-amorçage de taille m^+ tel que défini précédemment. Dans ce cas, M_t^+ suivra une distribution binomiale de paramètres m^+ et $p_t^+ = n_t^+ / n^+$: $M_t^+ \sim Bin(m^+, p_t^+)$. En conséquence, $TP_t^+ = M_t^+ / m^+$ sera la variable aléatoire pour le ratio de vrais positifs au seuil t. Puisque m^+ est fixe sur l'ensemble des échantillons, TP_t^+ a pour espérance et variance :

$$E(TP_t^+) = p_t^+, \quad Var(TP_t^+) = p_t^+(1-p_t^+)/m^+ . \qquad (4.3)$$

Similairement, pour les instances négatives, n_t^- représentera le nombre d'instances avec un score supérieur ou égal à t parmi les n^- du jeu initial. Aussi, M_t^- sera la variable aléatoire pour le nombre d'instances négatives, provenant d'un échantillon de taille m^-, ayant un score plus grand ou égal à t. Au seuil t, la probabilité qu'une instance négative obtienne un score d'au moins t est $p_t^- = n_t^- / n^-$. On aura donc $FP_t^- = M_t^- / m^-$ avec $M_t^- \sim Bin(m^-, p_t^-)$ comme variable aléatoire du ratio de faux positifs. L'espérance et la variance de cette variable seront donc :

$$E(FP_t^-) = p_t^-, \quad Var(FP_t^-) = p_t^-(1-p_t^-)/m^- . \qquad (4.4)$$

Puisque l'auto-amorçage stratifié tire $x^+ \in X^+$ et $x^- \in X^-$ indépendamment, il en résulte que TP_t^+ et FP_t^- sont aussi indépendants.

L'algorithme 4.1 de l'annexe B exprime comment calculer ces intervalles de confiance pour un ensemble de h seuils donnés. Le temps de calcul de cet algorithme est de l'ordre $O(n \ln(n))$. Le logarithme provient de la nécessité d'ordonner les instances en ordre décroissant. Si les instances sont fournies directement ordonnées, alors le temps devient linéaire par rapport à n.

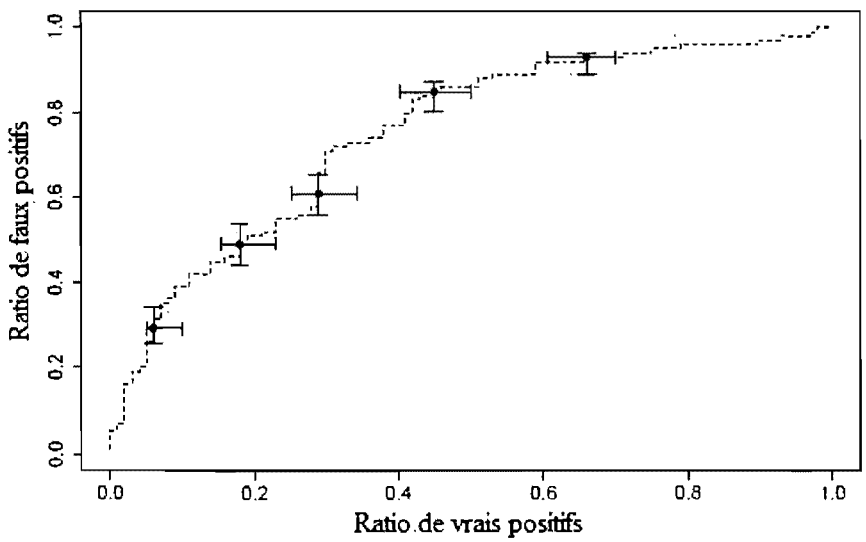

Fig. 4.1 Courbe ROC empirique avec les intervalles de confiance ponctuels pour la performance du modèle. Le moyennage par seuil est utilisé pour 5 seuils. Toutes les valeurs sont obtenues en utilisant l'algorithme 4.1 de l'annexe B.

Notons que, pour parvenir à ce temps de calcul, il faut considérer que la fonction de répartition binomiale (notée $B(k,n,p) = \sum_{j=0}^{k} b(j,n,p)$ où $b(j,n,p) = \binom{n}{j} p^j (1-p)^{n-j}$ est la probabilité d'obtenir j succès lors de n essais indépendants d'une binomiale de probabilité p) peut être approximer numériquement [Press et al., 2007] et donc être

obtenue dans un temps d'ordre constant O(1). En particulier, la fonction *pbinom* du langage R est obtenue en temps constant.

4.3.2 Jeux de données combinés (*paired designs*)

Lorsqu'on veut comparer la performance de deux modèles, il faut faire attention lors de la sélection des seuils puisque les scores obtenus pour ces modèles n'ont pas toujours la même signification. Par exemple, si un modèle assigne des scores entre 0 et 100, alors qu'un autre n'assigne que des scores supérieurs à 100, comparer leur performance pour des seuils fixés n'est pas pertinent. Même si les scores sont dans le même intervalle, leur distribution peut être différente, ce qui rend la comparaison inefficace. Une façon de régler ce problème est de calibrer les scores [Fawcett et Niculescu-Mizil, 2007; Platt, 2000; Zadrozny et Elkan, 2002]. Cette méthode consiste à trouver la bijection entre les scores initiaux d'un modèle et la probabilité qu'une instance soit positive. Cette dernière valeur devient le score calibré.

La calibration a pour but d'aligner les scores afin qu'on puisse faire une comparaison valable des performances pour des seuils fixés. L'avantage de cette méthode est qu'elle peut se faire de manière automatique, sans intervention de l'usager. Cependant, cette automatisation peut aussi être vue comme un désavantage si l'usager désire comparer la performance de modèles pour des seuils non alignés avant, ou même après, la calibration.

Afin de couvrir les deux approches, nous considérons des ensembles de h paires de seuils (t_1,t_2). Dans le cas où les scores sont calibrés, $t_1 = t_2$ pour chaque paire. Sinon, t_1 et t_2 représentent les seuils auxquels l'usager décide de comparer les deux modèles.

Une fois t_1 et t_2 fixés, les ratios de vrais et de faux positifs peuvent varier d'un modèle à l'autre. Il devient donc impossible, à moins de connaître les conditions d'opération, de choisir entre deux modèles si les deux ratios d'un modèle sont inférieurs à ceux de l'autre modèle. Cependant, il est possible d'évaluer la probabilité, notée $f_1(t_1,t_2)$, qu'aux seuils t_1 et t_2, le modèle 1 domine le modèle 2. On dit qu'il y a dominance du modèle 1 sur le modèle 2 lorsque ${}_1FP^-_{t_1}$ est plus petit ou égal à ${}_2FP^-_{t_2}$ alors que ${}_1TP^+_{t_1}$ est supérieur ou égal à ${}_2TP^+_{t_2}$. Le cas où les deux types de ratios sont égaux d'un modèle à l'autre doit être exclu afin d'avoir la dominance stricte. Puisque x^+ et x^- sont tirés indépendamment, la probabilité jointe peut être exprimée comme le produit des probabilités marginales :

$$f_1(t_1,t_2) = \Pr\{\Delta_{1.2}TP^+_{t_1,t_2} \geq 0\}\ \Pr\{\Delta_{1.2}FP^-_{t_1,t_2} \leq 0\} - \Pr\{\Delta_{1.2}TP^+_{t_1,t_2} = 0\}\Pr\{\Delta_{1.2}FP^-_{t_1,t_2} = 0\}. \quad (4.5)$$

L'algorithme 4.2 de l'annexe B permet de calculer les valeurs de $f_1(t_1,t_2)$.

Lorsqu'on s'intéresse à la différence entre les ratios de vrais positifs de deux modèles avec jeux de données combinés, les variations proviennent d'instances où les modèles sont en désaccord. En conséquence, il y aura trois catégories d'instances qui détermineront la distribution des différences des ratios de vrais positifs. Ces catégories sont :

(a) le modèle 1 génère un score supérieur ou égal à t_1 alors que le modèle 2 est inférieur à t_2;

(b) le modèle 1 génère un score inférieur à t_1 alors que le modèle 2 est supérieur ou égal à t_2; et

(c) les deux modèles sont en accord sur la classification de l'instance.

Soit $n^{\pm}_{t_1,t_2}$ et $n^{+}_{t_1,t_2-}$ le nombre d'instances positives des catégories (a) et (b), respectivement. Les instances négatives ont une notation similaire en remplaçant le signe + par le signe -. Lorsqu'on performe l'auto-amorçage stratifié, la distribution jointe du nombre d'instances positives tirées des trois catégories est un trinôme. La différence entre les ratios de vrais positifs sera 0 chaque fois qu'on aura $n^{\pm}_{t_1,t_2} = n^{+}_{t_1,t_2-}$. En conséquence, la probabilité que la différence des ratios de vrais positifs soit nulle est une somme de probabilités trinomiales et peut être calculée en temps d'ordre linéaire. Si on s'intéresse à une inégalité (la différence des ratios de vrais positifs est supérieure, supérieure ou égale, inférieure, inférieure ou égale, à 0), la différence sera obtenue à l'aide d'une double sommation de probabilités trinomiales. Les propriétés des probabilités binomiales permettent de développer une équation récursive permettant de calculer les probabilités des inéquations dans un temps d'ordre linéaire. On peut traiter les instances négatives de la même façon et dans le même temps afin d'obtenir la distribution de la différence des ratios de faux positifs. Puisqu'il faudra, pour obtenir la courbe ROC, calculer cela pour chaque seuil et ordonner les scores, le temps de calcul sera de l'ordre $O(n \cdot \max(h, \ln(n)))$.

Malheureusement, évaluer la supériorité de la performance d'un modèle par rapport à un autre à l'aide de l'équation (4.5) est un test avec bien peu de puissance. La cause principale de ce phénomène est que les probabilités $\Pr\{\Delta_{1,2}TP^{+}_{t_1,t_2} < 0, \Delta_{1,2}FP^{-}_{t_1,t_2} < 0\}$ et $\Pr\{\Delta_{1,2}TP^{+}_{t_1,t_2} > 0, \Delta_{1,2}FP^{-}_{t_1,t_2} > 0\}$ sont généralement suffisamment larges pour que les tests soient non concluants selon les standards habituels (c'est-à-dire avec des valeurs-p sous 1, 5 ou même 10%).

En définissant

$\tilde{p}^{+}_{t_1,t_2} = n^{+}_{t_1,t_2}/n^{+}$, $\tilde{p}^{\pm}_{t_1,t_2} = n^{\pm}_{t_1,t_2}/n^{+}$, $\tilde{\delta}_{t_1,t_2} = \tilde{p}^{+}_{t_1,t_2} - \tilde{p}^{\pm}_{t_1,t_2}$ *et* $\tilde{\gamma}_{t_1,t_2} = \tilde{p}^{+}_{t_1,t_2} + \tilde{p}^{\pm}_{t_1,t_2}$, on

obtient, pour la distribution des ratios de vrais positifs, les estimateurs suivants :

$$E\{\Delta_{1,2}TP^{+}_{t_1,t_2}\} = \tilde{\delta}_{t_1,t_2}, \tag{4.6}$$

$$Var\{\Delta_{1,2}TP^{+}_{t_1,t_2}\} = \frac{\tilde{\gamma}_{t_1,t_2} + \tilde{\delta}^{2}_{t_1,t_2}}{m^{+}}. \tag{4.7}$$

Les moments pour la distribution des ratios de faux positifs sont obtenus identiquement en utilisant les valeurs correspondantes pour les scores négatifs.

Lorsqu'on inverse le test de score afin d'obtenir l'intervalle de confiance pour $\delta_{t_1,t_2} = p^{+}_{t_1,t_2} - p^{\pm}_{t_1,t_2}$, la vraie différence entre les ratios de vrais positifs, on obtient les bornes suivantes :

$$\delta_{t_1,t_2} = \frac{\tilde{\delta}_{t_1,t_2} \pm z\sqrt{\left(\gamma_{t_1,t_2}(1 + z^2/m^{+}) - \tilde{\delta}^{2}_{t_1,t_2}\right)/m^{+}}}{1 + z^2/m^{+}}. \tag{4.8}$$

On remarquera que cette équation contient le paramètre de nuisance $\gamma_{t_1,t_2} = p^{+}_{t_1,t_2} + p^{\pm}_{t_1,t_2}$, une somme de probabilités non observées. Il est possible d'utiliser l'estimé d'auto-amorçage de $\tilde{\gamma}_{t_1,t_2}$, mais nous avons déterminé de manière empirique que la couverture était meilleure en utilisant l'estimateur lissé :

$$\hat{\gamma}_{t_1,t_2} = \frac{n^{+}_{t_1,t_2} + n^{\pm}_{t_1,t_2} + 2}{n^{+} + 4}. \tag{4.9}$$

4.4 Moyennage vertical

4.4.1 Jeux de données uniques (*single design*)

Il faut évaluer la distribution d'auto-amorçage exacte du ratio de vrais positifs pour chaque élément de l'ensemble de h ratios de faux positifs : $\{r_i/m^-,\ 1 \leqslant r_1 \leqslant r_2 \leqslant \ldots \leqslant r_h \leqslant m^- - 1,\ i = 1,2,\ldots,h\}$. Posons TP_r^+, la variable aléatoire du ratio de vrais positifs lorsque le ratio de faux positifs vaut r/m^- et M_r^+, la variable aléatoire du nombre d'instances positives correctement classifiées pour ce même ratio de faux positifs. Il en découle que $TP_r^+ = M_r^+ / n^+$. Il est important de remarquer que M_r^+ ne dépend pas seulement de x^+ mais aussi de x^- par l'intermédiaire du ratio de faux positifs.

Étant donné un échantillon d'auto-amorçage stratifié (x^+,x^-) et un ratio de faux positifs r / m^- fixé, il y a r instances négatives, celles ayant les plus hauts scores, qui sont mal classifiées. Le plus faible de ces scores, noté T_r^-, est le seuil qui donnera un ratio de faux positifs r / m^-. Une fois cette valeur déterminée, il devient possible de calculer la distribution des ratios de vrais positifs conditionnelle à la valeur T_r^-. En intégrant sur l'ensemble de la distribution de T_r^-, on obtiendra la distribution inconditionnelle de TP_r^+.

Soit $s_1 \geqslant s_2 \geqslant \ldots \geqslant s_{n^-}$, les scores des instances négatives en ordres décroissant. Étant donné un échantillon d'auto-amorçage stratifié $(x^+,\ x^-)$, si au moins r des m^- instances négatives de l'échantillon x^- sont tirées parmi les k instances négatives ayant le plus grand score, alors le seuil T_r^- est plus grand ou égal à s_k. Nous pouvons donc déduire l'équation suivante :

$$\Pr\{T_r^- \geq s_k\} = \sum_{j=r}^{m^-} b(k/n^-, j, m^-)$$

$$= 1 - B(k/n^-, r-1, m^-). \qquad (4.10)$$

La distribution du seuil T_r^- est donc :

$$\Pr\{T_r^- = s_k\} = \Pr\{T_r^- \geq s_k\} - \Pr\{T_r^- \geq s_{k-1}\}$$

$$= B\left(\frac{k-1}{n^-}, r-1, m^-\right) - B\left(\frac{k}{n^-}, r-1, m^-\right) \qquad (4.11)$$

Tel qu'expliqué à la sous-section 4.3.1, on peut supposer que le calcul de la distribution binomiale s'effectue en temps constant. On obtient donc $\Pr\{T_r^- = s_k\}$ dans un temps constant.

Posons maintenant n_k^+, le nombre d'instances positives, parmi les n^+ instances positives de l'ensemble de test, ayant un score supérieur ou égal à s_k. Posons aussi M_k^+, la variable aléatoire du nombre d'instances ayant un score supérieur ou égal à s_k parmi les m^+ de l'échantillon d'auto-amorçage stratifié x^+. Dans ce cas, $M_k^+ \sim \text{Bin}(m^+, p_k^+)$ où $p_k^+ = n_k^+ / n^+$. On a donc la distribution inconditionnelle des ratios de vrais positifs suivante :

$$\Pr\{TP_r^+ = l/\mathrm{m}^+\} = \sum_{k=1}^{n^-} \Pr\{T_r^- = s_k\} \Pr\{M_k^+ = l\}$$

$$= \sum_{k=1}^{n^-} \Pr\{T_r^- = s_k\} b(l, m^+, p_k^+). \qquad (4.12)$$

Puisque chaque probabilité est obtenue dans un temps d'ordre linéaire pour chaque valeur de l dans $\{0,1,\ldots,m^+\}$ et pour chaque ratio de faux positifs h, on peut obtenir la distribution du ratio de vrais positifs pour toutes les valeurs de h dans un temps de l'ordre $O(n^2 \cdot h)$.

Une fois ces informations obtenues, on peut représenter les distributions d'auto-amorçage exactes à l'aide de ces deux premiers moments et d'une approximation gaussienne. Au ratio de faux positifs r/m^-, on obtient :

$$E\{TP_r^+\} = \sum_{k=1}^{n^-} \Pr\{T_r^- = s_k\} E_X\{M_k^+ / m^+\}$$

$$= \sum_{k=1}^{n^-} \Pr\{T_r^- = s_k\} p_k^+ \qquad (4.13)$$

et comme second moment :

$$E\{TP_r^{+2}\} = \sum_{k=1}^{n^-} \Pr\{T_r^- = s_k\} \mathrm{E}_X\{(M_k^+ / m^+)^2\}$$

$$= \sum_{k=1}^{n^-} \Pr\{T_r^- = s_k\} \left[p_k^{+2} + p_k^+ (1 - p_k^+)/m_k^+ \right] \qquad (4.14)$$

Ces valeurs sont obtenues dans un temps d'ordre linéaire pour chaque ratio de faux positifs. Il faut encore une fois ordonner les instances. En conséquence, l'intervalle de confiance ponctuel est obtenu dans un temps d'ordre $O(n \cdot \max(h, \ln(n)))$.

L'algorithme 4.3 de l'annexe B présente comment faire ces calculs.

