Une méthode d'ensemble en apprentissage non supervisé quand on ne connaît rien sur la performance des experts ?

Antoine Cornuéjols, Christine Martin

AgroParisTech, département MMIP et INRA UMR-518
16, rue Claude Bernard
F-75231 Paris Cedex 5 (France)
{antoine.cornuejols,christine.martin}@agroparistech.fr
http://www.agroparistech.fr/mia/equipes:membres:page:antoine

Résumé. Les méthodes d'ensemble ou les méthodes collaboratives supposent que la performance de chaque « expert » soit mesurable (cas de l'apprentissage supervisé) ou soit estimée *a priori* (cas non supervisé) afin d'attribuer un poids ou une confiance aux informations échangées pour construire la fonction de décision finale.

Nous présentons ici une méthode d'apprentissage non supervisé, dans le cas à deux classes inconnues, s'appuyant sur une base d'experts qui sont des boîtes noires dont la performance par rapport aux régularités cibles est inconnue. Nous montrons comment sélectionner automatiquement, dans cette base, des experts performants sur la tâche cible et comment combiner leur résultats pour obtenir une fonction de décision généralement aussi bonne ou meilleure que le meilleur expert (inconnu) de la base. Les expériences réalisées confirment le bon fonctionnement de la méthode.

1 Introduction

Les méthodes d'ensemble en apprentissage artificiel consistent à chercher à profiter de l'expertise variée de diverses fonctions de décision pour obtenir par combinaison une fonction de décision finale que l'on espère meilleure que chacune des fonctions de décision considérée isolément. Ces méta-méthodes impliquent deux opérations principales, la première étant la *sélection* de fonctions de décision utiles, la seconde étant la *combinaison* des fonctions sélectionnées pour obtenir une fonction de décision finale agrégée. Les méthodes d'ensemble ont d'abord été développées en apprentissage supervisé dans divers contextes (*learning from expert advice* (Freund et Schapire (1997)), l'algorithme de majorité pondérée (Littlestone et Warmuth (1994)), l'apprentissage enligne (Cesa-Bianchi et Lugosi (2006)), ...). En apprentissage supervisé, en effet, il est possible d'estimer la performance de chaque fonction de décision, par exemple sur un ensemble de validation ou par validation croisée, ce qui permet de détecter les fonctions de décision utiles (ou faibles, c'est-à-dire au moins un peu meilleures que le hasard)

et ensuite de les combiner en prenant en compte la performance mesurée de chaque fonction retenue.

En apprentissage non supervisé, la situation est beaucoup moins claire. Comme en apprentissage supervisée, il n'existe pas de méthode universellement bonne, c'est-à-dire s'appliquant à tous les problèmes. Chaque méthode est biaisée vers la mise en évidence de certains types de régularités. Lorsqu'un nouveau problème est étudié, il n'est pas évident de savoir *a priori* si une méthode donnée est appropriée, et jusqu'à quel point, pour identifier les régularités espérées. Il est tout à fait possible qu'une méthode soit inadaptée pour identifier ce qui est recherché, comme il peut être illusoire d'attendre qu'un professeur de gymnastique sache détecter les élèves bons en math. Or, contrairement à l'apprentissage supervisé, il n'existe pas de juge de paix pour estimer la valeur des méthodes sur la tâche étudiée. Comment dès lors détecter les méthodes utiles et savoir en combiner les résultats ?

Les travaux existants sur les méthodes d'ensemble en apprentissage non supervisé n'affrontent pas ce problème dans toute sa généralité. L'hypothèse de départ est que les méthodes disponibles sont pertinentes à un degré ou un autre pour la tâche abordée. L'étape de sélection est ainsi court-circuitée. Le souci principal de ces méthodes d'ensemble est de réduire l'instabilité des méthodes de clustering, c'est-à-dire leur propension à retourner des résultats différents sur un même jeu de données, en combinant leurs résultats (Dimitriadou et al., 2001; Domeniconi et Al-Razgan, 2009). Si l'instabilité est effectivement réduite par cette approche, le résultat final n'est pas pour autant garanti d'être bon. L'agrégation d'avis d'"experts" dont certains peuvent ne pas être adaptés à la tâche n'a en effet pas de raison de donner miraculeusement un bon résultat. C'est pourquoi, des travaux récents proposent des méthodes pour sélectionner *a posteriori* les résultats à agréger en fonction de leur *qualité* et de leur *diversité* (Alizadeh et al., 2014; Fern et Lin, 2008). Malheureusement, les critères de qualité et de diversité traduisent eux-mêmes des *a priori* subjectifs et le problème de la sélection n'est donc pas véritablement résolu.

