Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés

Julien Thomas[*,**], Pierre-Emmanuel Jouve[**], Nicolas Nicoloyannis[*]

[*]Laboratoire ERIC, Université Lumière Lyon2, France
http://eric.univ-lyon2.fr
[**]Société Fenics Lyon, France
http://www.fenics-sas.com

Résumé. Les critères servant à l'évaluation de modèles d'apprentissage supervisé ainsi que ceux utilisés pour bâtir des arbres de décision sont, pour la plupart, symétriques. De manière pragmatique, cela signifie que chacune des modalités de la variable endogène se voit assigner une importance identique. Or, dans nombre de cas pratiques cela n'est pas le cas. Ainsi, on peut notamment prendre l'exemple de jeux de données fortement déséquilibrés pour lesquels l'objectif principal est l'identification des objets représentatifs de la modalité minoritaire (Aide au diagnostic, identification de phénomènes inhabituels : fraudes, pannes...). Dans ce type de situation il apparaît clairement qu'assigner une importance identique aux erreurs de prédiction ne constitue pas la meilleure des solutions. Nous proposons dans cet article un critère (pouvant servir à la fois pour l'évaluation de modèles d'apprentissage supervisé ou encore de critère utilisé pour bâtir des arbres de décision) prenant en compte cet aspect non symétrique de l'importance associée à chacune des modalités de la variable endogène. Nous proposons ensuite une évolution des modèles de type forêts aléatoires utilisant ce critère pour les jeux de données fortement déséquilibrés.

1 Introduction

L'évaluation des performances d'un modèle constitue l'étape finale de tout processus d'apprentissage supervisé. Elle est le retour nécessaire à l'utilisateur pour le guider dans la poursuite de sa fouille de données. Ces mesures, comme celles utilisées pour bâtir des arbres de décisions, sont généralement symétriques. De façon pratique, on entend par symétrique le fait que les erreurs sur chaque modalité de la variable endogène se voient attribuer une importance similaire. Or de nombreux exemples industriels nous montrent que cela n'est pas toujours le cas, en particulier lorsqu'on se trouve en présence de jeux de données fortement déséquilibrés : aide au diagnostic (Grzymala-Busse, 2000), identification de phénomènes inhabituels comme les fraudes lors des transactions par cartes bancaires (Chan, 2001) ou les pannes d'équipements de télécommunications (Weiss, 1998), et bien d'autres encore. Dans ce type de cas l'objectif principal est d'identifier les instances représentatifs de la classe minoritaire.
Il est pour cela nécessaire d'utiliser des méthodologies d'apprentissage adaptées (Weiss, 2004)

(Japkowicz, 2000), comme notamment les méthodologies sensibles au coût (Domingos, 1999) ou celles basées sur des techniques d'échantillonnage (Chawla, 2002), mais l'évaluation des performances des modèles résultant doit également prendre en considération cet aspect non symétrique de l'importance des modalités, sans se limiter à un simple taux de correction global. Une évaluation locale, c'est-à-dire par modalité, doit alors être conduite. Le taux de rappel et le taux de précision sont les deux indicateurs de base des performances d'un modèle vis-à-vis d'une modalité. Il est également possible de fusionner ces deux critères en utilisant par exemple la f-measure (Van Rijsbergen, 1979) ou de créer d'autres critères utilisant le dénombrement de chaque type d'erreurs (insertion ou omission) (Makhoul, 1999).
Nous proposons dans cet article un critère appelé PRAGMA (Precision and RecAll rates Guided Model Assessment), pouvant servir à la fois pour l'évaluation de modèles d'apprentissage supervisé ou encore de critère utilisé pour bâtir des arbres de décision, prenant en compte l'ensemble de ces aspects. Nous proposons ensuite une évolution des modèles de type forêts aléatoires utilisant ce critère pour les jeux de données fortement déséquilibrés.

2 PRAGMA : Precision and RecAll rates Guided Model Assessment

PRAGMA utilise deux principes : la notion d'importance d'une classe et la notion de préférence entre taux de rappel et taux de précision pour chaque classe.
Tout d'abord l'importance d'une classe est représentée par un coefficient θ_i fixé par l'utilisateur et utilisé en fin d'évaluation. Ensuite pour chaque classe, nous évaluons le modèle en fonction de son taux de rappel (r_i) et de son taux de précision (p_i). Cette fonction $f(r_i, p_i)$, que nous cherchons à minimaliser par analogie avec le nombre d'erreurs d'un modèle, doit avoir les propriétés suivantes :

(1) $f(0,0) = 1$, on fixe la valeur de la pire situation ($r_i = 0$ et $p_i = 0$).