4.4.2 Jeux de données combinés (*paired designs*)

Posons $T_{1,r}^-$ et $T_{2,r}^-$, les variables aléatoires représentant les seuils où le ratio de faux positifs est r/m^- dans le premier et le second modèle respectivement. Soit $s_{1,1} \geqslant s_{1,2} \geqslant \ldots \geqslant s_{1,n^-}$ et $s_{2,1} \geqslant s_{2,2} \geqslant \ldots \geqslant s_{2,q}$, les scores des instances négatives, en ordre décroissant, obtenus à l'aide du premier et du second modèle respectivement. Supposons $T_{1,r}^- = s_{1,k}$ et $T_{2,r}^- = s_{2,j}$. Il est important de réaliser que j n'est pas nécessairement égal à k. Par exemple, supposons un scénario où n^- = 4. Supposons aussi que les scores

associés à ces instances sont {1,2,3,4} selon le premier modèle et {11,12,14,13} selon le second. Un échantillon d'auto-amorçage de taille $m^- = 4$ est tiré et un ratio de faux positifs est fixé à $r / m^- = 25\%$, c'est-à-dire $r = 1$. Supposons maintenant que cet échantillon d'auto-amorçage est formé de la première instance deux fois, de la seconde instance puis de la troisième instance. Le premier modèle posera $T_{1,r}^-$ comme étant $s_{1,2} = 3$ puisque $s_{1,1} = 4$ ne fait pas partie de l'échantillon et que 3 devient donc la plus grande valeur de l'échantillon. Le second modèle, quant à lui, posera $s_{2,1} = 14$ comme valeur de $T_{2,r}^-$. On aura donc $k = 2$ et $j = 1$, et donc j n'égale pas k.

Le calcul de la distribution de ΔTP_r^+ se fera en trois étapes. Tout d'abord, dans un temps d'ordre linéaire, il faudra calculer la fonction de densité de chaque probabilité :

$$f_r(k, j) = \Pr\{T_{1,r}^- = s_{1,k}, T_{2,r}^- = s_{2,j}\}. \tag{4.15}$$

Puisque $f_r(k,j)$ est calculé pour chaque valeur de r, k et j, le temps de calcul est de l'ordre de $O(n^4)$. Ensuite, étant donné les seuils $T_{1,r}^-$ et $T_{2,r}^-$, chaque fonction de distribution conditionnelle pour ΔTP_r^+ est calculée en temps d'ordre linéaire :

$$g_{k,j}(d) = \Pr\{\Delta TP_r^+ = d/m^+ \mid T_{1,r}^- = s_k, T_{2,r}^- = s_j\}. \tag{4.16}$$

Comme $g_{k,j}(d)$ est calculé pour chaque valeur de k, j et d, le temps de calcul est encore de l'ordre $O(n^4)$. Finalement, on calcule, en temps d'ordre quadratique, chaque valeur de la distribution inconditionnelle de ΔTP_r^+ :

$$h_r(d) = \sum_{k,j} \left(f_r(k, j).g_{k,j}(d) \right) \tag{4.17}$$

Ici encore, comme $h_r(d)$ est calculée pour toutes les valeurs de r et de d, le temps de calcul est de l'ordre $O(n^4)$. L'algorithme 4.4 de l'annexe B résume l'obtention de ces étapes.

Malheureusement, des calculs de l'ordre $O(n^4)$ sont peu pratiques lors d'applications réelles. Afin d'améliorer l'efficacité de l'algorithme, il faudrait approximer la distribution jointe des seuils. De plus, nous avons observé que la majeure partie des entrées de la matrice de fonction de masse jointe prend des valeurs très près de 0. Les calculs pourraient donc être accélérés en approximant la pleine matrice par une version creuse (*sparse*).

4.5 Simulations numériques

La suite de ce chapitre sera divisée en quatre parties. Chacune d'entre elles sera composée d'expériences traitant l'une des quatre situations couvertes dans les sections 4.3 et 4.4. Ces expériences mesurent la performance en terme de précision de couverture. Elles sont inspirées de Hall et al. (2004) et de Macskassy et al. (2005) à la différence que, dans le cas du moyennage par seuil, on rapporte la précision de couverture en fonction du ratio total de positifs (le nombre d'instances du test classifiées positives divisé par le nombre total d'instance du jeu de données) plutôt qu'en fonction des seuils. Ceci nous permet de faire des comparaisons indépendantes des échelles de scores.

4.5.1 Moyennage par seuil et jeux de données uniques

Comme première expérience, nous reproduisons celles utilisées par Hall et al. (2004) afin de déterminer la robustesse de la précision de couverture sous leur approche de lissage par noyau en fonction de la forme de la distribution des scores. Nous l'appellerons l'expérience de forme.

Posons G^+ et G^-, les distributions des scores pour les instances positives et négatives respectivement. Nous avons testé les 5 paires de distributions suivantes :

– Exemple 1: $G^+ \sim N(1, 1)$, $G^- \sim N(0, 1)$;

– Exemple 2: $G^+ \sim N(2, 2)$, $G^- \sim N(0, 1)$;

– Exemple 3: $G^+ \sim \beta(2, 4)$, $G^- \sim \beta(2, 3)$;

– Exemple 4: $G^+ \sim \beta(1.2, 2)$, $G^- \sim \beta(1.2, 3)$;

– Exemple 5: $G^+ \sim \exp(3)$, $G^- \sim \exp(2)$.

Il est à noter que la courbe ROC théorique associée aux distributions de l'exemple 3 est inadmissible. Cet exemple sera tout de même considéré afin de vérifier la robustesse des techniques considérées lorsque utilisées dans le contexte d'une courbe inadmissible.

Nous avons fait 1000 simulations pour chacun des 5 exemples ci-dessus. Pour chaque simulation, un échantillon de 100 instances positives tirées de G^+ et de 100 instances négatives tirées de G^- est considéré. Les intervalles de confiance, calculés selon l'algorithme 4.1 au niveau de confiance $\alpha = 5\%$ avec un ensemble de 99 seuils correspondant aux ratios totaux de positifs $\{1\%,2\%,\ldots,99\%\}$, ont été construits pour chaque simulation. La figure 4.2 expose les résultats obtenus. On remarque que, dans tous les cas, la couverture est assez proche de la valeur cible sauf aux extrémités où il

semble y avoir une plus grande fluctuation. Ce phénomène est expliqué dans notre deuxième expérience.

Cette seconde expérience sera une reproduction de celle de Macskassy et al. (2005). Il s'agit de l'application de 4 méthodes pour obtenir des intervalles de confiance ponctuels. Cette fois, les scores des instances positives, comme des négatives, suivront des distributions normales. Cependant, les paramètres de ces distributions varieront. Macskassy et al. (2005) fixent les paramètres d'échelles à 3,75 pour les instances positives et à 3,00 pour les négatives. Le niveau de confiance est α = 10%. Le paramètre de localisation θ pour les instances positives varient dans l'ensemble {0,75; 1,5; 3,0; 5,0}. Celui des instances négatives est posé égal à $-\theta$. La taille des échantillons est fixée à 10 000, c'est-à-dire qu'un échantillon de 10 000 instances est tiré de la distribution d'instances positives et qu'un autre échantillon de 10 000 est tiré de la distribution des négatives. Pour chaque valeur de θ, 1000 simulations sont produites. Nous appellerons cette expérience : l'expérience de dispersion. La figure 4.3 présente les résultats de cette expérience.

Premièrement, il faut noter que l'asymétrie dans les graphiques est due à la différence entre les paramètres d'échelle. De plus, on remarque que les résultats sont meilleurs lorsque la section où les deux distributions se superposent est plus importante (lorsque θ est petit). Cela suggère que la couverture est plus précise pour les problèmes « difficiles ». Ce résultat contre-intuitif peut être expliqué par un exemple numérique simple.

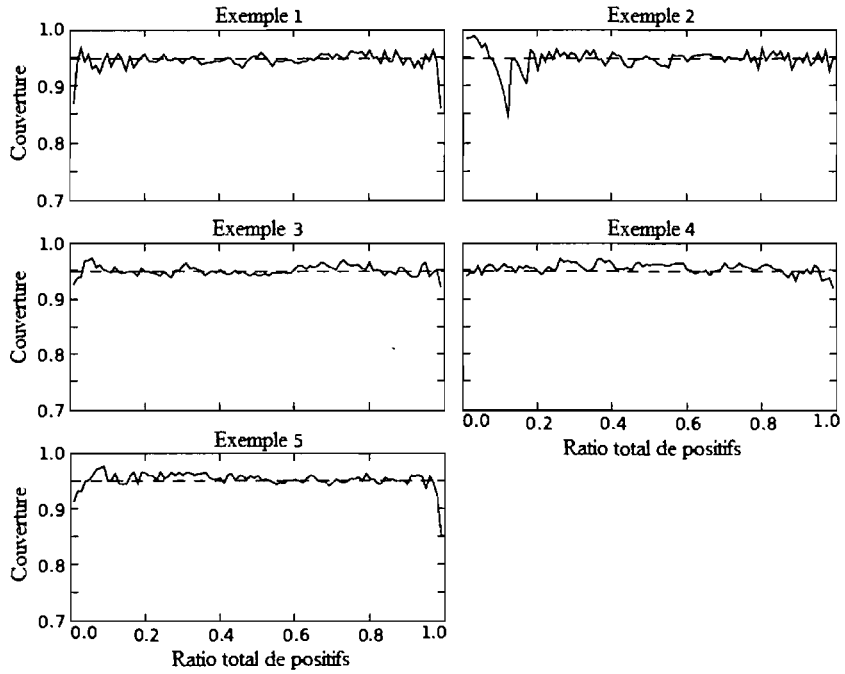

Fig. 4.2 Effet de la forme des distributions sur la couverture. Les intervalles de confiances sont construits au niveau de confiance $\alpha = 5\%$. La proportion de couverture (ligne pleine) pour 1000 simulations et la couverture cible (en pointillé) sont illustrées en fonction du ratio total de positifs.

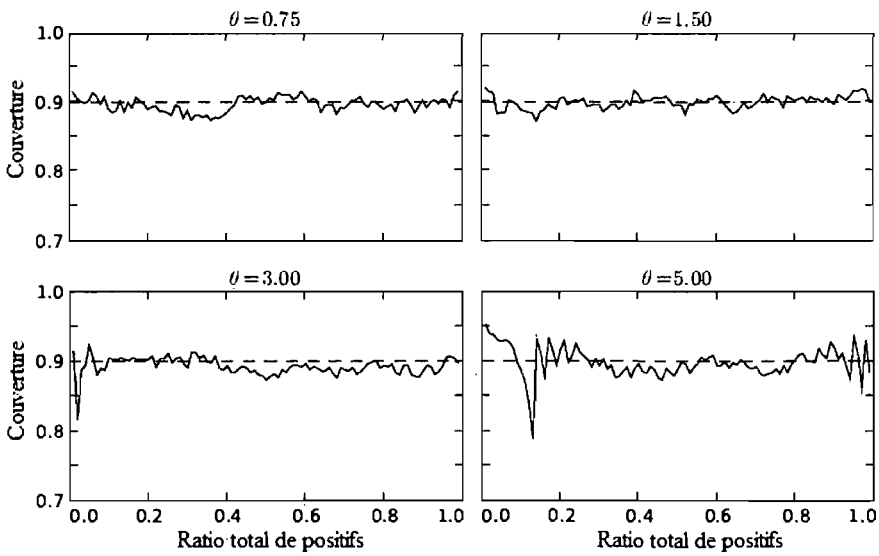

Fig. 4.3 Effet de la dispersion entre les distributions normales sur la couverture. Le paramère de localisation varie entre {0,75; 1,50; 3,00; 5,00}. La taille des échantillons est de 10 000. L'intervalle de confiance est calculé au niveau $\alpha = 10\%$. La proportion de couverture (ligne pleine), pour 1000 simulations, et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction du ratio total de positifs.

Considérons le cas où $\theta = 5$. La couverture la plus faible se trouve au ratio total de positifs 0,13. Dans ce cas, les valeurs théoriques cibles étaient de 0,25998 pour le ratio de vrais positifs et de $1{,}7549e^{-5}$ pour le ratio de faux positifs. Le seuil est de 7,4127. En d'autres termes, la probabilité qu'une instance négative ait un score supérieur à 7,4127 est $1{,}7549e^{-5}$. Pour un échantillon de taille 10 000, la probabilité d'obtenir 0 faux positifs est donc de $(1 - 1{,}7549e^{-5})^{10\,000} = 0{,}8390$ et la probabilité d'obtenir un seul faux positifs est $10\,000 \cdot 1{,}7549e^{-5}(1 - 1{,}7549e^{-5})^{9999} = 0{,}1472$. L'intervalle de confiance de score d'un échantillon de 10 000, performé tel que dans l'algorithme 4.1, pour le ratio de faux positifs est $[0;\ 3{,}37965e^{-4}]$ si aucun faux positif n'est observé et $[1{,}7794e^{-5};\ 5{,}6178e^{-4}]$ si un seul faux positif est observé. En conclusion, le premier intervalle contient le vrai ratio de faux positifs alors que le second ne le contient pas. La probabilité de couverture du vrai ratio de faux positifs est donc de 0,8390. La probabilité que le ratio de vrais positifs soit couvert est, ici, de 0,9487. On obtient donc une couverture des deux valeurs avec probabilité 0,7960. Cette valeur est près de la valeur observée par nos simulations (0,788), mais considérablement plus basse que la valeur ciblée de 0,9.

Si on refait cette expérience pour un ratio total de positifs de 0,14, on obtient une valeur couverte dès que le nombre observé de faux positifs est inférieur à 2. Cela s'exprime par la hausse abrupte de probabilité de couverture du graphique.

En d'autres termes, cette situation se produit parce que la dispersion entre les échantillons est telle qu'on considère des ratios de faux positifs excessivement petits et qu'on assigne donc trop de poids à un petit intervalle de faux positifs observés. Par exemple, dans notre situation, la couverture n'a lieu que si aucun faux positif n'est

observé. Au fur et à mesure qu'on considère de plus larges valeurs de ratios de faux positifs, la couverture s'obtient pour un plus grand intervalle de valeurs observées et la fluctuation de la couverture a donc moins d'amplitude.

Il faut toutefois noter que notre interprétation stipulant que les résultats sont meilleurs pour des problèmes « difficiles » est directement liée à notre façon de comparer les graphiques. Nous avons choisi de comparer les performances en fonction du ratio total de positifs. Cela nous force à regarder le problème à différents intervalles de seuils. Par exemple, avec $\theta = 0{,}75$, les seuils se trouvent dans l'intervalle [-7,8430; 8,5299] alors qu'avec $\theta = 5$, l'intervalle s'élargit à [-11,1618; 12,7016]. Les conclusions seraient probablement différentes si on comparait les graphiques en fonction de seuils égaux.

Dans une troisième expérience, nous considérons l'effet de la taille de l'échantillon sur la précision de couverture. Cette expérience provient aussi de Macskassy et al. (2005). Elle est en tout point similaire à l'expérience de dispersion à deux exceptions près. Tout d'abord, le paramètre de localisation ne varie plus et est fixé à $\theta = 3{,}0$. Ensuite, la taille de l'échantillon n'est plus fixe mais varie dans l'ensemble {25, 250, 2500, 10 000}. Nous la nommons : l'expérience de taille. La figure 4.4 présente les résultats. Cette fois, l'amplitude des fluctuations dans la précision de la couverture est plus grande pour les échantillons de petites tailles. Ce qui n'est pas surprenant.

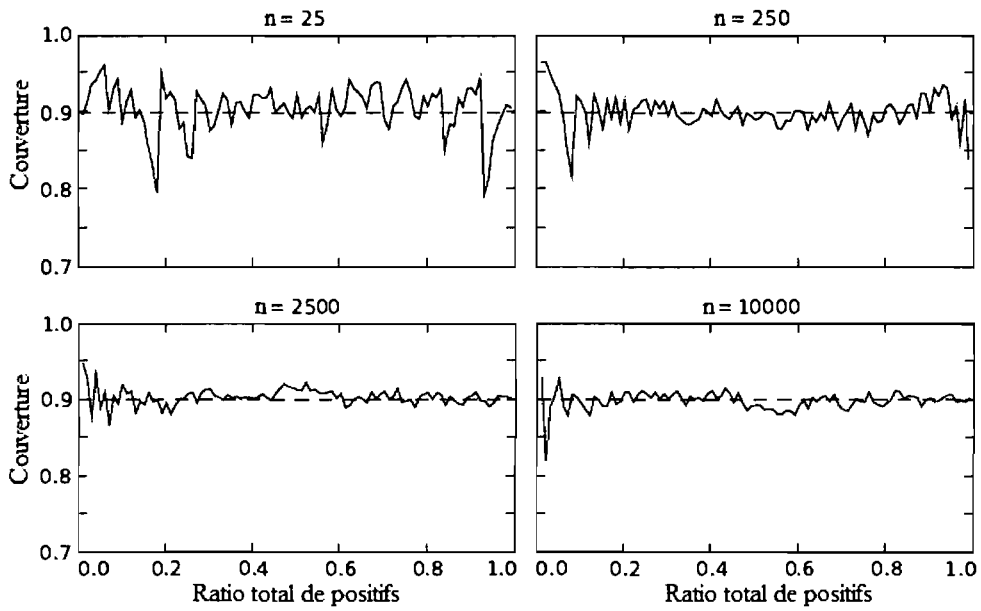

Fig. 4.4 Effet de la taille de l'échantillon sur la couverture. Les tailles varient dans l'ensemble {25, 250, 2500, 10 000}. Les intervalles de confiance sont obtenus au niveau de confiance $\alpha = 10\%$. Le paramètre de localisation pour les instances positives est $\theta = 3{,}0$. La proportion de couverture (ligne pleine) pour 1000 simulations et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction du ratio total de positifs.

4.5.2 Moyennage par seuil et jeux de données combinés

Cette fois, il est question de la précision de couverture pour les intervalles de confiance sur la différence entre deux courbes ROC à des seuils fixés. Étant donné un niveau de confiance α et puisque $\Delta_{1,2}TP^{+}_{t_1,t_2}$ et $\Delta_{1,2}FP^{-}_{t_1,t_2}$ sont indépendants, un intervalle de confiance de taille $\sqrt{1-\alpha}$ est défini pour chaque variable. L'intersection de ces deux intervalles de confiance est une région de confiance bidimensionnelle rectangulaire de taille 1 - α. Les intervalles de confiance de score sont obtenus à l'aide des équations (4.8) et (4.9).