Les tâches les plus étudiées en apprentissage non supervisé comprennent (i) la découverte de classes dans un ensemble d'exemples, ce que l'on appelle le *clustering* ou *catégorisation*, (ii) la *sélection d'attributs* qui cherche à identifier les descripteurs informatifs pour un objectif donné et qui peut être considérée comme un cas particulier de clustering dans lequel on suppose *a priori* l'existence de deux classes : celle des attributs utiles et celle des autres attributs, et (iii) la recherche d'exemples "nouveaux" dans une population d'exemples jugés "nominaux" ou "normaux" quand les deux classes "nominal" et "nouveau" sont *a priori* non définies.

Dans la suite, nous nous intéresserons à des problèmes de clustering à deux classes, auxquels peuvent se rapporter également la sélection d'attributs et la détection de nouveautés, c'est-à-dire d'une classe nouvelle par rapport à une classe nominale. Cependant l'approche présentée peut s'étendre au tri ("ranking") d'un ensemble d'éléments, ce qui fera l'objet d'autres publications.

Dans le contexte de l'utilisation d'une méthode d'ensemble pour ce type de tâches, à quelles méthodes de base peut-on avoir recours ?

La plupart des travaux considèrent un ensemble de méthodes de clustering. Par exemple l'ensemble peut être constitué d'algorithmes de k-moyenne avec des valeurs de

k différentes, éventuellement aussi des distances différentes, en plus d'algorithmes utilisant d'autres principes. Cependant, prendre comme méthode de base des algorithmes de clustering peut comporter des risques.

Par exemple, lorsque l'une des classes à trouver est très minoritaire (e.g. moins de 5%), les méthodes classiques de clustering sont souvent de performance médiocre car, sauf dans le cas de classes très contrastées, elles ont tendance à produire des classes d'effectifs comparables. C'est pourquoi, en particulier en sélection d'attributs mais pas seulement, on a souvent recours à des méthodes d'évaluation des "objets" : attributs ou exemples, qui retournent soit un poids associé à chaque objet, soit un classement général de ces objets. C'est le cas par exemple des méthodes dites de *filtre* en sélection d'attributs, telles que Relief (Kononenko, 1994), Anova, ou *Symmetrical Uncertainty* qui est basée sur une mesure de gain d'entropie (voir (Press et al., 2007) p.761).

L'un des avantages de ces méthodes d'évaluation est qu'elles sont de complexité calculatoire réduite et sont donc rapides. Une des limites est que là aussi il n'existe évidemment pas de fonction d'évaluation universellement bonne pour trier de manière appropriée à toute tâche les objets. En l'absence d'information supplémentaire, l'utilisateur est donc réduit à espérer que la fonction d'évaluation qu'il emploie est effectivement "alignée" avec les régularités qu'il cherche à mettre en évidence. Il doit aussi en général fixer *a priori* un seuil de sélection ou un nombre d'objets à retenir.

Curieusement, à notre connaissance, il n'existe que très peu de publications sur l'emploi de méthodes d'ensemble dans le contexte de la sélection d'attributs. Si les mérites comparés des approches "*filter*", "*wrapper*" et "*embedded*" donnent lieu à des débats nourris (Blum et Langley, 1997; Cornuéjols et Miclet, 2010; Guyon et Elisseeff, 2003; Kohavi et John, 1997), il ne semble pas que l'on ait tenté de comparer des méthodes de sélection d'attributs à des méthodes faibles, comme en apprentissage supervisé (Schapire et Freund, 2012; Zhou, 2012), et que l'on ait dès lors cherché à les sélectionner et les combiner dans l'espoir d'améliorer la performance globale. Ainsi, dans (Saeys et al., 2008), l'une des rares publications sur ce sujet, les auteurs préconisent une approche d'ensemble, mais c'est ici aussi pour améliorer la stabilité de la sélection d'attributs par rapport à des variations de l'ensemble des exemples. Les méthodes de sélection considérées sont engendrées en modifiant l'ensemble des exemples pris en compte par *une même méthode* de sélection et l'approche générale est celle du bagging. Si la stabilité est améliorée, la performance en terme de classification s'appuyant sur les attributs sélectionnés n'est pas modifiée de manière significative (moins de 1% dans le meilleur des cas). De plus, l'approche présuppose que la méthode de sélection de base, par exemple Relief, est appropriée. Ici aussi, il n'y a pas de véritable stratégie de sélection des méthodes de base. Elles sont toutes considérées *a priori* pertinentes et combinées par simple vote.