(2) $f(1,1) = 0$, on fixe la valeur de la meilleure situation ($r_i = 1$ et $p_i = 1$).

(3) $\frac{df(r,p)}{dr} < 0, r \in [0;1]$, à taux de précision égal, la mesure doit diminuer lorsque le taux de rappel augmente.

(4) $\frac{df(r,p)}{dp} < 0, p \in [0;1]$, à taux de rappel égal, la mesure doit diminuer lorsque le taux de précision augmente.

Une telle fonction peut avoir la forme suivante : $f(r_i, p_i) = 1 - 0.5(r_i + p_i)$

Pour prendre en compte les souhaits de l'utilisateur en terme de préférence entre le taux de rappel et le taux de précision, nous décidons de pondérérer à la fois r_i et p_i :
$f(r_i, p_i) = 1 - 0.5(\lambda \times r_i + \Omega \times p_i) = 1 + \alpha \times r_i + \beta \times p_i$

Le ratio α/β détermine la préference entre le rappel et la précision (plus celui-ci est grand (supérieur à 1), plus le rappel est préféré ; plus celui-ci est petit (inférieur à 1), plus la précision

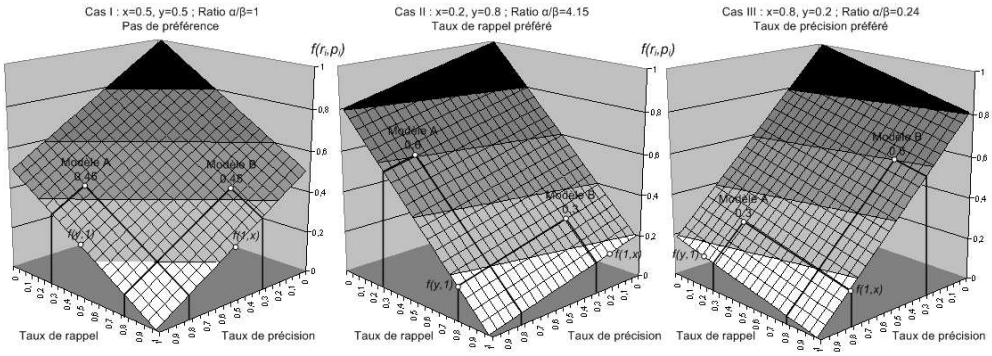

FIG. 1 – *2 modèles, A($r = 0.3; p = 0.8$) et B($r = 0.8; p = 0.3$), sont localement évalués à l'aide de 3 différentes $f(r_i, p_i)$: Cas I (symétrique) A et B sont équivalents ; Cas II (rappel préféré) B est meilleur ; Cas III (précision préférée) A est meilleur.*

est préférée ; s'il est égal à 1, cela signifie qu'aucune distinction n'est faite entre le rappel et la précision).
Pour déterminer, de manière instinctive et compréhensible, ces deux paramètres, l'utilisateur doit définir deux situations extrêmes qu'il juge de qualité équivalente. En pratique, ces deux situations sont : (a) celle où le taux de rappel est parfait ($r_i = 1$) et (b) celle où le taux de précision est parfait ($p_i = 1$). Il implique donc à l'utilisateur de définir deux valeurs x et y tel que $f(1, x) = f(y, 1)$. Choisir ces deux valeurs peut être considéré comme répondre aux deux questions suivantes :
Quel compromis êtes-vous prêt à faire vis-à-vis de la précision pour avoir un taux de rappel de 1 ? (répondre à cette question permet de définir x, avec $0 \leq x < 1$)(a)
Quel compromis êtes-vous prêt à faire vis-à-vis du rappel pour avoir un taux de précision de 1 ? (répondre à cette question permet de définir y, avec $0 \leq y < 1$)(b)
Avec cette dernière contrainte (5) $f(1, x) = f(y, 1)$, nous pouvons déterminer les paramètres α et β :

$$\alpha = \frac{-1}{1+\frac{(1-y)}{(1-x)}} \text{ et } \beta = \frac{1}{1+\frac{(1-y)}{(1-x)}} - 1$$