La modélisation de l'expérience est similaire à celles utilisées pour l'expérience de taille et de distribution. Les scores sont distribués selon une loi binormale avec un

paramètre d'échelle de 3,75 pour les instances positives et de 3,00 pour les négatives. Les intervalles de confiance sont obtenus au niveau $\alpha = 10\%$. Les paramètres de localisation sont déterminés ainsi : pour les instances positives du premier modèle, nous considérons deux valeurs : $\theta \in \{1,0; 3,0\}$. Les instances négatives des deux modèles sont posées égales à $-\theta$. Finalement, les instances positives du second modèle prennent trois valeurs : θ, $\theta + 2,0$ et $\theta + 4,0$. La différence entre les paramètres de localisation des distributions d'instances positives pour les deux modèles est, soit 0,0, soit 2,0 ou encore 4,0 et on y fera référence comme étant le paramètre de changement. Afin d'inclure de la dépendance entre les scores des deux modèles, trois facteurs de corrélations sont considérés : $\rho \in \{0,3; 0,6; 0,9\}$. La figure 4.5 illustre les résultats.

Une tendance très claire émerge dans tous les graphiques : les intervalles de confiance deviennent trop conservateurs lorsque le ratio total de positif se rapproche de 0 ou 1. De plus, nous observons que la couverture est moins stable lorsque la distribution et/ou la corrélation est plus grande. En contrepartie, le paramètre de changement semble n'avoir virtuellement aucun effet sauf lorsque $\rho = 0,9$ où on peut voir que les couvertures sans paramètre de changement (lignes bleues) sont encore plus conservatrices.

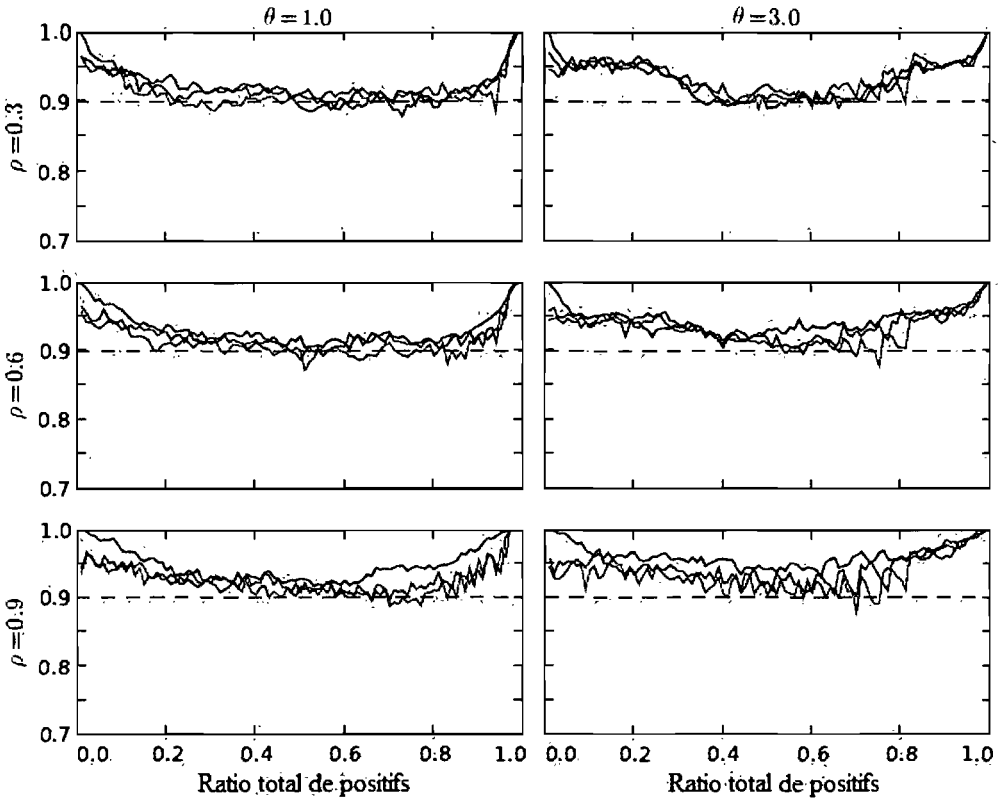

Fig. 4.5 Précision de couverture de régions de confiance bidimensionnelles pour la différence de performance entre deux courbes ROC sous le moyennage par seuil. La taille des échantillons est 100 et le niveau de confiance est $\alpha = 10\%$. Le paramètre de localisation pour les instances positives du premier modèle est $\theta = 1{,}0$ (à gauche) et $\theta = 3{,}0$ (à droite). Le facteur de corrélation est $\rho = 0{,}3$ (en haut), $\rho = 0{,}6$ (au centre) et $\rho = 0{,}9$ (en bas). Le paramètre de localisation pour le score des instances positives du second modèle est θ (en bleu), $\theta + 2{,}00$ (en vert) et $\theta + 4{,}00$ (en rouge). La proportion de couverture pour 1000 simulations et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction du ratio total de positifs.

4.5.3 Moyennage vertical et jeux de données uniques

Cette section couvrira la précision de couverture pour les intervalles de confiance utilisant l'algorithme 4.3. Les expériences de forme, de distribution et de taille, telles que décrites en 4.5.1 sont présentées. Nous comparons ici les résultats de l'approche proposée avec ceux provenant du lissage par noyau. Les noyaux gaussiens sont utilisés avec des paramètres de fréquence sélectionnés à l'aide du sélecteur de référence proposé par Hall et Hyndman (2003).

Les résultats pour les expériences de forme, de distribution et de taille sont présentés dans les figures 4.6, 4.7 et 4.8 respectivement. Dans le cas de l'expérience de forme, les deux méthodes ont des performances similaires. Cependant, pour les deux autres expériences, on remarque que lorsque la méthode d'auto-amorçage exacte proposée souffre de grande fluctuation, la couverture avec la méthode de lissage par noyau chute drastiquement. La méthode proposée emmène donc une amélioration réelle (une fluctuation plutôt qu'une brisure totale).

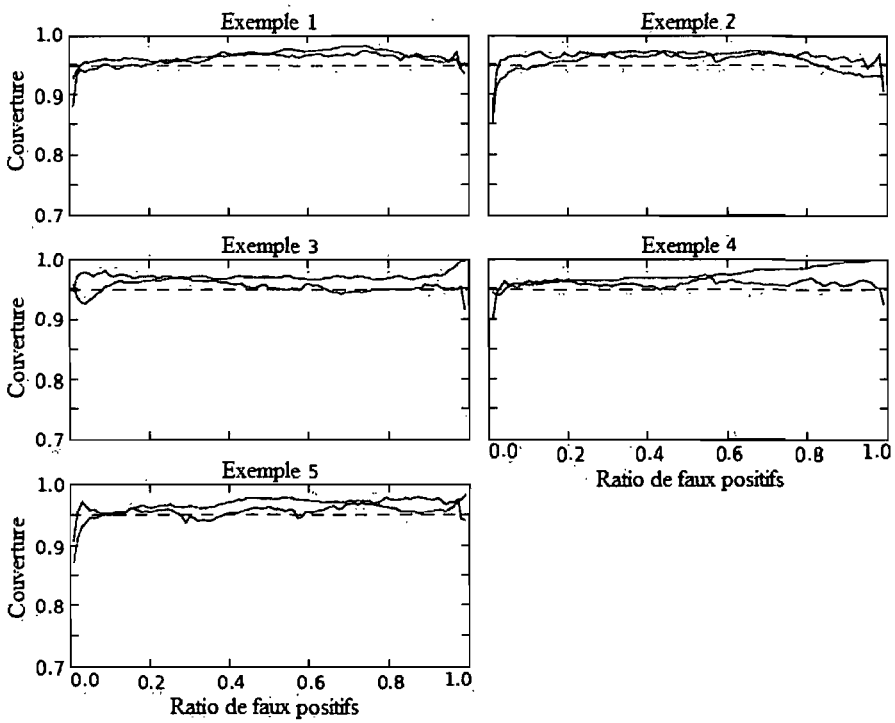

Fig. 4.6 Effet de la forme de la distribution sur la couverture. Les intervalles de confiance sont construits au niveau de confiance $\alpha = 5\%$. La proportion de couverture pour 1000 simulations utilisant l'approche d'auto-amorçage exacte proposée (en noir), la méthode de lissage par noyau (en bleu) et la couverture cible de 95% (ligne pointillée) son illustrées en fonction du ratio de faux positifs.

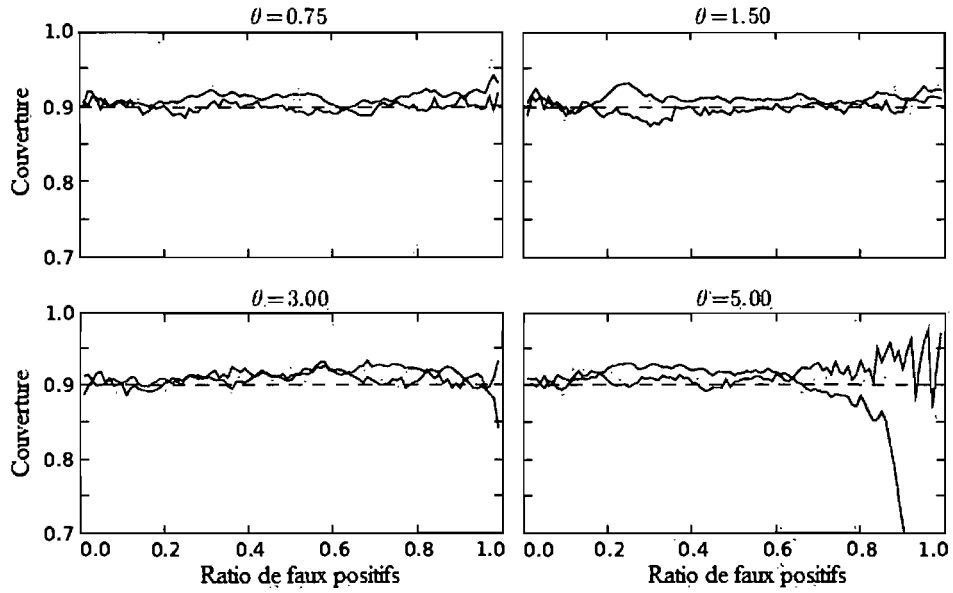

Fig. 4.7 Effet de la dispersion entre les distributions sur la couverture. Le paramètre de localisation pour les instances positives est posé égal à 0,75 (en haut à gauche), à 1,50 (en haut à droite), à 3,00 (en bas à gauche) et à 5,00 (en bas à droite). La taille de l'échantillon est de 10 000. Les intervalles de confiance sont construits au niveau de confiance $\alpha = 10\%$. La proportion de couverture pour 1000 simulations en utilisant l'approche d'auto-amorçage exacte proposée (en noir), la méthode de lissage par noyau (en bleu) et la couverture cible de 95% (ligne pointillée) sont illustrées en fonction du ratio de faux positifs.

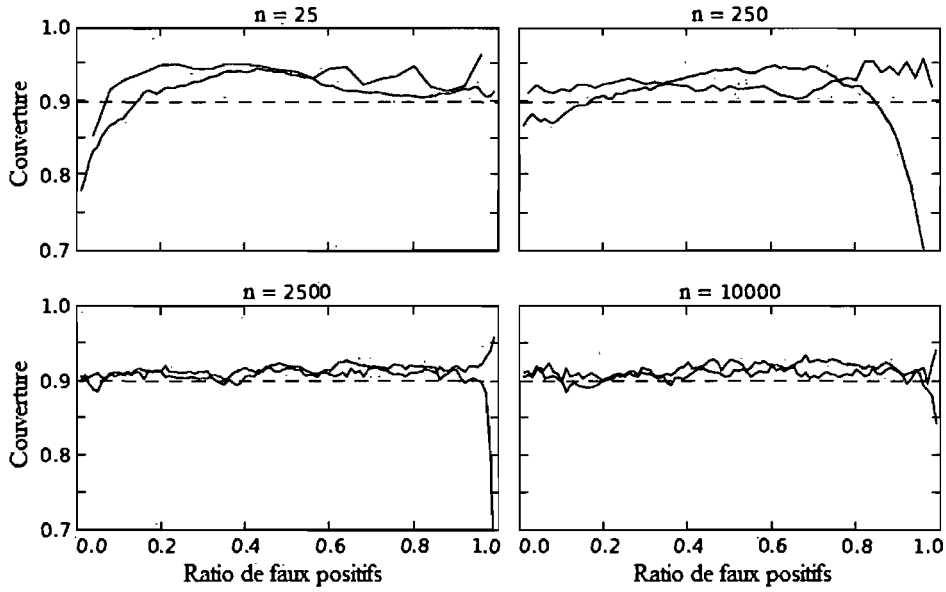

Fig. 4.8 Effet de la taille de l'échantillon sur la couverture. Les tailles considérées sont 25 (en haut à gauche), 250 (en haut à droite), 2500 (en bas à gauche) et 10 000 (en bas à droite). Les intervalles de confiance sont construits au niveau de confiance $\alpha = 10\%$. Le paramètre de localisation pour les instances positives est $\theta = 3{,}0$. La proportion de couverture pour 1000 simulations en utilisant l'approche d'auto-amorçage exacte proposée (en noir), la méthode de lissage pas noyau (en bleu) et la couverture cible de 95% (ligne pointillée) sont illustrées en fonction du ratio de faux positifs.

4.5.4 Moyennage vertical et jeux de données combinés

Cette section traitera des résultats sur la précision de couverture pour les intervalles de confiance obtenus à l'aide de l'algorithme 4.4 sur la différence entre deux courbes ROC, à un ratio de faux positifs fixé. L'expérience suit la même approche qu'à la section 4.5.2. Les résultats sont illustrés par la figure 4.9. La précision de la couverture est généralement trop conservatrice, particulièrement pour les valeurs de dispersion ($\theta = 3$) et de corrélation ($\rho = 0{,}9$) très grandes.

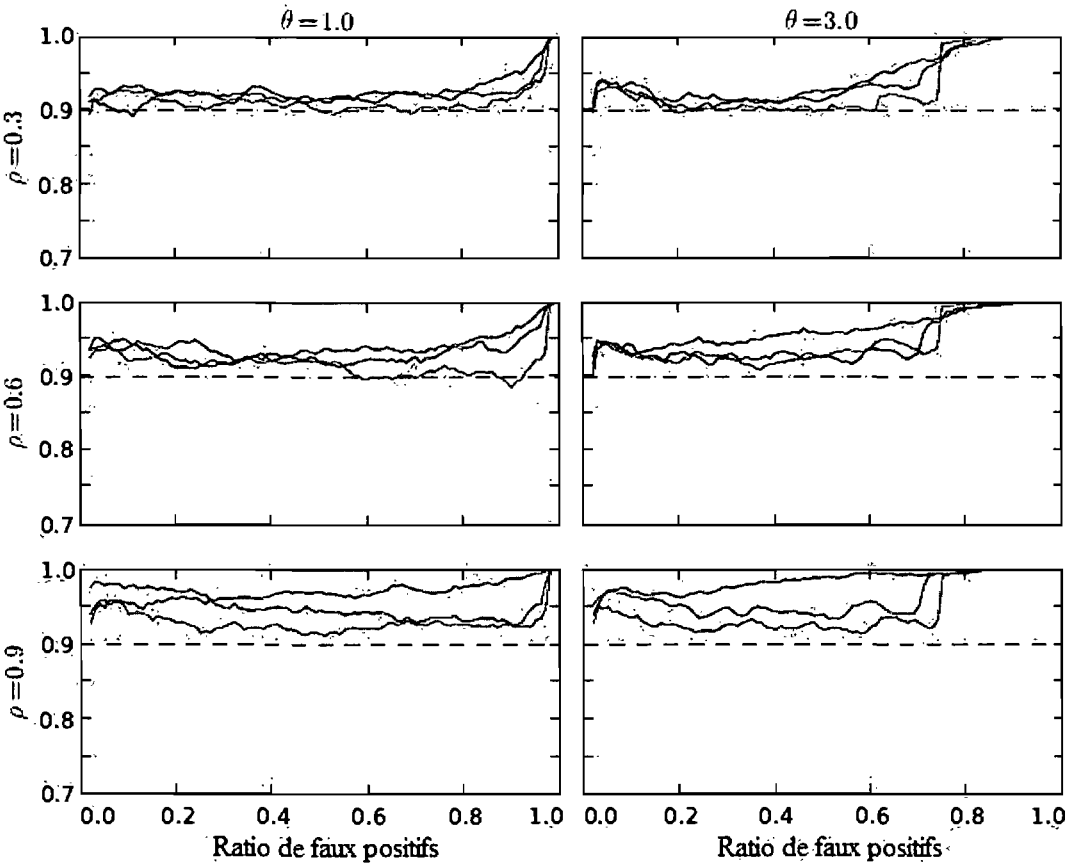

Fig. 4.9 Précision de couverture pour la différence entre les ratios de vrais positifs. La taille des échantillons est de 100 et le niveau de confiance est $\alpha = 10\%$. Le paramètre de localisation pour les instances positives du premier modèle est $\theta = 1{,}0$ (à gauche) et $\theta = 3{,}0$ (à droite). Le facteur de corrélation est $\rho = 0{,}3$ (en haut), $\rho = 0{,}6$ (au centre) et $\rho = 0{,}9$ (en bas). Le paramètre de localisation pour le score des instances positives du second modèle est θ (en bleu), $\theta + 2{,}00$ (en vert) et $\theta + 4{,}00$ (en rouge). La proportion de couverture pour 1000 simulations et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction du ratio de faux positifs.

4.6 Impact de l'utilisation de l'auto-amorçage stratifié

Tous les calculs et expériences que nous avons effectués sont sous l'hypothèse qu'on utilise l'auto-amorçage stratifié, c'est-à-dire que nos échantillons sont formés de deux échantillons indépendants, dont l'un est composé strictement d'instances positives et l'autre strictement de négatives. Il est intéressant de se demander si les intervalles de confiance que nous avons proposés sont robustes face à un changement dans les proportions de faux et de vrais positifs. En d'autres termes, est-ce que ces intervalles fonctionnent aussi bien si on utilise l'auto-amorçage classique, où on rééchantillonne parmi toutes les instances simultanément? Afin de répondre à cette question, nous avons reproduit l'expérience de taille de la section 4.5.1. Comme le montre la figure 4.10, il ne semble pas y avoir de différence significative entre les couvertures des deux approches.