La fonction f utilisée pour évaluer localement un modèle selon son rappel et sa précision sur une modalité est la suivante :

$$f(r_i, p_i) = \frac{-1}{1+\frac{(1-y)}{(1-x)}} \times r_i + (\frac{1}{1+\frac{(1-y)}{(1-x)}} - 1) \times p_i + 1$$

Cette fonction correspond à l'équation d'un plan où l'axe défini par les points (0,0,1) et (1,1,0) est fixe, et où le ratio α/β détermine l'orientation du plan autour de cet axe (la préférence entre le rappel et la précision)(fig. 1).
Ces évaluations locales (propres à chaque modalité) sont ensuite combinées lors de l'évaluation finale à l'aide d'une moyenne pondérée où les coefficients d'importance de chaque classe

constituent les pondérations :

$PRAGMA = \frac{1}{\sum_{i=1}^{i=n} \theta_i} \sum_{i=1}^{i=n} \theta_i \times f_i(r_i, p_i)$, n étant le nombre de classes.

3 Evolution des modèles de type forêts aléatoires

Une forêt aléatoire (ou Random Forest RF) (Breiman, 2001) est un ensemble d'arbres de classification de profondeur maximale, chacun construit à partir d'un échantillon bootstrap du jeu d'apprentissage. De plus, pour l'obtention de chaque noeud on limite la recherche de la meilleure discrimination à k variables tirées au sort. La prédiction pour un objet est obtenue en comptabilisant les prédictions de chaque arbre pour l'objet (chaque arbre vote pour une modalité) puis en choisissant la modalité ayant reçu le plus de voix parmi tous les arbres de la forêt (vote à la majorité).
Les performances d'une forêt sont sensiblement supérieures à celles d'un arbre seul tel que C4.5 (Leon, 2004). Elle est également plus robuste au bruit et présente de meilleures facultés de généralisation (Breiman, 2001). Cependant, celle-ci n'est pas spécifiquement adaptée aux jeux de données déséquilibrés, et ses deux paramètres (le nombre d'arbres et le nombre k de variables à tirer au sort) ne permettent pas à l'utilisateur de spécifier ses préférences en termes de taux de rappel et de précision selon chaque modalité.
L'évolution que nous proposons ici consiste à remplacer l'étape du vote classique à la majorité par une nouvelle stratégie de vote pondéré où la recherche automatique des poids optimaux se fait à l'aide de PRAGMA.

Stratégie de vote. Notre stratégie de vote consiste à donner plus ou moins d'importance aux voix attribuées par les arbres (fig. 2). Une pondération par classe est déterminée (soit par l'utilisateur, soit automatiquement), laquelle multiplie le nombre de voix reçues par l'individu pour cette classe. Ainsi la modalité assignée à un objet n'est pas toujours celle dont il a reçu le plus de voix, mais celle dont le nombre de voix multiplié par son poids est le plus grand. Ceci permet d'augmenter les taux de rappel des classes minoritaires en leur affectant des pondérations fortes, ou plus généralement de jouer sur les taux de rappel et de précision de chaque classe en modifiant leur pondération.

Recherche automatique. Il peut être assez difficile de trouver manuellement les pondérations ajustant au mieux les résultats du modèle aux besoins de l'utilisateur. Si pour un problème à deux modalités tout peut se ramener à un déplacement de la frontière, en terme de nombre de votes, entre les deux classes, dès qu'il y a plus de trois modalités le nombre de possibilités de paramétrage, c'est-à-dire de ratios entre chaque couple de pondérations, devient bien plus conséquent, et il apparaît nécessaire de rendre automatique la recherche des pondérations.
L'algorithme utilisé pour automatiser la recherche des pondérations est construit autour d'un recuit simulé (Kirkpatrick, 1983) cherchant à optimiser la mesure PRAGMA paramétrée selon les souhaits de l'utilisateur. Ce procédé est parfaitement adapté et efficace pour ce type d'optimisation. Le surcoût calculatoire (comparé à une forêt aléatoire classique) est extrêmement

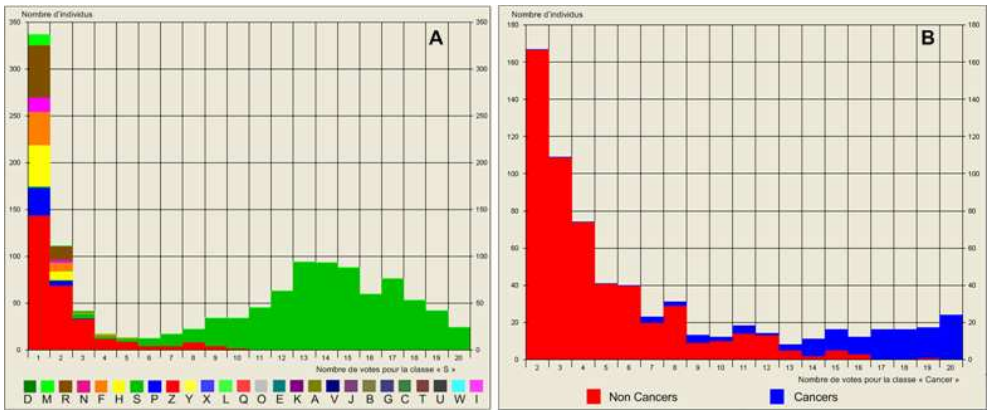

FIG. 2 – *Exemples de distribution de votes pour : (A) le jeu de données Letters(Hettich, 1999)(les objets sans vote pour la modalité 'S' ne sont pas représentés) ; (B) le jeu de données Mammo (voir Expérimentations)(les objets ayant moins de 2 votes pour la modalité 'Cancer' ne sont pas représentés).*

faible. En effet, la forêt n'est construite qu'une fois, seul le résultat du vote après pondération est mis à jour pour en permettre l'évaluation par PRAGMA. De plus, en conservant pour chaque individu le nombre de votes non pondérés qu'il a reçu pour chaque modalité, il suffit de mettre à jour uniquement la matrice de confusion après pondération du vote pour évaluer le modèle et passer à l'itération suivante.

4 Expérimentations

Nous présentons dans cette section les résultats obtenus par pondération automatique des votes d'une forêt aléatoire. Deux types de situations ont été envisagés :
(1) Utilisation de l'optimisation sur des jeux de données équilibrés. Notre but ici, en tant qu'utilisateur, est de favoriser un maximum le taux de rappel de certaines classes jugées 'prioritaires'. Les tests sont réalisés sur les jeux de données de référence Autos et Letters (Hettich, 1999) dont les variables endogènes possèdent respectivement 6 et 26 modalités.
(2) Utilisation de l'optimisation des jeux de données déséquilibrés à 2 modalités. Notre but dans cette situation est de favoriser un maximum le taux de rappel de la classe minoritaire. Les tests sont réalisés sur les jeux Hypothyroïd et Satimage (Hettich, 1999) réduits à 2 classes (minoritaire ; fusion des autres classes), ainsi que sur le jeu Mammo, issu de la mise au point d'un système d'aide au diagnostic du cancer du sein.

4.1 Jeux de données équilibrés

Nous supposerons ici que l'utilisateur cherche à maximiser les taux de rappel des classes '_3' et '_2' pour le jeu Autos, et les taux de rappel des voyelles pour le jeu Letters. Ceci se

traduit par le paramétrage de la fonction PRAGMA suivant : coefficient d'importance 10 et couple $(x; y) = (10; 90)$ pour les classes prioritaires, coefficient d'importance 1 et couple $(x; y) = (80; 80)$ pour les autres classes. Nous utilisons des forêts aléatoires de 20 arbres, avec respectivement 5 et 4 variables pour la randomisation. Les résultats présentés dans les tables 1 et 2 sont issus d'une 10-CrossValidation. La figure 3 montre les résultats détaillés sur le jeu Letters. Notez que la classe '__2' du jeu Autos ne contient que 3 objets, les résultats propres à cette modalité sont peu significatifs. Les différentes moyennes réalisées sont toujours pondérées par les effectifs des différentes classes.