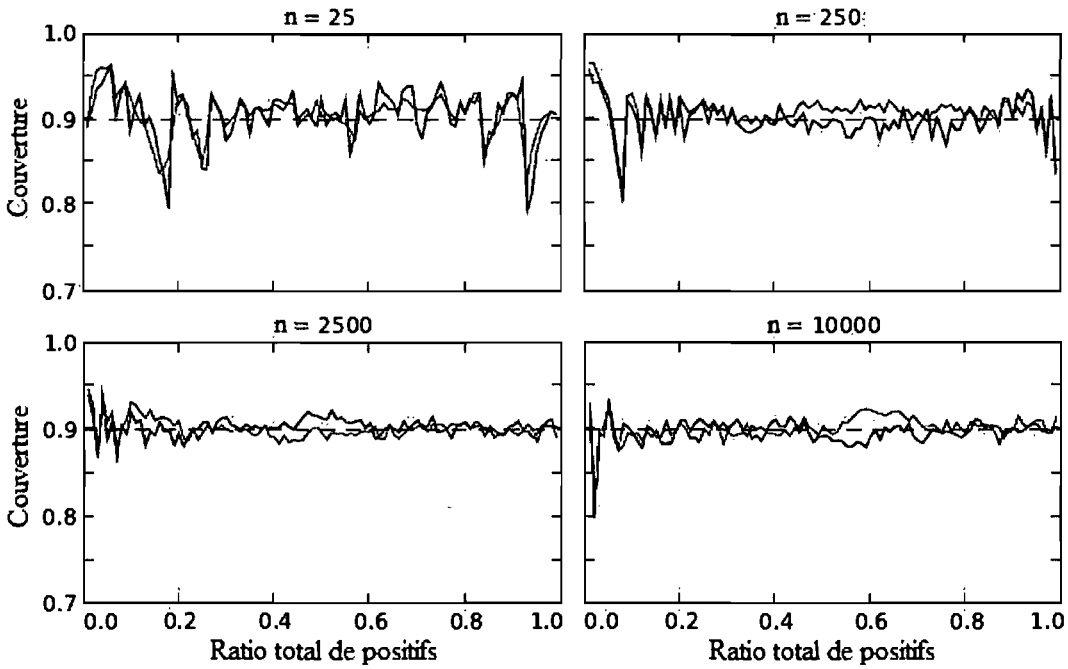

Fig. 4.10 Échantillonnage complet (vert) versus stratifié (bleu). Les deux approches sont comparées à l'aide de l'expérience de taille de la section 4.5.1.

CHAPITRE 5

Distribution d'auto-amorçage exacte ponctuelle des courbes de coûts

Le chapitre 3 de ce mémoire expliquait qu'une courbe de coût [Drummond & Holte, 2000; Drummond et Holte, 2006] est une représentation graphique du coût espéré d'un classifieur en fonction des conditions d'opération. Comme pour les courbes ROC, il est intéressant de pouvoir évaluer si un classifieur est *significativement* plus performant qu'un autre. Il est donc intéressant de pouvoir créer des intervalles de confiance pour la performance des classifieurs. Encore une fois, l'une des méthodes fréquemment utilisées pour y parvenir est le rééchantillonnage d'auto-amorçage [Efron & Tibshirani, 1993]. Nous apportons donc une avancée en traitant de la distribution d'auto-amorçage exacte obtenue lorsqu'on fait tendre le nombre d'échantillons d'auto-amorçage vers l'infini. Nous présenterons ici deux procédures d'auto-amorçage. Nous traiterons tout d'abord le rééchantillonnage d'auto-amorçage stratifié, tel que défini au chapitre 4, où les proportions d'instances positives et négatives sont maintenues d'un échantillon à l'autre. Ensuite, nous utiliserons l'approche d'auto-amorçage complet, selon lequel les échantillons sont tirés complètement aléatoirement du jeu de données initial.

Les différences et similitudes entre ces deux approches ont été présentées au chapitre 4. Plus particulièrement, la section 4.6 démontrait que la couverture des intervalles de confiance des courbes ROC était similaire sous les deux hypothèses. Rappelons que l'utilité principale de l'auto-amorçage stratifié est d'éviter le problème de division par zéro qui se produit lorsque les instances d'un échantillon ne proviennent que d'une seule des deux classes. L'auto-amorçage complet, quant à lui, permet de varier les

proportions entre les instances. Du point de vue de l'usager, les deux méthodes offrent donc des informations différentes et vont plus loin qu'une simple dérivation mathématique. Pour ces raisons, nous traiterons des deux méthodes.

5.1 Rééchantillonnage d'auto-amorçage stratifié

En utilisant les mêmes définitions et les mêmes notations qu'au chapitre 3 et 4, on obtient la formule suivante :

$$C_t^T = (1 - TP_t^+)\, p(+)C(-|+) + FP_t^- \, p(-)C(+|-), \tag{5.1}$$

où C_t^T est le coût total lié à un classifieur sous l'auto-amorçage stratifié. Il faut noter que, dans ce cas, TP_t^+ et FP_t^- sont indépendants. Les conditions d'opération sont définies comme au chapitre 3:

$$w = \frac{p(+)C(-|+)}{p(+)C(-|+) + p(-)C(+|-)}. \tag{5.2}$$

Ainsi, on obtient le coût normalisé :

$$C_t^N = w(1 - TP_t^+) + (1 - w)FP_t^-, \tag{5.3}$$

où $w \in [0;1]$. L'espérance et la variance du coût normalisé sont donc :

$$E[C_t^N] = w(1 - p_t^+) + (1 - w)\, p_t^-, \tag{5.4}$$

$$Var[C_t^N] = w^2 \, p_t^+ (1 - p_t^+) + (1 - w)^2 \, p_t^- (1 - p_t^-). \tag{5.5}$$

À l'aide de ces moments, nous pouvons utiliser l'approximation gaussienne de la distribution de C_t^N pour obtenir des intervalles de confiance. Au niveau du temps de calcul, le tout est dominé par la mise en ordre des instances en fonction de leur score et sera donc de l'ordre $O(n \cdot \ln(n))$.

Afin de déterminer si la différence entre les performances de deux classifieurs est statistiquement significative, il faut obtenir la distribution de la différence de leurs coûts normalisés :

$$\Delta_{1,2}C^N_{t_1,t_2} = {}_2C^N_{t_2} - {}_1C^N_{t_1}$$

$$= w({}_1TP^+_{t_1} - {}_2TP^+_{t_2}) + (1-w)({}_2FP^-_{t_2} - {}_1FP^-_{t_1}). \quad (5.6)$$

Comme pour les courbes ROC, ${}_1C^N_{t_1}$ et ${}_2C^N_{t_2}$ ne sont pas indépendants puisque les scores assignés par deux classifieurs différents peuvent être corrélés : par exemple, les transactions frauduleuses évidentes obtiendront des scores élevés avec tous les classifieurs. Aussi, une erreur commise par les deux classifieurs s'annulera lors du calcul de la différence des coûts. Posons ${}_1n^+_{t_1}$, le nombre d'instances positives classifiées correctement par le premier classifieur et mal classifiées par le second. Inversement, posons ${}_2n^+_{t_2}$, le nombre d'instances positives correctement classifiées par le 2e classifieur et mal classifiées par le premier. Il faut noter que les seuils t_1 et t_2 associés aux conditions d'opération w peuvent changer d'un classifieur à l'autre puisque la distribution et l'échelle des scores peuvent varier. Nous définissons ${}_1n^-_{t_1}$ et ${}_2n^-_{t_2}$ identiquement pour les instances négatives.

Soit ${}_1N^+_{t_1}$, ${}_2N^+_{t_2}$, ${}_1N^-_{t_1}$ les variables aléatoires pour les nombres d'instances correspondantes dans l'échantillon d'auto-amorçage stratifié. Les valeurs ${}_1N^+_{t_1}$ et ${}_2N^+_{t_2}$ suivent conjointement une distribution multinomiale. Il en est de même pour leurs homologues négatifs. En conséquence, on obtient les deux premiers moments suivants pour $\Delta_{1,2}C^N_{t_1,t_2}$:

$$E[\Delta_{1,2}C^N_{t_1,t_2}] = w\left({}_1p^+_{t_1} - {}_2p^+_{t_2}\right) + (1-w)\left({}_2p^-_{t_2} - {}_1p^-_{t_1}\right), \tag{5.7}$$

$$Var[\Delta_{1,2}C^N_{t_1,t_2}] = \frac{w^2}{n^+}\left[{}_1p^+_{t_1} + {}_2p^+_{t_2} - \left({}_1p^+_{t_1} - {}_2p^+_{t_2}\right)^2\right] + \frac{(1-w)^2}{n^-}\left[{}_1p^-_{t_1} + {}_2p^-_{t_2} - \left({}_1p^-_{t_1} - {}_2p^-_{t_2}\right)^2\right], \tag{5.8}$$

où ${}_1p^+_{t_1} = {}_1n^+_{t_1}/n^+$, ${}_2p^+_{t_2} = {}_2n^+_{t_2}/n^+$, ${}_1p^-_{t_1} = {}_1n^-_{t_1}/n^-$ et ${}_2p^-_{t_2} = {}_2n^-_{t_2}/n^-$. Encore une fois, le temps de calcul total est de l'ordre $O(n \cdot \ln(n))$.

5.2 Rééchantillonnage d'auto-amorçage complet

Sous l'auto-amorçage complet, les proportions d'instances positives et négatives peuvent varier d'un échantillon à l'autre. En conséquence, nous utilisons les lettres majuscules N^+ et N^- pour représenter ces valeurs qui suivent des distributions binomiales : $N^+ \sim \text{Bin}(n, p^+)$ et $N^- \sim \text{Bin}(n, p^-)$. Il en découle que, sous l'auto-amorçage complet, $P^+ = N^+/n$ et $P^- = N^-/n$ et suivent donc des binomiales. Ces distributions pourraient toutefois être facilement substituées si on avait des raisons de croire que d'autres distributions seraient plus appropriées dans la pratique.

L'équation (5.1) est toujours valide sous l'auto-amorçage complet à l'exception que P^+ et P^- sont maintenant traitées comme des variables aléatoires. Puisque ces valeurs ne sont plus fixes, la normalisation de cette équation se fera en divisant par la moyenne pondérée la plus large possible. Cette dernière sera simplement le coût maximum des deux erreurs de classification : $C_{max} = \max[C(-|+), C(+|-)]$. Le premier cas s'obtiendra lorsque $N^+ = n$ et $TP_t^+ = 0$. De manière similaire, on aura le second cas lorsque $N^- = n$ et $FP_t^- = 1$. L'équation (5.3) devient donc :

$$C^N_t = \frac{N^+C(-|+)(1-TP^+_t) + N^-C(+|-)FP^-_t}{nC_{\max}}. \tag{5.9}$$

L'espérance et la variance de C_t^N peuvent donc être obtenues à l'aide d'espérances itérées :

$$E[C_t^N] = E_{N^+}\{E[C_t^N \mid N^+]\}$$

$$= \frac{C(-|+)(n^+ - n_t^+) + C(+|-)n_t^-}{nC_{\max}}, \tag{5.10}$$

$$Var[C_t^N] = Var_{N^+}\{E[C_t^N \mid N^+]\} + E_{N^+}\{Var[C_t^N \mid N^+]\}$$

$$= \frac{C(-|+)^2\alpha_t^+ + C(+|-)^2\alpha_t^- + \delta_t^2}{(nC_{\max})^2} \tag{5.11}$$

où

$$\alpha_t^+ = n_t^+ - \frac{(n_t^+)^2}{n^+},$$

$$\alpha_t^- = n_t^- - \frac{(n_t^-)^2}{n^-} \text{ et}$$

$$\delta_t^2 = \left(C(-|+)\frac{n^+ - n_t^+}{n^+} - C(+|-)\frac{n_t^-}{n^-}\right)^2 \frac{n^+ n^-}{n}.$$

Encore une fois, nous obtenons les intervalles de confiance à l'aide d'une distribution gaussienne ayant comme deux premiers moments ceux trouvés précédemment. Le tout se fait dans un temps de l'ordre $O(n \cdot \ln(n))$.

Pour la différence entre la performance de deux classifieurs sous l'auto-amorçage complet, l'équation (5.6) devient :

$$\Delta_{1,2}C_{t_1,t_2}^N = {}_2C_{t_2}^N - {}_1C_{t_1}^N$$

$$= \frac{C(-|+)({}_1N_{t_1}^+ - {}_2N_{t_2}^+) + C(+|-)({}_2N_{t_2}^- - {}_1N_{t_1}^-)}{nC_{\max}}. \tag{5.12}$$

On obtient l'espérance et la variance de cette différence :

$$E[\Delta_{1,2} C^N_{t_1,t_2}] = E_{N^+}\{E[\Delta_{1,2} C^N_{t_1,t_2} \mid N^+]\}$$

$$= \frac{C(-|+)({}_1n^+_{t_1} - {}_2n^+_{t_2}) + C(+|-)({}_2n^-_{t_2} - {}_1n^-_{t_1})}{nC_{\max}}, \quad (5.13)$$

$$Var[\Delta_{1,2} C^N_{t_1,t_2}] = Var_{N^+}\{E[\Delta_{1,2} C^N_{t_1,t_2} \mid N^+]\} + E_{N^+}\{Var[\Delta_{1,2} C^N_{t_1,t_2} \mid N^+]\}$$

$$= \frac{C(-|+)^2 \alpha^+_{t_1,t_2} + C(+|-)^2 \alpha^-_{t_1,t_2} + \delta^2_{t_1,t_2}}{(nC_{\max})^2} \quad (5.14)$$

où

$$\alpha^+_{t_1,t_2} = {}_1n^+_{t_1} + {}_2n^+_{t_2} - \frac{({}_1n^+_{t_1} - {}_2n^+_{t_2})^2}{n^+},$$

$$\alpha^-_{t_1,t_2} = {}_1n^-_{t_1} + {}_2n^-_{t_2} - \frac{({}_1n^-_{t_1} - {}_2n^-_{t_2})^2}{n^-} \text{ et}$$

$$\delta^2_{t_1,t_2} = \left(C(-|+)\frac{{}_1n^+_{t_1} - {}_2n^+_{t_2}}{n^+} - C(+|-)\frac{{}_2n^-_{t_2} - {}_1n^-_{t_1}}{n^-} \right)^2 \frac{n^+ n^-}{n}.$$

Encore une fois, le temps global de calcul est de l'ordre $O(n \ln(n))$.

5.3 Simulations numériques

Cette section présente les résultats de quelques expériences que nous avons menées afin de vérifier la performance des intervalles de confiance décrits dans ce chapitre. La première expérience est inspirée de celle qu'utilise Macskassy et al. (2005) pour comparer les intervalles de confiance de courbes ROC. Les scores des instances positives et négatives suivent des distributions normales dans lesquelles les paramètres varient. Le paramètre d'échelle est fixé à 3,00 pour les deux types d'instances, mais le paramètre de localisation θ des instances positives varie dans l'ensemble {0,75; 1,5; 3,0; 5,0}, alors que celui des instances négatives est posé comme valant $-\theta$. La taille des

échantillons est fixée à 1000. Le nombre de simulations pour l'auto-amorçage stratifié est aussi fixé à 1000 pour chaque valeur de θ. Nous nommerons cette expérience : l'expérience de dispersion. Pour le calcul des intervalles de confiance, nous utiliserons un niveau de confiance $\alpha = 10\%$. La figure 5.1 présente les résultats. On remarque que les résultats sont meilleurs lorsque les distributions se superposent peu, c'est-à-dire pour de grandes valeurs de θ. On note une certaine tendance à être trop conservateur lorsque *w* est proche des extrêmes. La section 5.5 traitera de ce phénomène récurrent.

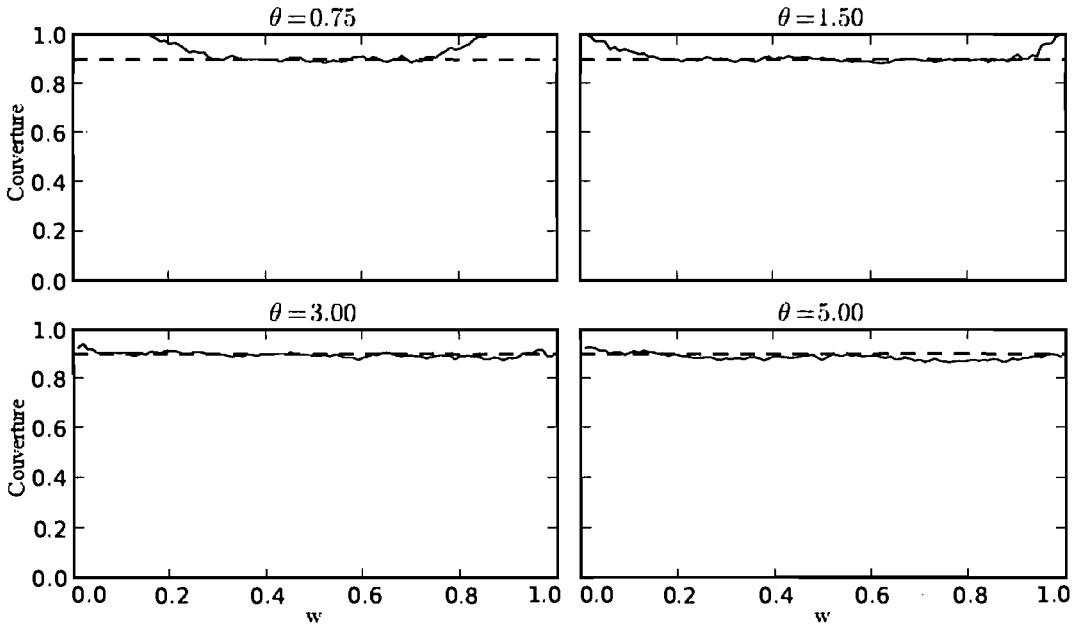

Fig. 5.1 Effet de la dispersion entre les distribution sur la couverture. L'échantillonnage d'auto-amorçage stratifié est utilisé. Les intervalles de confiance sont trouvés pour le coût d'un classifieur. Le paramètre de dispersion des instances positives varie entre les valeurs 0,75 (en haut à gauche), 1,50 (en haut à droite), 3,00 (en bas à gauche) et 5,00 (en bas à droite). La taille des échantillons est de 1000. Le niveau de confiance est de 10%. La proportion de couverture pour 1000 simulations (ligne pleine) et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction des conditions d'opération.