	'_3'	'_2'	'_1'	'_0'	'__1'	'__2'	MP	MA	MG
RF Class. Rappel	81.4	68.7	74.1	85.1	81.8	33.3	74.6	79.4	78.0
RF Class. Précision	84.6	84.6	72.7	77.0	81.8	50.0	84.6	75.6	78.2
RF Opt. Rappel	92.6	71.9	74.1	83.6	68.2	66.7	81.4	77.4	78.5
RF Opt. Précision	80.6	82.1	76.9	75.7	83.3	100.0	81.5	77.8	78.8
Evolution Rappel	**+11.2**	**+3.2**	0.0	-1.5	-13.6	+33.4	**+6.8**	**-2.0**	**+0.5**
Evolution Précision	-4.0	-2.5	+4.2	-1.3	+1.5	+50.0	**-3.1**	**+2.2**	**+0.6**

TAB. 1 – *Résultats pour Autos : le taux de correction global de la RF Classique est de 78.0, celui de la RF Optimisée est de 78.5, soit une amélioration +0.5pts. Légende : MP Moyenne des classes Prioritaires ; MA Moyenne des Autres classes ; MG Moyenne Globale.*

	Moyenne Consonnes	Moyenne Voyelles	Moyenne Globale
RF Class. Rappel	88.0	88.5	88.1
RF Class. Précision	87.9	90.8	88.5
RF Opt. Rappel	84.8	95.0	87.1
RF Opt. Précision	90.9	78.9	88.1
Evolution Rappel	**-3.2**	**+6.5**	**-1.0**
Evolution Précision	**+3.0**	**-11.9**	**-0.4**

TAB. 2 – *Résultats pour Letters : le taux de correction global de la RF Classique est de 88.1, celui de la RF Optimisée est de 87.2, soit une perte -0.9pts.*

Ces différents résultats montrent la capacité de l'optimisation à retranscrire les volontés de l'utilisateur. Pour les deux jeux de données, les taux de rappel des classes ciblées ont augmenté. Il en résulte également (de manière logique) : (1) une baisse du taux de précision pour ces mêmes classes ; (2) une baisse du taux de rappel et une augmentation du taux de précision (en moyenne) pour les classes où aucune préférence n'avait été spécifiée. Notons également que ces changements n'entraînent pas forcément une diminution du taux de correction global. Celui-ci peut augmenter ou diminuer selon les jeux de données et le paramétrage de la mesure PRAGMA (ici augmentation du taux correction global pour Autos et diminution sur Letters).

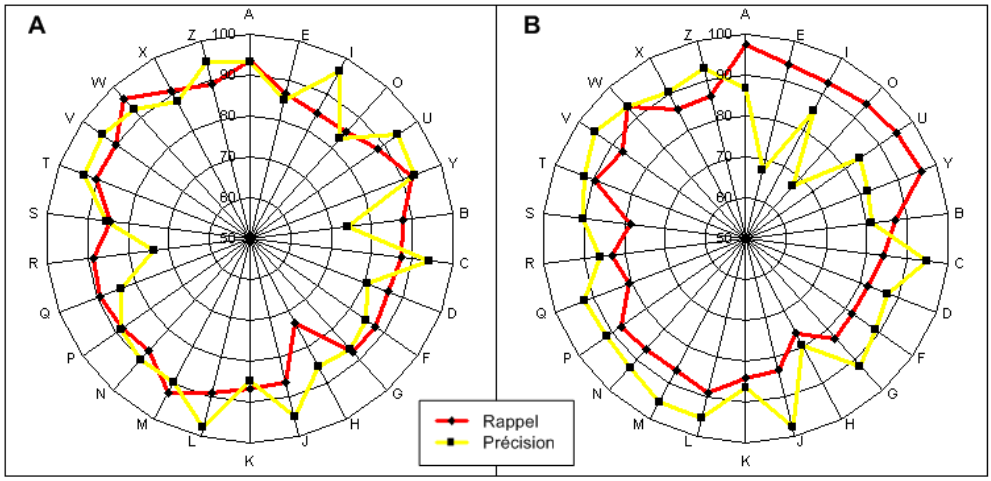

FIG. 3 – *Résultats détaillés pour Letters : (A) RF Classique ; (B) RF Optimisée, le rappel et la précision "s'organisent" selon les préférences de l'utilisateur.*