On notera aussi que l'amélioration des résultats lorsque la dispersion augmente est contraire aux résultats qu'on obtenait pour les courbes ROC. Nous avons vu au chapitre précédent que la courbe ROC évaluait un plus grand intervalle de seuils lorsque la dispersion augmentait. Cela était dû au fait que la courbe ROC nécessite des seuils qui sont tels que les ratios de faux et de vrais positifs varient de 0 à 1. Donc, plus la

dispersion est grande, plus l'éventail des seuils devra être grand afin d'obtenir la courbe ROC. À l'inverse, une augmentation de la dispersion fera rétrécir l'intervalle des seuils considérés pour les courbes de coûts. Nous avons vérifié numériquement ce fait, mais un exemple tout simple nous permettra de se convaincre de ce phénomène. Supposons que la dispersion soit totalement nulle, c'est-à-dire que les distributions des instances positives et négatives sont totalement centrées au même point. Dans ce cas, le seuil obtenant le moindre coût dépendra grandement des conditions d'opérations et génèrera un grand intervalle. À l'inverse, si la dispersion est assez grande pour que les distributions soient totalement disjointes, le seuil se trouvant entre les deux distributions serait un classifieur parfait (il classifie tout les négatifs et les positifs dans leur classe respective) et, en conséquence, serait le seuil optimal peu importe les conditions d'opérations. Ce cas-ci aurait donc un intervalle de seuil de mesure nulle. La taille de l'intervalle des seuils considérés varie donc inversement entre les courbes ROC et les courbes de coûts lorsqu'on change la dispersion. De plus, lorsque l'intervalle des seuils considérés augmente, la probabilité d'obtenir un seuil qui se trouve dans les ailes externes des distributions augmente. Puisque c'est lorsque nous observons ces seuils extrêmes que la théorie des courbes ROC et des courbes de coûts devient instable, il en découle que la précision de couverture est affectée par la dispersion de manière opposée entre les courbes ROC et les courbes de coûts.

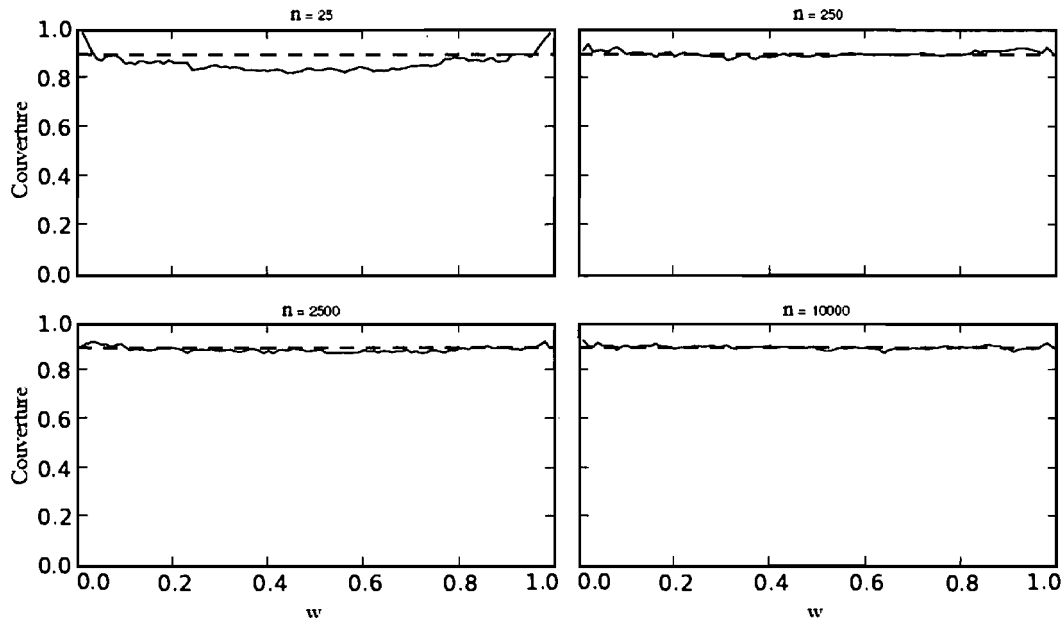

Fig. 5.2 Effet de la taille des échantillons sur la couverture. Le rééchantillonnage d'auto-amorçage stratifié est utilisé. Les intervalles de confiance sont trouvés pour le coût d'un classifieur. La taille des échantillons varient entre les valeurs 25 (en haut à gauche), 250 (en haut à droite), 2500 (en bas à gauche) et 10 000 (en bas à droite). Le niveau de confiance est de 10%. Le paramètre de localisation des instances positives est de 3,0. La proportion de couverture (ligne pleine) pour 1000 simulations et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction des conditions d'opération.

La seconde expérience, dite l'expérience de taille, considère l'effet de la taille des échantillons sur la précision de la couverture. Elle est très similaire à l'expérience de dispersion à deux exceptions près. Tout d'abord, le paramètre de localisation ne varie plus et est fixé à $\theta = 3,0$. Ensuite, la taille des échantillons n'est plus fixée à 1000 mais varie dans l'ensemble {25, 250, 2500, 10 000}. La figure 5.2 présente les résultats. On remarque qu'au fur et à mesure où la taille des échantillons augmente, la précision de couverture s'améliore. Cependant, il vaut la peine de souligner que, même pour des échantillons aussi petit que 25, la couverture est relativement bonne.

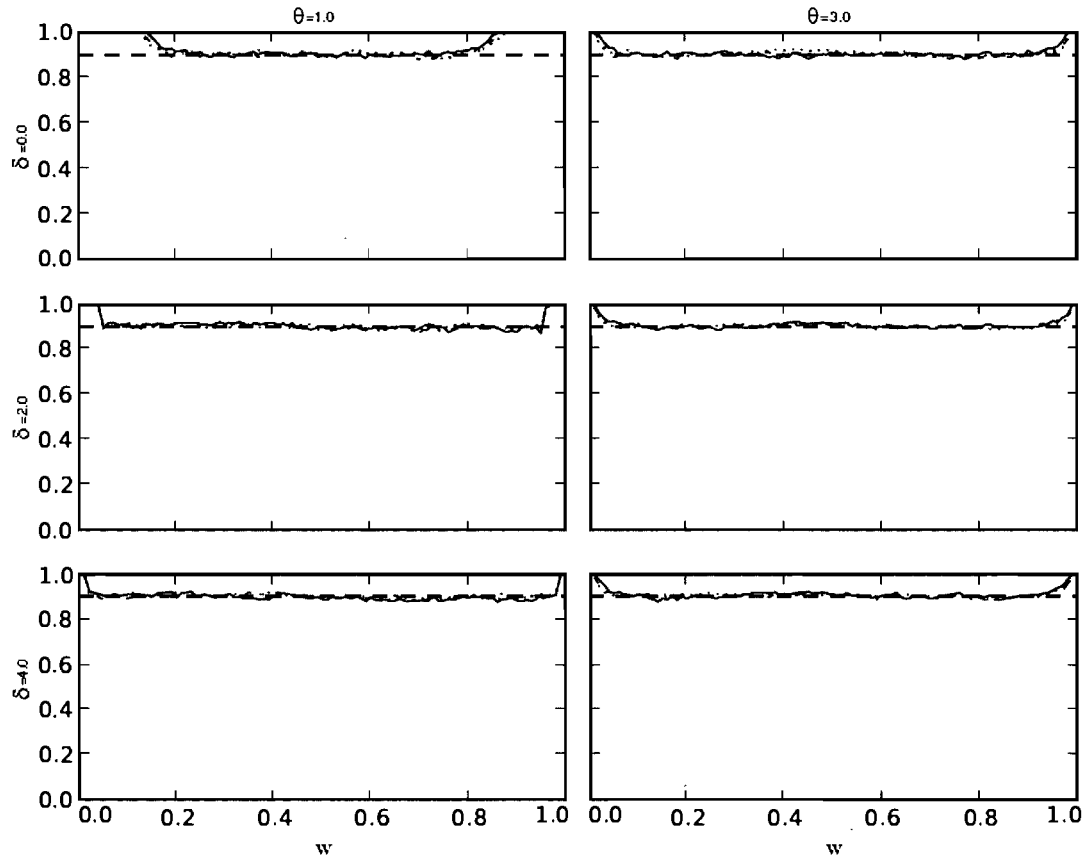

Fig. 5.3 Précision de couverture des intervalles de confiance pour la différence de performance entre deux classifieurs. L'auto-amorçage stratifié est utilisé. La taille des échantillons est de 1000. Le niveau de confiance est de 10%. Le paramètre de localisation pour les instances positives du premier classifieur est $\theta = 1$ (à gauche) et $\theta = 3$ (à droite). Le paramètre de localisation pour les instances positives du second classifieur est θ (en haut), $\theta + 2$ (au centre) et $\theta + 4$ (en bas). Sur chaque graphique, le coefficient de corrélation est 0,3 (ligne pointillée), 0,6 (ligne semi-pointillée) et 0,9 (ligne pleine). Les proportions de couverture pour 1000 simulations et la couverture cible de 90% (trait horizontal) sont illustrées en fonction des conditions d'opération.

Notre troisième expérience, dite l'expérience des différences, traite des intervalles de confiance pour la différence de performance entre deux classifieurs. La structure de l'expérience est semblable à celle des deux précédentes. Les scores sont distribués selon des distributions binormales avec 3,0 comme paramètre d'échelle. Les intervalles de confiance sont obtenus au niveau de confiance $\alpha = 10\%$. Les paramètres de localisation sont choisis comme suit : pour les instances positives du premier classifieur, nous considérons deux valeurs : $\theta \in \{1{,}0 ; 3{,}0\}$. Pour les instances négatives de deux classifieurs, le paramètre vaut $-\theta$. Quant au paramètre pour les instances

positives du second classifieur, il prendra trois valeurs : θ, $\theta + 2$ et $\theta + 4$. La différence entre les paramètres de localisation des deux classifieurs (soit 0, 2 ou 4) se nomme, comme au chapitre précédent, le paramètre de changement δ. On inclus aussi une forme de dépendance entre les scores de chaque classifieur. Les facteurs de corrélation considérés sont : $\rho \in \{0{,}3; 0{,}6; 0{,}9\}$.

Les résultats sont présentés à la figure 5.3. Comme pour toutes les expériences, on remarque une tendance conservatrice pour les valeurs extrêmes de *w*. En regardant la figure par colonne, on remarque que le paramètre de dispersion θ influence la précision. Les plus grandes valeurs de θ améliore la couverture. On note aussi que le paramètre de changement améliore la précision au fur et à mesure qu'il augmente. De plus, le coefficient de corrélation n'a presque pas d'impact sur la précision de couverture. Il s'agit d'une propriété très intéressante puisqu'elle implique que la performance des intervalles de confiance est indépendante de la corrélation entre les scores.

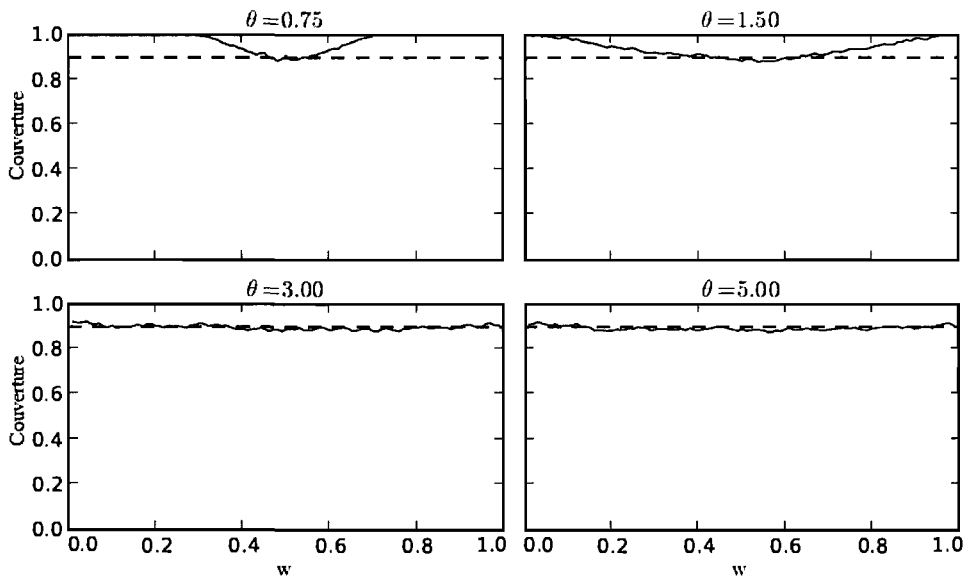

Fig. 5.4 Effet de la dispersion entre les distributions sur la couverture. Le rééchantillonnage d'auto-amorçage complet est utilisé. Les intervalles de confiance sont trouvés pour le coût d'un classifieur. Le paramètre de localisation pour les instances positives varie entre les valeurs 0,75 (en haut à gauche), 1,50 (en haut à droite), 3,00 (en bas à gauche) et 5,00 (en bas à droite). La taille des échantillons est de 1000. Le niveau de confiance est de 10%. La proportion de couverture pour 1000 simulations (ligne pleine) et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction des conditions d'opération.

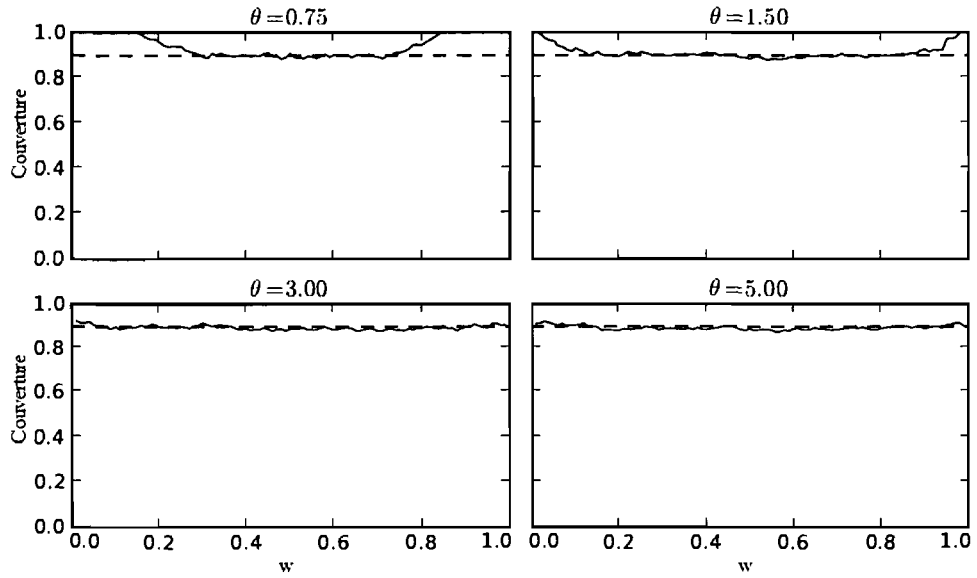

Fig. 5.5 Effet de la dispersion entre les distributions sur la couverture. Le rééchantillonnage d'auto-amorçage complet est utilisé. Les intervalles de confiance sont trouvés pour le coût d'un classifieur en utilisant les formules sous l'hypothèse d'auto-amorçage stratifié. Le paramètre de localisation pour les instances positives varie entre les valeurs 0,75 (en haut à gauche), 1,50 (en haut à droite), 3,00 (en bas à gauche) et 5,00 (en bas à droite). La taille des échantillons est de 1000. Le niveau de confiance est de 10%. La proportion de couverture pour 1000 simulations (ligne pleine) et la couverture cible de 90% (ligne pointillée) sont illustrées en fonction des conditions d'opération.

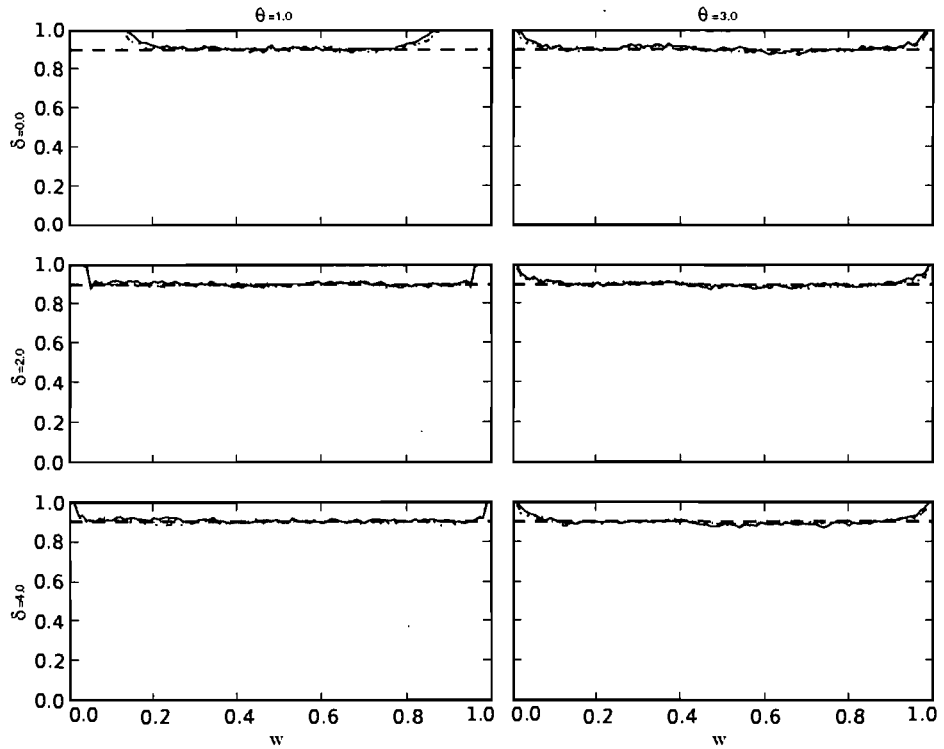

Fig. 5.6 Précision de couverture des intervalles de confiance pour la performance entre deux classifieurs. Le rééchantillonnage d'auto-amorçage complet est utilisé. Les intervalles de confiance sont calculés avec les formules sous l'hypothèse d'auto-amorçage stratifié. La taille des échantillons est de 1000. Le niveau de confiance est de 10%. Le paramètre de localisation des instances positives du premier classifieur est $\theta = 1$ (à gauche) et $\theta = 3$ (à droite). Le paramètre de localisation des instances positives du second classifieur est θ (en haut), $\theta + 2$ (au centre) et $\theta + 4$ (en bas). Pour chaque graphique, le coefficient de corrélation est 0,3 (ligne pointillée), 0,6 (ligne semi-pointillée) et 0,9 (ligne pleine). Les proportions de couverture pour 1000 simulations et la couverture cible de 90% (trait horizontal) sont illustrées en fonction des conditions d'opération.