4.2 Jeux de données déséquilibrés

Ce type de jeux de données assez courant dans le milieu industriel (détection de phénomènes anormaux, fraudes, pannes, aide au diagnostic...) constitue un réel challenge pour l'apprentissage automatique. L'objectif principal est de détecter un maximum d'objets de la classe minoritaire (taux de rappel élevé) sans présenter trop de faux positifs (taux de précision correct) ce qui aurait pour effet de rendre de tels systèmes inutilisables.
Nos tests sont réalisés sur 3 jeux de données (table 3) : Hypothyroïd et Satimage (Hettich, 1999) réduits à deux classes en fusionnant les classes non minoritaires et Mammo issu de la mise au point d'un système d'aide au diagnostic du cancer du sein. Notons que ce dernier a été réduit en terme de variables et d'objets pour le rendre plus difficile et ne pas dévoiler des résultats industriels confidentiels.
La volonté de maximiser le taux de rappel de la classe minoritaire se traduit par le paramétrage de PRAGMA suivant : coefficient d'importance 10 et couple $(x; y) = (10; 90)$ pour la classe minoritaire, coefficient d'importance 1 et couple $(x; y) = (80; 80)$ pour la classe majoritaire. Nous présentons les résultats obtenus en 10-CrossValidation avec C4.5 (témoin de référence des difficultés pouvant présenter les jeux de données), une forêt aléatoire classique, et une forêt aléatoire optimisée par pondération des votes à l'aide de la mesure PRAGMA (table 4). Les forêts sont composées de 20 arbres, avec respectivement 5, 6 et 15 variables utilisées lors de la randomisation pour les 3 jeux de données.

Les taux du rappel des classes minoritaires les plus élevés sont systématiquement obtenus par la forêt aléatoire optimisée, ceci sans provoquer de fortes baisses des taux de précision. La mesure PRAGMA guide en cela parfaitement le modèle vers les performances souhaitées par l'utilisateur. On remarque également que selon les cas l'optimisation permet également parfois

Jeux de données	Effectif	Nombre de variables exogènes	Fréquence de la classe minoritaire
Hypothyroïd	3772	27	7.71
Satimage	6435	36	9.73
Mammo	3528	134	5.00

TAB. 3 – *Composition des jeux de données déséquilibrés utilisés.*

d'améliorer le taux de précision de la classe minoritaire ou le taux de correction globale.

Des résultats détaillés obtenus en 10-CrossValidation sur le jeu Mammo sont présentés en figure 4. Ils permettent une meilleure description de l'effet de la pondération des votes et de l'utilisation de la mesure PRAGMA sur les performances du modèle. Quatre indices (taux de rappel, taux de précision, nombre d'erreurs, mesure PRAGMA) sont évalués pour différentes valeurs du ratio R : pondération de la classe 'Cancer' / pondération de la classe 'Non Cancer'. On remarque que les plus fortes variations pour les taux de rappel et de précision se produisent pour la classe 'Cancer' de par son effectif faible. Le graphe du nombre d'erreurs présente deux caractéristiques notables : (1) celui-ci est asymétrique, car une baisse légère du taux de rappel sur la classe majoritaire due à une forte pondération de la classe minoritaire crée logiquement plus d'erreurs qu'une faible variation du taux de rappel de la classe minoritaire ; (2) pour les ratios $2 \leq R \leq 5$ le nombre d'erreurs total varie très peu alors que la nature des erreurs change (voir les graphes des taux de rappel et de précision). Une sorte de transfert d'erreurs se produit :
$R \leq 2$: les objets mal classés appartiennent majoritairement à la classe 'Cancer'
$3 \leq R \leq 4$: les proportions d'objets mal classés pour chacune des 2 classes sont similaires
$5 \leq R$: les objets mal classés appartiennent majoritairement à la classe 'Non Cancer'
La mesure PRAGMA permet de faire différentes observations : (1) l'asymétrie est inversée, montrant ainsi que les variations du taux de rappel de la classe 'Cancer' constituent la principale influence de la mesure (ceci s'expliquant par le fort coefficient d'importance et le paramétrage orienté vers le taux de rappel pour la classe 'Cancer') ; (2) pour les ratio $2 \leq R \leq 5$

		Hypothyroïd	Satimage	Mammo
C4.5	Rappel classe minoritaire	96.9	55.0	8.3
	Précision classe minoritaire	95.2	59.0	66.7
	Taux de correction globale	**99.4**	91.9	98.1
RF Classique	Rappel classe minoritaire	95.1	50.9	29.2
	Précision classe minoritaire	94.2	**83.2**	**84.5**
	Taux de correction globale	99.1	94.2	98.5
RF Optimisée	Rappel classe minoritaire	**99.5**	**62.0**	**43.5**
	Précision classe minoritaire	**97.7**	73.9	82.0
	Taux de correction globale	99.2	**94.3**	**98.7**

TAB. 4 – *Résultats obtenus en 10-CrossValidation pour Hypothyroïd, Satimage et Mammo.*

le plateau observé pour le graphe du nombre d'erreurs disparaît au profil de variations plus importantes. Ceci montre que la nature des erreurs est prise en considération par la mesure PRAGMA pour laquelle une erreur de classement d'individus de la classe 'Cancer' fait davantage augmenter la mesure que celle d'individus de la classe 'Non Cancer'. Les souhaits de l'utilisateur de rendre la classe 'Cancer' plus importante et de favoriser son taux de rappel vis-à-vis de son taux de précision sont ainsi visibles à travers la lecture du graphe de la mesure PRAGMA.