Les figures 5.4 et 5.5 reprennent l'expérience de dispersion vue précédemment, mais dans le cas de rééchantillonnage d'auto-amorçage complet. La figure 5.4 a été générée en utilisant les formules théoriques d'auto-amorçage complet vues à la section 5.2. On remarque que la précision de couverture aux extrémités est trop conservatrice, surtout pour les petites valeurs de θ. La figure 5.5 représente la même expérience que la figure 5.4, mais cette fois on calcule les intervalles de confiance à l'aide des formules provenant de l'auto-amorçage stratifié, section 5.1, malgré le fait que le rééchantillonnage soit fait selon l'auto-amorçage complet. Bien que le phénomène des extrémités conservatrices soit toujours présent, son effet est moindre. On a donc une meilleure précision à l'aide des formules de la section 5.1. La figure 5.6 reprend, quant à elle, l'expérience des différences dans le cas de l'auto-amorçage complet. Tout comme pour l'expérience de dispersion, les formules de la section 5.2 (dont le graphique n'est pas présenté ici) génèrent des intervalles de confiance plus conservateur que celles de la section 5.1.

Nous n'avons pas reproduit l'expérience de taille puisque, dans le cas de l'auto-amorçage complet, un échantillon de petite taille peut prendre un nombre d'instances positives ou négatives beaucoup trop faible. Cette expérience devient donc insignifiante pour la performance des intervalles de confiance.

En regardant les figures 5.5 et 5.6, on remarque que, sous les équations d'auto-amorçage stratifié, les deux méthodes de rééchantillonnage performe similairement.

5.4 Approximation de Wald ajustée

Dans ce chapitre, les intervalles de confiance ont été calculés à l'aide d'une approximation gaussienne selon l'approche d'Agresti et Coull (1998) et d'Agresti et Min (2005). Il s'agit d'une approche de Wald ajustée.

Tableau 5.1 : Ajustement d'Agresti pour un jeu de données unique

	Instances positives	Instances négatives
Score ≥ t	2	2
Score < t	2	2

Dans un scénario à jeu de données unique, la méthode d'Agresti consiste à ajouter 2 instances à chacun des éléments de la matrice de confusion, voir le tableau 5.1. La méthode ajoute donc 8 instances à l'ensemble des données. L'intervalle de confiance est ensuite calculé selon la méthode de Wald classique. Cependant, les instances supplémentaires garantissent une variance non nulle même pour les cas extrêmes. [Agresti et Coull, 1998].

Dans le cas où on a des jeux de données combinés, Agresti ajuste, comme le montre le tableau 5.2, en ajoutant ½ instances pour chaque possibilité [Agresti et Min, 2005]. Puisque notre étude considérait des jeux de données combinés lorsque nous comparions deux classifieurs, il y a 8 possibilités. Encore une fois, lorsque ces instances sont ajoutées, l'intervalle de confiance est calculé comme sous l'hypothèse de Wald classique.

Tableau 5.2 : Ajustement d'Agresti pour des jeux de données combinés

	Instances positives	Instances négatives
Score $\geq t_1$ Score $\geq t_2$	$\frac{1}{2}$	$\frac{1}{2}$
Score $\geq t_1$ Score $< t_2$	$\frac{1}{2}$	$\frac{1}{2}$
Score $< t_1$ Score $\geq t_2$	$\frac{1}{2}$	$\frac{1}{2}$
Score $< t_1$ Score $< t_2$	$\frac{1}{2}$	$\frac{1}{2}$

5.5 Discussion

Nous nous rappelons que, comme vu au chapitre 4, l'approche de Wald classique provoquait, pour des raisons de variance trop faible ou nulle, des bris au niveau de la couverture dans les extrémités de la courbe ROC. Bien que non représenté graphiquement dans ce mémoire, nous avons vérifié que le même phénomène se produit pour les courbes de coûts lorsque les conditions d'opération prennent des valeurs près de 0 ou 1. Bien que l'approche d'Agresti ait parfois des niveaux de précision trop conservateurs à ces mêmes extrêmes, elle performe, malgré tout, significativement mieux que l'approche de Wald classique. Il existe d'autres alternatives, mais les solutions d'Agresti génèrent habituellement de bons résultats et restent les plus simples à ce jour.

Il est intéressant de relever que, lorsque les conditions d'opération atteignent des valeurs extrêmes, les valeurs de l'espérance de coût (et de la différence par la même occasion) chutent à zéro. L'estimation qu'on en fait est alors inefficace (si la valeur 0 n'est pas incluse) ou parfaite (si 0 est contenue dans notre intervalle). C'est ce phénomène qui provoque les brisures dans le cas de Wald et l'effet conservateur de l'approche d'Agresti.

Un autre point d'intérêt est le choix du nombre d'instances qu'ajoute la méthode d'Agresti. Les valeurs 2 (jeu de données unique) et ½ (jeux de données combinés) sont choisies arbitrairement par Agresti. L'idée est que si ce nombre est trop faible, l'impact sera trop petit et l'ajustement ne règlera pas les problèmes de variance que génère l'approche de Wald. Si, à l'opposé, ces valeurs sont trop grandes, le biais ajouté rajoutera trop d'erreur et la précision de couverture en sera aussi amoindrie. Nous avons donc fait une dernière expérience, dite l'expérience d'ajustement. Cette dernière reprend les expériences de dispersion et de différences, mais en y faisant varier la quantité d'instances ajoutées. Puisque le rééchantillonnage d'auto-amorçage complet et celui stratifié réagissent de la même façon, nous ne présentons ici que les graphiques sous l'hypothèse d'auto-amorçage stratifié. La figure 5.7 reprend donc l'expérience de dispersion de la figure 5.1. Les figures 5.8 et 5.9 reprennent la même expérience mais en modifiant, à la baisse et à la hausse respectivement, le nombre d'instances ajoutées. Sur la figure 5.8, on se rapproche de l'expérience de Wald et on peut voir la brisure (surtout pour un θ petit) lorsque les conditions d'opérations sont faibles. L'inégalité des variances pour les distribution des instances positives et des instances négatives de notre jeu de données initiales provoque l'asymétrie entre le côté droit et le côté gauche du graphique. La brisure n'est pas visible, mais elle le deviendrait si l'ajustement était assez petit. La figure 5.9, quant à elle, présente clairement des cassures des deux côtés lorsque l'ajustement est trop grand.

De la même manière, la figure 5.10 reprend l'expérience des différences de la figure 5.3. Les figures 5.11 et 5.12 reprennent cette expérience avec un ajustement à la baisse et à la hausse respectivement. Dans le premier cas, on remarque des brisures des

deux côtés. Dans le second, la réaction est moins flagrante mais les courbes sont un peu plus conservatrices.

Nous concluons en soulignant que, bien que l'impact soit surtout visible pour de petites valeurs de θ, les valeurs suggérées par Agresti (soit 2 et ½) semblent générer de meilleurs résultats.

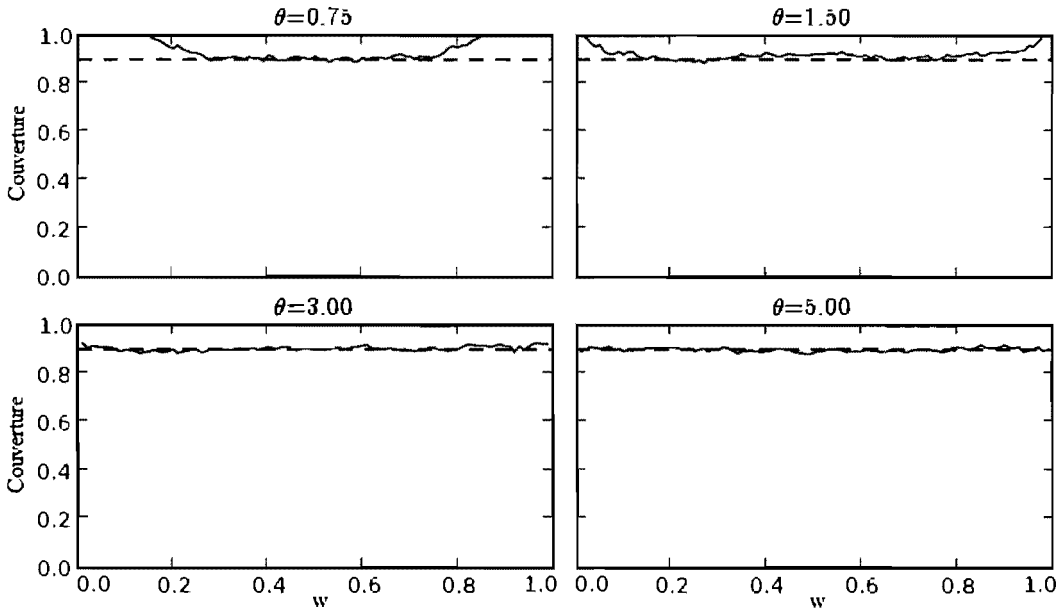

Fig. 5.7 Effet de la dispersion sur la précision de couverture. L'auto-amorçage stratifié est utilisé. L'ajustement d'Agresti ajoute 2 instances à chaque possibilité.

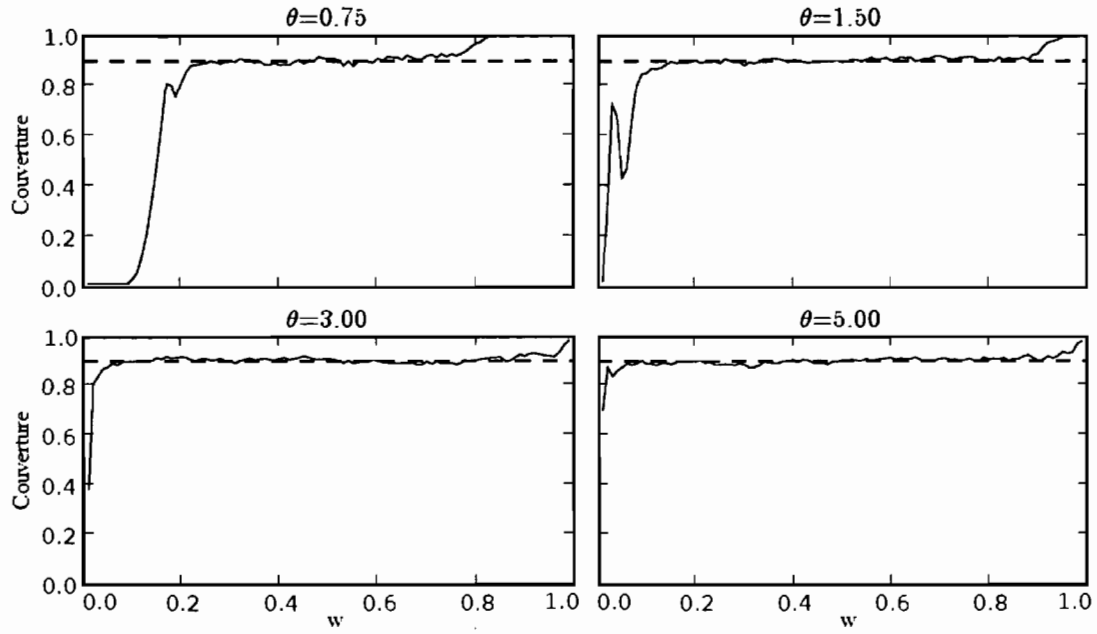

Fig. 5.8 Effet de la dispersion sur la précision de couverture. L'auto-amorçage stratifié est utilisé. L'ajustement d'Agresti ajoute 0.5 instance à chaque possibilité.

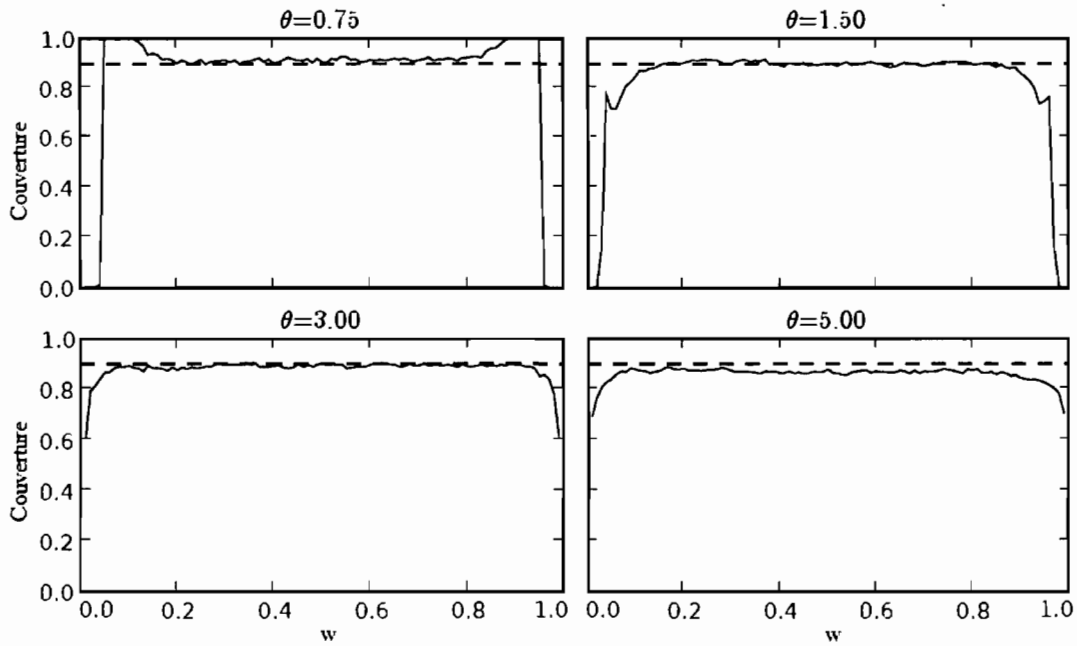

Fig. 5.9 Effet de la dispersion sur la précision de couverture. L'auto-amorçage stratifié est utilisé. L'ajustement d'Agresti ajoute 3 instances à chaque possibilité.

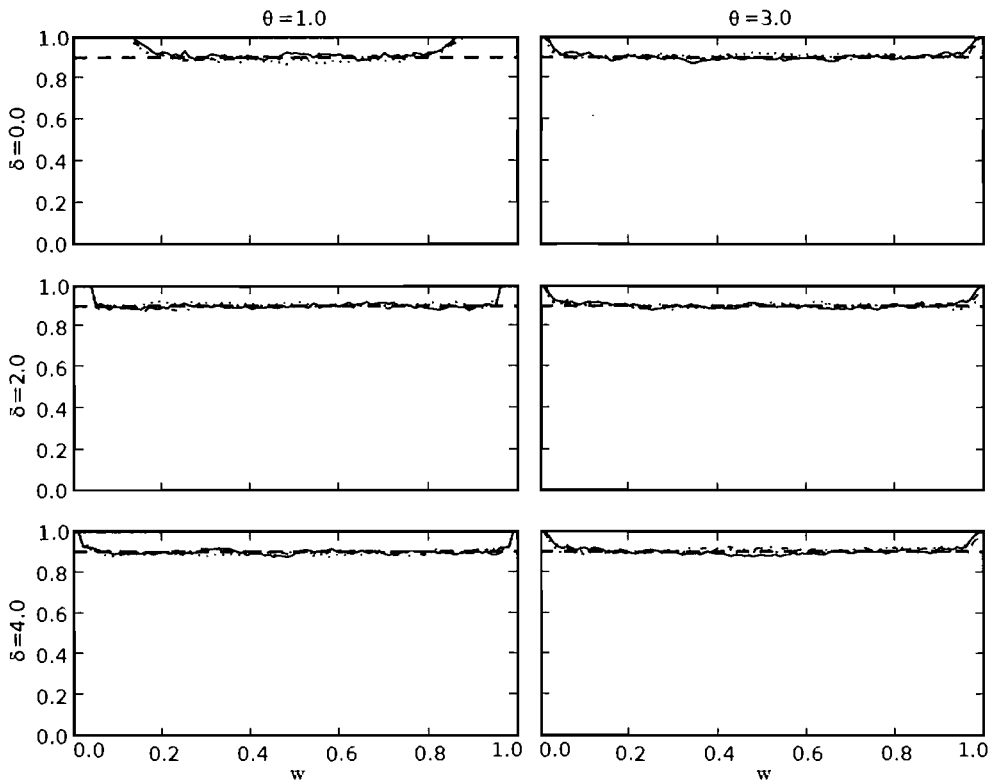

Fig. 5.10 Reprise de l'expérience des différences sous rééchantillonnage d'auto-amorçage stratifié. L'ajustement d'Agresti ajoute 1/2 instance à chaque possibilité.

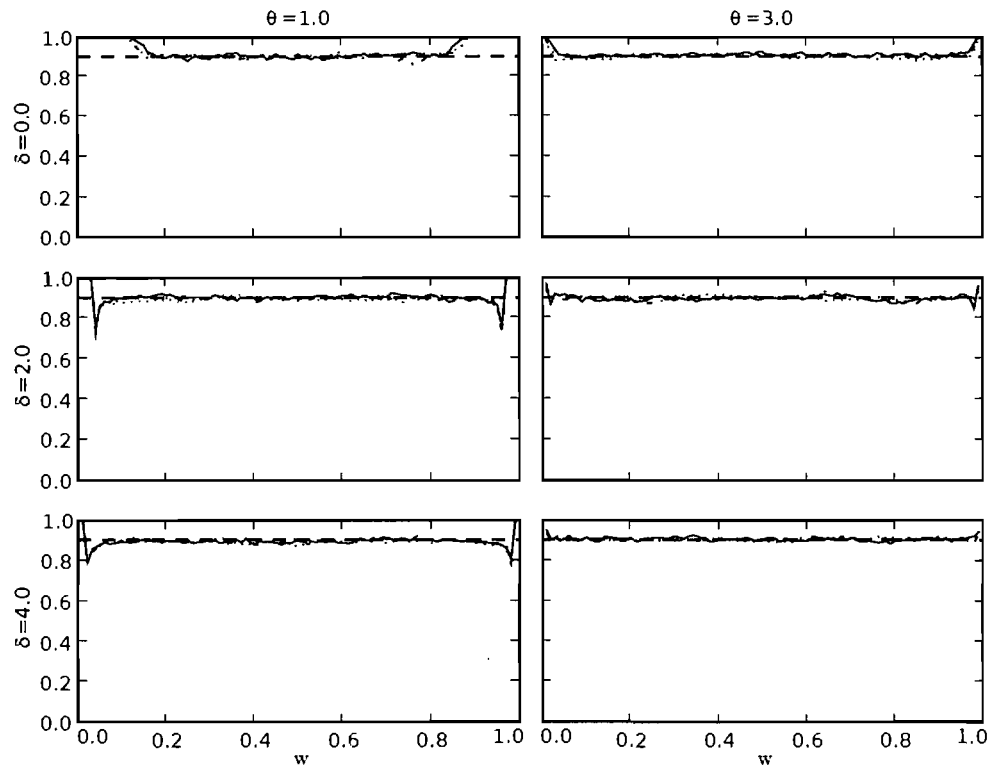

Fig. 5.11 Reprise de l'expérience des différences sous rééchantillonnage d'auto-amorçage stratifié. L'ajustement d'Agresti ajoute 1/8 instance à chaque possibilité.

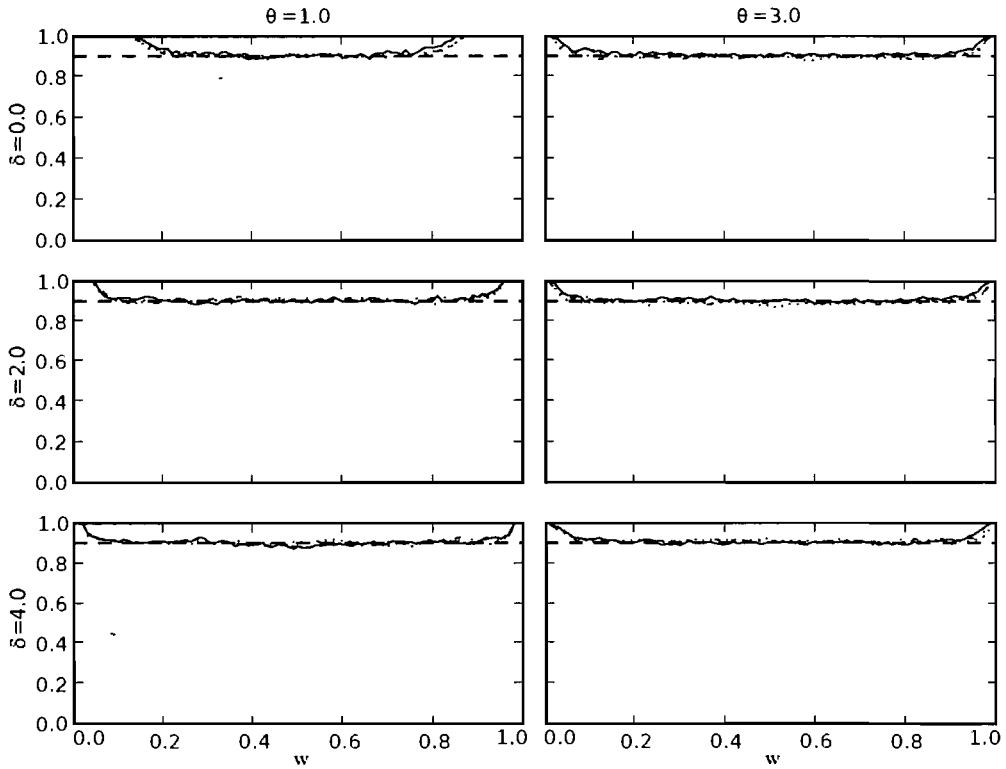

Fig. 5.12 Reprise de l'expérience des différences sous rééchantillonnage d'auto-amorçage stratifié. L'ajustement d'Agresti ajoute 1 instance à chaque possibilité.

CHAPITRE 6

Conclusion

L'objectif de ce mémoire était de présenter des outils servant à valider l'efficacité d'un test. Pour y parvenir, nous avons présenté une méthode permettant de calculer les distributions d'auto-amorçage exactes ponctuelles des courbes ROC et des courbes de coûts. Nous avons aussi présenté une approche afin de calculer les intervalles de confiance de ces distributions et tester la précision de couverture de ces mêmes intervalles.

La classification binaire étant bidimensionnelle, la prise de la moyenne des courbes ROC lors de l'auto-amorçage pouvait causer quelques inconvénients. Nous avons donc proposé deux méthodes : le moyennage vertical et le moyennage par seuil. Ces deux méthodes indiquent en fait des informations différentes à l'usager et le choix doit donc être fait en fonction de l'utilisation qu'on compte faire du modèle.

De plus, deux approches d'auto-amorçage ont été présentées : l'auto-amorçage stratifié, où les rééchantillonnages sont faits indépendamment pour les instances positives et négatives, et l'auto-amorçage complet, où le rééchantillonnage est fait à la fois sur les deux types d'instances. Les résultats indiquent que, généralement, les deux approchent performent de façon similaire.

Afin de diminuer les temps de calcul, une approximation gaussienne a été faite sur les résultats afin de calculer les intervalles de confiance. Dans ces scénarios, il a été question de l'approche idéale pour maximiser la précision de couverture. Les résultats indiquent que l'approche classique de Wald a une précision de couverture beaucoup trop faible dans la majorité des cas. Cependant, les approches d'Agresti et Coull (1998),

d'Agresti et Min (2005) et de score règle ce problème. Leurs résultats sont très satisfaisants bien qu'il reste parfois des situations où les résultats soient trop conservateurs lorsque les conditions d'opération se rapprochent trop des extrêmes.

En général, les résultats sont concluants : l'approche d'auto-amorçage exacte permet de calculer les distributions avec une précision adéquate. Il s'agit d'un avancement intéressant, principalement pour la communauté d'apprentissage machine, qui utilise fréquemment les courbes ROC et les courbes de coûts, mais qui était en manque d'outils pour évaluer leur performance. Toutefois, certains cas, spécialement celui de la section 4.4.2 sur les jeux de données combinés pour les courbes ROC, nécessitent des calculs d'ordres très élevés rendant leur application réelle très difficile.

Il serait intéressant, lors de travaux futurs, d'étudier d'autres combinaisons d'approches permettant d'augmenter la précision de couverture en même temps que de diminuer le temps de calcul, particulièrement dans le cas de jeux de données combinés lors du moyennage vertical des courbes ROC. Une autre façon d'y arriver serait de simplement ajouter une approximation pour la distribution jointe des seuils, mais il faudrait en vérifier l'impact sur la précision de couverture.

On pourrait aussi s'intéresser à diminuer l'effet conservateur lorsque les conditions d'opération sont extrêmes en tentant d'extrapoler les distributions des scores au-delà des données observées.

Bibliographie

Agresti A. et Coull B.A. Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician*, 52(2):119-226, 1998.

Agresti A. et Min Y. Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*. 24:729-740, 2005.

Bandos A. Nonparametric methods in comparing two correlated ROC curves. *PhD thesis, Graduate School of Public Health, University of Pittsburgh*, 2005.

Bradley A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145-1159, 1997.

Drummond C. et Holte R. Explicitly representing expected cost: an alternative to ROC representation. *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.* 198-207, 2000.

Drummond C. et Holte R. Cost curves: an improved method for visualizing classifier performance. *Machine Learning*. 65:95-130, 2006.

Efron B. et Tibshirani R.J. An introduction to the bootstrap. *New York: Chapman et Hall.* 1993.

Egan J.P. Signal Detection Theory and ROC Analysis, Series in Cognition and Perception. *New York: Academic Press.* 1975.

Fawcett T. et Provost F. Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*. 1(3):291-316, 1997.

Fawcett T. ROC graphs: Notes and practical considerations for researchers. *Technical Report, HP Laboratories*. 2004.

Fawcett T. et Flach A. A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*. 58(1):33-38, 2005.

Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 27(8):861-874, 2006a.

Fawcett T. ROC graphs with instance varying costs. *Pattern Recognition Letters*. 27(8):882-891, 2006b.

Fawcett T. et Niculescu-Mizil A. PAV and the ROC convex hull. *Machine Learning*. 68(1):97-106, 2007.

Flach P. et Wu S. Repairing concavities in ROC curves. *Proc. 2003 UK Workshop on Computational Intelligence.* 38-44. 2003.

Green D.M. et Swets J.M. Signal detection theory and psychophysics. *New York: John Wiley and Sons Inc. Journal of Sound and Vibration.* 5(3):519-521, 1966.

Hall P. et Hyndman R. Improved methods for bandwidth selection when estimating ROC curves. *Statistics and Probability Letters.* 64:181-189, 2003.

Hall P., Hyndman R. et Fan Y. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika.* 91:743-750, 2004.

Hanley J.A. et McNeil B.J. The meaning and use of the Area under Receiver Operating Characteristic (ROC) Curve. *Radiology.* 143:29-36, 1982.

Hilden J. et Glasziou P. Regret graphs, diagnostic uncertainty, and vouden's index. *Statistics in Medicine.* 15:969-986, 1996.

Holte R. Very simple classification rules perform well on most commonly used datasets. *Machine Learning.* 11(1):63-91, 1993.

Macskassy S., Provost F. et Rosset S. ROC confidence bands : An empirical evaluation. *Proceedings of the 22nd International Conference on Machine Learning. Boon, Germany.* 2005.

Metz C.E., Herman B.A. et Roe C.A. Statistical comparison of two ROC curve estimates obtained from partially paired datasets. *Medical Decision Making.* 18:110-121, 1998.

Mueller S.T. et Zhang J. Upper and Lower Bounds of Area Under ROC Curves and Index of Discriminability of Classifier Performance. *Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning. Pittsburgh, USA.* 2006.

Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers. MIT Press* 61-74, 2000.

Pottmann H. Basics of projective geometry. *An institute for mathematics and its applications tutorial. Geometric Design: Geometric for CAGD. Accessible sur internet (dernière verification 2008/11/09): http://www.ima.umn.edu/multimedia/spring/tut7.html.* 2001.

Preparata F.P. et Shamos M.I. Computational Geometry, An Introduction. *Text and Monographs in Computer Science. New York: Springer-Verlag.* 1988.

Press W.H., Teukolsky S.A., Vetterling W.T. et Flannery B.P. Numerical Recipes: The Art of Scientific Computing. *Cambridge University Press.* 2007.

Provost F., Fawcett T. et Kohavi R. The case Against Accuracy Estimation for Comparing Induction Algorithms. *J. Shavlik (ed.): Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA.* 445-453, 1998.

Quinlan J.R. Learning decision trees. *Machine Learning.* 1(1):1-25, 1987.

Spackman K.A. Signal detection theory: Valuable tools for evaluating inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning, San Mateo, CA. Morgan Kaufman.* 160-163, 1989.

Swets J.A. Measuring the accuracy of diagnostic systems. *Science.* 240:1285-1293, 1988.

Swets J.A., Dawes R.M. et Monahan J. Better decisions through science. *Scientific American.* 283(4):82-87, 2000.

Webb G.I. et Ting K.M. On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning.* 58(1):25-32, 2005.

Zadrozny B. et Elkan C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *Proceedings of the Eighteenth International Conference on Machine Learning.* 609-616, 2001.

Zadrozny B. et Elkan C. Transforming classifiers scores into accurate multiclass probability estimates. *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.* 694-699, 2002.

Zhou X.H., Obuchowski N.A. et McClish D.K. Statistical methods in diagnostic medicine. *New York: Wiley and Sons Inc.* 2002.

ANNEXE A

Exemple de cas où les ratios de vrais et de faux positifs ne sont pas indépendants de la distribution des classes. Traduction libre de Webb et Ting (2005).

Nous présentons un exemple simple afin d'illustrer comment des altérations de la distribution des attributs sans égard à la distribution des classes peuvent à la fois modifier la distribution des classes et changer les ratios de vrais èt de faux positifs. Considérons une tâche d'apprentissage machine inspirée d'un exemple simple de Quinlan (1987) afin de décider si on devrait aller jouer au golf. Nous cherchons à prédire le comportement d'un enthousiaste du golf dénommé John. Son comportement peut être précisément prédit en se référant à deux attributs ayant chacune deux valeurs possibles. Ces attributs sont les *conditions atmosphériques*, pouvant être *plaisantes* ou *désagréables*, et la *disponibilité*, pouvant être *occupé* ou *libre*. Les classes sont donc *joue* ou *ne joue pas* et représente le fait que John joue, ou non, au golf. La première classe étant définie comme étant la classe positive. Le concept sous-jacent est qu'on classifie dans *joue* si et seulement si on a *plaisantes* et *libre*, c'est-à-dire qu'on suppose que John joue au golf dès que les conditions sont plaisantes et que qu'il est libre. Comme nous ne considérons pas de cas où le concept peut dériver, nous posons que cette base restera inchangée. Pour rendre l'exemple aussi simple que possible, nous assumerons aussi que les deux attributs sont indépendants, c'est-à-dire que la disponibilité de John n'affecte pas la température et vice-versa. Il serait toutefois possible de créer un exemple sans cette hypothèse. Les données de base sont des observations prisent sur une année durant laquelle les fréquences des quatre combinaisons d'attributs possibles sont les mêmes. Dans le tableau ci-dessous, la colonne titrée *Initial* représente la fréquence à laquelle chaque combinaison apparaît dans ce scénario initial. Afin de retirer le problème d'erreur d'échantillonnage, nous assumerons que les fréquences des échantillons sont exactement pareil à celle des vrais probabilités.

Afin de pouvoir utiliser l'analyse ROC, il nous faut un model à analyser. Pour illustrer notre point, dans le cas où la classe est uniquement déterminée par les valeurs des attributs, il faut seulement qu'au moins une des classe soit parfois, mais pas toujours, mal classée. Si ces conditions minimales ne sont pas respectées, les ratios de vrais et de faux positifs seront invariants peu importe la distribution des données. Supposons qu'on applique la méthode d'apprentissage par souche pour la prise de décision *(decision stump learning)* [Holte, 1993]. Nous pourrions alors former un modèle qui classifierait un événement comme étant *joue* si et seulement si on a *plaisantes*. Pour ce modèle, le ratio de vrais positifs du scénario de base est 1,0 (tous les événements positifs sont correctement assignés) et le ratio de faux positifs de ce scénario est un tiers (*plaisantes,* mais *occupé* mal classifiés).

Supposons maintenant que nous utilisions les données visées telles que la distribution des classes est différente de celle des données initiales. L'analyse ROC devrait s'appliquer sans tenir compte de la distribution des classes. Afin d'illustrer notre point, nous allons augmenter la fréquence de la classe *joue* dans les données visées au niveau 0,5. Notons que la valeur 0,5 choisie n'est pas spécifiquement importante pour notre exemple. Le même effet serait apparent pour tout changement dans la distribution des classes. Tout ce qui changerait pour différentes distributions serait l'amplitude de l'effet. Il faut aussi noter que nous traitons ici d'une situation où le changement entre les

données de base et les données visées est dû à un changement dans les distributions desquelles les données sont tirées, en opposition avec un changement dans la manière d'échantillonner les données.

Tableau A.1. Exemple de distributions des données

Objet	Initial	Retraité	Intermédiaire	Propice	Paradisiaque
plaisantes, libre, joue	0,25	0,50	0,50	0,50	0,50
plaisantes, occupé, ne joue pas	0,25	0,00	0,21	0,17	0,50
désagréables, libre, ne joue pas	0,25	0,50	0,21	0,25	0,00
désagréables, occupé, ne joue pas	0,25	0,00	0,08	0,08	0,00

Lorsqu'on augmente la fréquence de la classe *joue* à 0,5, il faut donc que P(*plaisantes* et *libre* | données visées) = 0,5. Comme *plaisantes* et *libre* sont indépendants, il en découle que P(*plaisantes* | données visées) x P(*libre* | données visées) = 0,5. Cela veut dire que, puisque les *conditions atmosphériques* et la *disponibilité* de John déterminent la valeur de la variable classe, c'est seulement en variant les deux qu'il est possible d'obtenir une distribution des classes particulière. Quatre de l'infinité de combinaisons possibles sont :

1. P(*plaisantes* | données visées) reste 0,5 alors que P(*libre* | données visées) devient 1,0 (John prend sa retraite!) (colonne « Retraité » du tableau 1);
2. P(*plaisantes* | données visées) et P(*libre* | données visées) deviennent toutes deux 0,71 (colonne « Intermédiaire » du tableau 1);
3. P(*plaisantes* | données visées) devient 0,67 et P(*libre* | données visées) devient 0,75 (colonne « Propice » du tableau 1);
4. P(*plaisantes* | données visées) devient 1,0 alors que P(*libre* | données visées) reste 0,5 (John déménage au paradis!) (colonne « Paradisiaque » du tableau 1).