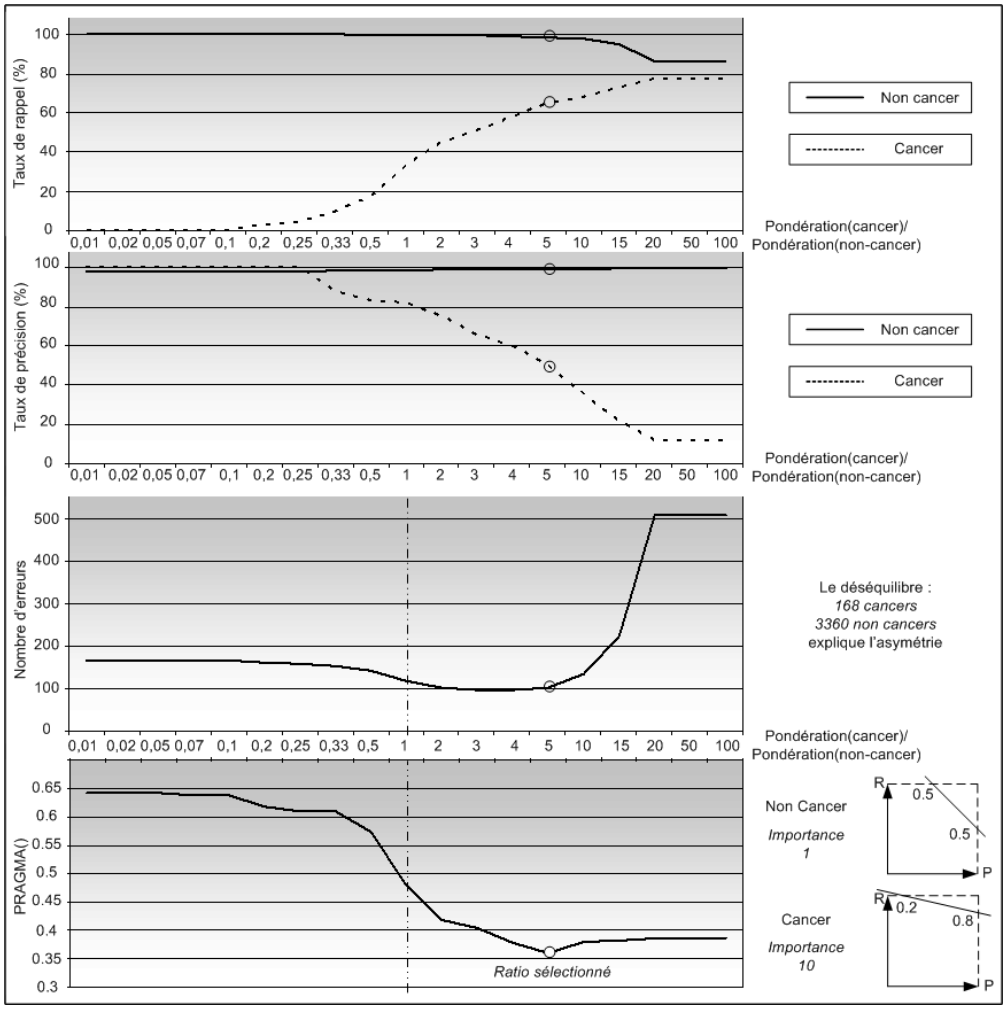

FIG. 4 – *Résultats détaillés pour le jeu de données Mammo.*

5 Conclusion

Nous proposons dans cet article une nouvelle mesure de qualité des performances des modèles d'apprentissage supervisé appelée PRAGMA (Precision and RecAll rates Guided Model Assessment). Ce critère permet à l'utilisateur d'évaluer ses modèles vis-à-vis de ses attentes en termes de taux de rappel, de taux de précision et d'importance de chaque classe sous la forme d'une mesure unique. Nous montrons ensuite comment il est possible d'utiliser cette mesure comme critère à optimiser pour orienter les performances d'un modèle. L'exemple présenté ici est l'optimisation des forêts aléatoires par pondération (selon les différentes modalités) des votes. Les résultats montrent que cette adaptation permet à l'utilisateur d'orienter de manière effective les performances des forêts aléatoires selon ses souhaits.
Dans nos travaux futurs nous projetons de tester d'autres types de fonctions que celle d'un plan pour l'évaluation locale d'une modalité vis-à-vis de son taux de rappel et de son taux de précision. Nous travaillons d'ores et déjà sur l'utilisation de la mesure PRAGMA comme critère de construction des arbres de décision en remplacement des différentes mesures d'entropie classiquement utilisées, mais nos différents tests ne sont pas encore finalisés.

Remerciements Nous tenons à remercier les membres de la société Fenics Sas (France), en particulier Simon Marcellin, Jérémy Clech, et Anne-Sophie Darnand, ainsi qu'Elie Prudhomme de l'université Lumière Lyon2 pour leur aide et leurs nombreux conseils qui ont permis à cet article de voir le jour. Ce travail a été réalisé dans le cadre d'une thèse cofinancée par le Ministère de la Recherche et de l'Industrie.

Références

Breiman L. (2001). *Random Forests*. In Machine Learning, vol 45(1), pp 5-32.

Chan P.K., Stolfo S.J. (2001). *Toward scalable learning with non-uniform class and cost distributions : a case study in credit card fraud detection*. Proc. of the 4th International Conf. on Knowledge Discovery and DataMining, 164-168.

Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002). *Smote : Synthetic minority oversampling technique*. Journal of Artificial Intelligence Research, 16, pp 321-357.

Domingos P. (1999). *Metacost : A general method for making classifiers cost-sensitive*. Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 155-164, San Diego, ACM Press.