Le ratio de vrais positifs restera 1,0 pour toutes les possibilités. Cela se produit puisque notre modèle est une génération large du vrai concept. Cependant, de toutes les combinaisons possibles de P(*plaisantes* | données visées) et P(*libre* | données visées) générant la nouvelle distribution des classes, seulement le scénario « Propice » maintiendrait un ratio de faux positifs d'un tiers. Dans cet exemple, pour que l'analyse ROC prédise efficacement la performance d'un classifieur lors de changements de distributions des classes, il faut que l'univers soit organisé de manière à ce que la *disponibilité* d'un golfeur ne puisse (ou soit plus probable) que changer en concordance avec des changements spécifiques dans les *conditions atmosphériques*.

Si John cesse d'être *occupé* mais que les *conditions atmosphériques* ne changent pas (le scénario « Retraité ») le ratio de faux positifs devient 0,0 et l'analyse ROC la surestimera. Pour le scénario « Intermédiaire » dans lequel il y a les mêmes probabilités pour les *conditions atmosphériques* et pour la

disponibilité de John, le ratio de faux positifs est de 0,41 et l'analyse ROC la sous-estime. Si les *conditions atmosphériques* s'améliorent et deviennent invariablement *plaisantes* alors que la *disponibilité* de John ne change pas (le scénario « Paradisiaque »), le ratio de faux positifs devient 1,0 et l'analyse ROC la surestimera encore une fois.

Exemple de cas où les ratios de vrais et de faux positifs ne sont pas indépendants de la distribution des classes. Article original de Webb et Ting (2005)

We provide a simple example to illustrate how alterations to the distribution of the attributes without regard to the distribution of the class may both alter the distribution of the class and alter true and false positive rates. Consider a learning task inspired by Quinlan's (1987) classic example of deciding whether to play golf. We seek to predict the behavior of a golf enthusiast called John. John's behavior can be accurately predicted with reference to two attributes, *Playing Conditions*, with the two values *Pleasant* and *Unpleasant*, and *Other Commitments* with the two values *Busy* and *Free*. The classes are *Play* and *Don't Play*, representing respectively whether John plays golf or does not, with the former considered the positive class. The underlying concept is *Play* if and only if *Pleasant* and *Free*. That is, John plays golf whenever the weather is pleasant and he has no other commitments. As we do not consider concept drift, we do not allow this concept to alter. To make the example as simple as possible, we assume that the attributes are independent of each other. That is, John's commitments do not affect the weather and the weather does not affect John's commitments. Our ability to construct an example in no way depends upon this simplifying assumption, however. The base data are taken from observations drawn over a year for which the frequencies of each of the four combinations of attribute values are equal. Table 1 displays the four combinations of X values together with the associated class. The column titled *Initial* shows the frequency with which each combination appears in the base data. To remove sampling error as an issue, we assume that the sample frequencies exactly match the true probabilities.

In order to cast light on ROC analysis we require a model to analyze. In order to demonstrate our point, in the case where the class is uniquely determined by the attribute values, we require only that at least one class is sometimes, but not always, misclassified. If these minimal conditions are not satisfied, $TP(\cdot)$ and $FP(\cdot)$ must be invariant no matter what the data distribution. Assume we apply decision stump learning (Holte, 1993). We might form a model that classifies an occassion as *Play* if and only if *Pleasant*. For this model, $TP(base) = 1.0$ (all Play objects are correctly labelled) and $FP(base) = 1/3$ (pleasant but busy days are misclassified).

Suppose now we move to target data for which there is a different class distribution from that of the base data. ROC analysis is supposed to apply irrespective of the class distribution. For the sake of illustration we will increase the frequency of *Play* in the target data to 0.5. Note, however, that this particular frequency is not important to our example. The same effect will be apparent for any change in the class distribution. All that alters with different distributions is the magnitude of the effect. Note also that we are addressing here the situation where the change from the base to the target data represents a change in the underlying distributions from which the data are drawn, rather than a change in the way in which the data are sampled.

Table A.2. Example data distributions.

Object	Initial	Retire	Intermediate	Propitious	Paradise
Pleasant, Free, Play	0.25	0.50	0.50	0.50	0.50
Pleasant, Busy, Don't Play	0.25	0.00	0.21	0.17	0.50
Unpleasant, Free, Don't Play	0.25	0.50	0.21	0.25	0.00
Unpleasant, Busy, Don't Play	0.25	0.00	0.08	0.08	0.00

As we are increasing the frequency of *Play* to 0.5 we require that $P(Pleasant \wedge Free \mid target) = 0.5$. As *Pleasant* and *Free* are independent, it follows that we require that $P(Pleasant \mid target) \times P(Free \mid target) = 0.5$. That is, because the weather and John's commitments determine the value of the class variable, it is only by varying both the weather and John's commitments precisely in conjunction that it is possible to obtain a particular class distribution. Four of the infinite number of combinations of $P(Pleasant \mid target)$ and $P(Free \mid target)$ for which the desired class distribution are obtained are:

1. $P(Pleasant \mid target)$ remains 0.5 while $P(Free \mid target)$ rises to 1.0 (John retires!), illustrated in the *Retire* column of Table 1;
2. $P(Pleasant \mid target)$ and $P(Free \mid target)$ both rise to 0.71, illustrated in the *Intermediate* column of Table 1;
3. $P(Pleasant \mid target)$ rises to 0.67 and $P(Free \mid target)$ rises to 0.75, illustrated in the *Propitious* column of Table 1; and
4. $P(Pleasant \mid target)$ rises to 1.0 while $P(Free \mid target)$ remains 0.5 (John moves to paradise!), illustrated in the *Paradise* column of Table 1.

For all alternatives the true positive rate will remain 1.0. This is because our model happens to be an overgeneralization of the true concept. However, of all the infinite number of combinations of $P(Pleasant \mid target)$ and $P(Free \mid target)$ for which the new class distribution are obtained, only for exactly those propitious values $P(Pleasant \mid target) = 2/3$ and $P(Free \mid target) = 0.75$ does the false positive rate remain at 1/3. For this example, for ROC analysis to successfully predict classification performance under a change of class distribution requires that the world is organized so that a golfer's commitments can only (or are most likely) to change only in conjunction with specific changes in the weather.

If John ceases having other commitments but the weather does not change (the retirement scenario), the false positive rate becomes 0.0 and the ROC analysis will overestimate it. For the intermediate scenario in which there are equal increases in the frequencies both of pleasant weather and of John having no commitments, the false positive rate rises to 0.41 and the ROC analysis will underestimate it. If the weather improves so as to be invariably pleasant but John's commitments do not change (the paradise scenario), the false positive rate becomes 1.0 and the ROC analysis will again underestimate it.

ANNEXE B
Algorithmes

Voici la liste des algorithmes mentionnés dans ce mémoire. Les termes provenant directement des langages de programmation (en gras dans la section codage) sont laissés en anglais par souci de compréhension et de simplification.

Algorithme 2.1 : Générer une des points ROC [Fawcett, 2006a]
Entrées : *L, le jeu de données; f(i), la probabilité que le classifieur estime que l'exemple i est une instance positive; P et N, le nombre d'instances positives et négatives.*
Sorties : *R, une liste de point ROC en ordre croissant de ratio de faux positifs*
Pré requis : *P>0 et N>0*

1: $L_{ordonnée} \leftarrow$ *L ordonnée en f scores décroissants*
2: $FP \leftarrow TP \leftarrow 0$
3: $R \leftarrow <>$
4: $f_{précédent} \leftarrow -\infty$
5: $i \leftarrow 1$
6: ***while*** $i \leqslant |L_{ordonnée}|$ ***do***
7: ***if*** $f(i) \neq f_{précédent}$ ***then***
8: ***Push*** *(FP/N, TP/P)* ***onto*** *R*
9: $f_{précédent} \leftarrow f(i)$
10: ***if*** $L_{ordonnée}[i]$ *est positif* ***then***
11: $TP \leftarrow TP+1$
12: ***else*** /* $L_{ordonnée}[i]$ *est négatif* */
13: $FP \leftarrow FP+1$
14: $i \leftarrow i + 1$
15: ***Push*** *(FP/N, TP/P)* ***onto*** *R* /* *le point* (1,1) */
16: **end**

Algorithme 4.1 : Estimés d'auto-amorçage exacts stratifiés pour le moyennage par seuil des courbes ROC, avec intervalles de confiance ponctuels. Temps : O($n \cdot \ln(n)$)
Entrées : *Les scores pour les instances positives et négatives, la taille m, un jeu de h seuils.*
Sorties : *L'ensemble des intervalles de confiance d'auto-amorçage exacts stratifiés*

1: *Ordonner, de façon décroissante, les scores des instances positives du jeu de données et les identifier* $s^+(1), s^+(2),..., s^+(n^+)$.

2: *Ordonner, de façon décroissante, les scores des instances négatives du jeu de données et les identifier* $s^-(1), s^-(2),..., s^-(n^-)$.

3: $n_t^+ \leftarrow 0, n_t^- \leftarrow 0$
4: $\alpha_{2D} \leftarrow 1-\sqrt{1-\alpha}$
5: $z \leftarrow z_{1-\alpha_{2D}/2}$
6: ***for*** $j=1,2,....,h$ ***do***
7: $t \leftarrow j^e$ *plus grand seuil*
8: ***while*** $n_t^+ < n^+$ *et* $s^+(n_t^+ + 1) \geq t$ ***do***
9: $n_t^+ \leftarrow n_t^+ + 1$
10: **end while**

11: $p_t^+ \leftarrow n_t^+ / n^+$

12: $L_t^+ \leftarrow \dfrac{p_t^+ + z^2/2m^+ - z\sqrt{p_t^+(1-p_t^+)/m^+ + \left(z/2m^+\right)^2}}{1+z^2/m^+}$

13: $U_t^+ \leftarrow \dfrac{p_t^+ + z^2/2m^+ + z\sqrt{p_t^+(1-p_t^+)/m^+ + \left(z/2m^+\right)^2}}{1+z^2/m^+}$

14: *Performer des calculs similaires, en utilisant les scores des instances négatives, afin d'obtenir les bornes* L_t^- *et* U_t^-.

15: *Illustrer les régions de confiance rectangulaires autour des ratios estimés de faux et de vrais positifs* (p_t^-, p_t^+), *en utilisant* L_t^-, U_t^-, L_t^+ *et* U_t^+.

16: ***end for***

Algorithme 4.2 : Estimés d'auto-amorçage exacts stratifiés pour la probabilité de dominance de courbes ROC moyennées par seuil. Temps : O($n \cdot h'$)

Entrées : *Les scores pour les instances positives et négatives, la taille m, un jeu de h seuils.*

Sorties : *L'ensemble des probabilités de dominance d'auto-amorçage exactes stratifiées.*

1: ***for*** $j=1,2,\ldots,h$ ***do***

2: $(t_1,t_2) \leftarrow j^e$ *paire de seuils*

3: *calculer les valeurs de* $n^+_{t_1,t_2}$, $n^+_{t_1,t_2}$, $n^-_{t_1,t_2}$ *et* $n^-_{t_1,t_2}$.

4: $u \leftarrow n^+_{t_1,t_2} / n^+$, $v \leftarrow (1-u) n^+_{t_1,t_2} / n^+$

5: $b_u \leftarrow (1-u)^{m^+}$, $b_v \leftarrow (1-v)^{m^+}$, $c \leftarrow 1$

6: $\Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \geq 0\} \leftarrow 1$, $\Pr\{\Delta_{1,2}TP^+_{t_1,t_2} = 0\} \leftarrow 0$.

7: ***for*** $i=0,1,\ldots,\lfloor m^+/2 \rfloor$ ***do***

8: $\Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \geq 0\} \leftarrow \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \geq 0\} - b_u c$

9: $\Pr\{\Delta_{1,2}TP^+_{t_1,t_2} = 0\} \leftarrow \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} = 0\} + b_u b_v$

10: $b_u \leftarrow b_u \dfrac{u}{1-u} \dfrac{m^+ - i}{i+1}$

11: $c \leftarrow c - b_v \left(1 + \dfrac{m^+ - 2i}{m^+ - i} \dfrac{v}{1-v}\right)$

12: $b_v \leftarrow b_v \dfrac{v}{(1-v)^2} \dfrac{(m^+ - 2i)(m^+ - 2i - 1)}{(i+1)(m^+ - i)}$

13: ***end for***

14: $\Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \geq 0\} \leftarrow \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \geq 0\} + \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} = 0\}$

15: *Performer des calculs similaires pour les instances négatives en remplaçant* $\Delta_{1,2}TP^+_{t_1,t_2}$, n^+, m^+, $n^+_{t_1,t_2}$ *et* $n^+_{t_1,t_2}$ *par* $\Delta_{1,2}FP^-_{t_1,t_2}$, n^-, m^-, $n^-_{t_1,t_2}$ *et* $n^-_{t_1,t_2}$ *respectivement.*

16: $f_1(t_1,t_2) \leftarrow \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \geq 0\}\Pr\{\Delta_{1,2}FP^-_{t_1,t_2} \leq 0\} - \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} = 0\}\Pr\{\Delta_{1,2}FP^-_{t_1,t_2} = 0\}$

17: $f_2(t_1,t_2) \leftarrow \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} \leq 0\}\Pr\{\Delta_{1,2}FP^-_{t_1,t_2} \geq 0\} - \Pr\{\Delta_{1,2}TP^+_{t_1,t_2} = 0\}\Pr\{\Delta_{1,2}FP^-_{t_1,t_2} = 0\}$

18:***end for***

19:***return*** $f_1(t_1,t_2)$ et $f_2(t_1,t_2)$ pour chaque paire de seuils.

Algorithme 4.3 : Estimés d'auto-amorçage exacts stratifiés pour le moyennage vertical des courbes ROC, avec intervalles de confiance ponctuels ajustés pour une distribution gaussienne. Temps : O($n \cdot h'$)

Entrées : *Les scores pour les instances positives et négatives, la taille m, un jeu* $\{r_i/m^-, i = 1,2,\ldots,h\}$ de *h ratios de faux positifs, un niveau de confiance* α.

Sorties : *Les intervalles de confiance d'auto-amorçage exacts stratifiés pour les ratios de vrais positifs pour des ratios de faux positifs donnés.*

1*: Ordonner, de façon décroissante, les scores des instances positives du jeu de données.*

2*: Ordonner, de façon décroissante, les scores des instances négatives du jeu de données.*

3:***for*** $k=1,2,\ldots,n^-$ ***do***

4: *Calculer* n_k^+

5: $p_k^+ \leftarrow n_k^+/n^+$

6: $E\{M_k^+/m^+\} \leftarrow p_k^+$

7: $E\{(M_k^+/m^+)^2\} \leftarrow (p_k^+)^2 + p_k^+(1-p_k^+)/m^+$

8:***end for***

9: $z \leftarrow (1-\alpha/2)^e$ *quantile de la distribution gaussienne standard*

10:***for*** $i=1,2,\ldots,h$ ***do***

11: $E\{TP_{r_i}^+\} \leftarrow 0$

12: $E\{(TP_{r_i}^+)^2\} \leftarrow 0$

13: ***for*** $k=1,2,\ldots,n^-$ ***do***

14: $\Pr\{T_{r_i}^- = s_k\} \leftarrow B\left(\frac{k-1}{n^-}, r_i - 1, m^-\right) - B\left(\frac{k}{n^-}, r_i - 1, m^-\right)$

15: $E\{TP_{r_i}^+\} \leftarrow E\{TP_{r_i}^+\} + \Pr\{T_{r_i}^- = s_k\}\, E\{M_k^+/m^+\}$

16: $* E\{(TP_{r_i}^+)^2\} \leftarrow E\{(TP_{r_i}^+)^2\} + \Pr\{T_{r_i}^- = s_k\} E\{(M_k^+/m^+)^2\}$

17: ***end for***

18: $e_i \leftarrow E\{TP_{r_i}^+\}$

19: $v_i \leftarrow E\{(TP_{r_i}^+)^2\} - e_i^2$

20: $TP_{\inf,r_i}^+ \leftarrow \dfrac{e_i + z^2/2m^+ - z\sqrt{v_i + (z/2m^+)^2}}{1 + z^2/m^+}$

21: $TP_{\sup,r_i}^+ \leftarrow \dfrac{e_i + z^2/2m^+ + z\sqrt{v_i + (z/2m^+)^2}}{1 + z^2/m^+}$

22:***end for***

23:***return*** *les intervalles de confiance* $(TP_{\inf,r_i}^+, TP_{\sup,r_i}^+)$

Algorithme 4.4 : Distribution d'auto-amorçage exacte stratifiée pour la différence entre les ratios de vrais positifs de deux courbes ROC en cas de moyennage vertical.
Temps : $O(n^4)$
Entrées : *Les scores des deux modèles pour les instances positives et négatives, la taille m.*
Sorties : *La distribution, pour chaque ratio de faux positifs, de la différence entre les ratios de vrais positifs des deux modèles.*

1: *Calculer les valeurs de* $n^-_{k,j}$, $n^-_{k,\bar{j}}$, $n^-_{\bar{k},j}$, $n^-_{\bar{k},\bar{j}}$, $n^+_{k,j}$, $n^+_{k,\bar{j}}$, $n^+_{\bar{k},j}$ *et* $n^+_{\bar{k},\bar{j}}$ *pour chaque valeur de k et de j.*

2: ***for*** $r = m^-, m^- - 1, \ldots, 1$ ***do***
3: ***for*** $k = 1,2,\ldots,n^-$ ***do***
4: ***for*** $j = 1,2,\ldots,n^-$ ***do***
5: *Calculer la fonction de densité jointe* $f_r(k,j)$
6: ***end for***
7: ***end for***
8: ***end for***
9: ***for*** $k = 1,2,\ldots,n^-$ ***do***
10: ***for*** $j = 1,2,\ldots,n^-$ ***do***
11: ***for*** $d = -1, \frac{m^+ - 1}{m^+}, \ldots, \frac{-1}{m^+}, 0, \frac{1}{m^+}, \ldots, \frac{m^+ - 1}{m^+}, 1$ ***do***
12: *Calculer la fonction de distribution conditionnelle* $g_{k,j}(d)$
13: ***end for***
14: ***end for***
15: ***end for***
16: ***for*** $r = 1,2,\ldots,m^-$ ***do***
17: *Calculer la fonction de distribution inconditionnelle* $h_r(d)$
18: ***end for***
19: ***return*** h