Grzymala-Busse J.W., Zheng Z., Goodwin L. K. (2000). *An approach to imbalanced data sets based on changing rule strength*. Learning from Imbalanced Data Sets : Papers from the AAAI Workshop, pp 69-74, AAAI Press Technical Report WS-00-05.

Hettich S., Bay S.D. (1999). *The UCI KDD Archive*. Irvine, University of California, USA, Department of Information and Computer Science.

Japkowicz N. (2000). *The Class Imbalance Problem : Significance and Strategies*. Proc. of International Conf. on Artificial Intelligence (IC-AI'2000).

Kirkpatrick S., Gelatt Jr. C.D., Vecchi M.P. (1983). *Optimization by Simulated Annealing*. Science Volume 220, Number 4598.

Leon F., Zaharia M.H., Gâlea D. (2004). *Performance Analysis of Categorization Algorithms*. Proc. of the 8th International Symposium on Automatic Control and Computer Science, ISBN 973-621-086-3.

Makhoul J., Kubala F., Schwartz R., Weischedel R. (1999). *Performance measures for information extraction*. Proc. of DARPA Broadcast News Workshop, Herndon, VA.

Van Rijsbergen C. J. (1979). *Information Retrieval*. London, Butterworth, pp.174.

Weiss G.M. (2004). *Mining with Rarity : A Unifying Framework*. SIGKDD Explorations, 6(1) :7-19.

Weiss G.M., Hirsh H. (1998). *Learning to predict rare events in event sequences*. Proc. of the 4th International Conference on Knowledge Discovery and Data Mining, pp 359-363.

Summary

Evaluation of supervised learning models, as well as decision trees building, are mostly done with symetrical criteria. Pragmatically that means that each modality of the target attribute has the same importance. However, this is not the case in many practical situations. Thus, one obvious example are strongly imbalanced datasets, in this case the aim is mainly the identification of objects representing the minority class (computer aided diagnosis, identification of unusual phenomena: frauds, breakdowns...). In this situations, assigning the same importance to each kind of prediction error does not constitute the best solution. We propose in this paper a criterion (that may be used for evaluation of supervised learning as well as for decision trees building) which takes into account this nonsymetrical aspect of the importance associated to each modality of the target attribute. Afterwards, we propose an evolution of random forests that uses this criterion and which is better adapted to strongly imbalanced datasets.