Dans le travail présenté ici, nous considérons des méthodes de base qui sont des fonctions d'évaluation. Nous supposons qu'est disponible une librairie $\mathcal{F}$ de fonctions d'évaluation d'objets, attributs ou exemples, pour lesquelles on ne connaît pas *a priori* leur pertinence pour la tâche en cours. Chaque fonction d'évaluation $f_i : \mathcal{X} \to \mathbb{R}$ permet d'associer un nombre à un objet $\mathbf{x}$. Appliquée à un échantillon de données $\mathcal{S}$, chacune de ces fonctions induit un tri sur $\mathcal{S}$ en plaçant par exemple en haut du classement produit les objets $\mathbf{x}$ d'évaluations les mieux évalués. Étant donné un seuil, il est alors possible

de distinguer les objets '+' (au-dessus du seuil dans le classement) des objets '−'. On peut ainsi traiter des problèmes de clustering à deux classes, de sélection d'attributs ou de découverte de nouveautés.

La question posée alors est la suivante : étant donné un échantillon de m objets $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ pour lesquels la seule supposition est qu'il existe une classe d'objets "nominaux" et une classe d'objets "nouveaux" ou "anormaux" (ou bien il s'agit d'attributs pertinents et non pertinents), est-il possible de sélectionner automatiquement des fonctions d'évaluation f_i utiles pour les identifier et peut-on combiner leurs résultats pour obtenir un résultat aussi bon, voire meilleur que la meilleure fonction d'évaluation dans $\mathcal{F}$ qui est inconnue *a priori* ?

Par exemple, la tâche étudiée pourrait être celle d'une administration fiscale cherchant à détecter les citoyens fraudeurs grâce à un fichier décrivant les contribuables, ou bien celle d'un biologiste voulant découvrir le sous-ensemble de gènes associés à la réponse d'un organisme à une condition environnementale donnée. Ces utilisateurs pourraient disposer de fonctions d'évaluation dont ils ne connaissent pas la pertinence pour la tâche visée. Leur serait-il alors possible d'utiliser une approche d'ensemble pour obtenir un classement final aussi bon, voire meilleur, que celui qui serait produit par la meilleure fonction d'évaluation disponible mais dont on en connaît pas l'identité ?

Dans la suite, nous montrons comment sélectionner des fonctions d'évaluation utiles, sans aucune information *a priori* sur leur mérite, et comment les combiner pour faire ressortir les exemples '+'. Nous avons appliqué l'algorithme résultant à des données artificielles inspirées d'un problème réel en génomique, montrant que, malgré la grande diversité des fonctions dévaluation initiales, seules des fonctions utiles sont sélectionnées, à l'aide desquelles un tri final sur $\mathcal{S}$ est réalisé, tendant à mettre en tête de classement les exemples '+'.

La méthode présentée est une méthode d'ensemble (Dietterich, 2000; Zhou, 2012) dont le principe est présenté en section 2. Pour être performante, une méthode d'ensemble doit recourir à des fonctions d'évaluation "diverses", c'est-à-dire peu corrélées. La section 3 présente une mesure de corrélation entre fonctions. Il faut ensuite savoir identifier les fonctions utiles (section 4), puis les combiner pour obtenir la fonction globale finale (section 6). La section 5 rapporte les expériences réalisées pour tester la validité de l'approche et les résultats obtenus. Les conclusions et les perspectives pour aller plus loin sont présentées en section 7.

2 Le principe de la méthode

On suppose que l'échantillon d'objets à analyser $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ est issu d'un mélange de distributions sur l'espace d'entrée $\mathcal{X}$: la distribution $\mathbf{P}^-_{\mathcal{X}}$ des exemples nominaux et la distribution $\mathbf{P}^+_{\mathcal{X}}$ des exemples d'intérêt dont on cherche à identifier les représentants dans $\mathcal{S}$.

On suppose de plus que, dans l'ignorance d'une fonction permettant de distinguer sûrement les exemples de la classe '+' de ceux de la classe '−', on dispose d'une librairie $\mathcal{F}$ de fonctions d'évaluation ou fonctions de score $f_i : \mathcal{X} \to \mathbb{R}$ pour lesquelles on ne sait rien *a priori* de leur pertinence pour distinguer les deux classes, c'est-à-dire de leur

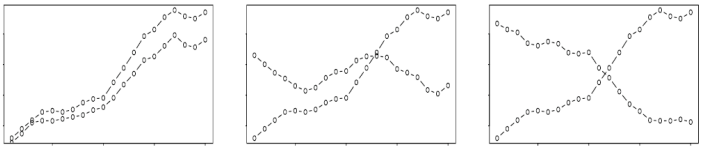

FIG. 1 – *Illustration de la mesure de corrélation de Pearson sur des couples de courbes.*

propension à donner un score supérieur (ou inférieur) aux objets de la classe '+' par rapport aux objets de la classe '−'.

2.1 Fonctions d'évaluation faibles

On espère que parmi les fonctions de $\mathcal{F}$ se trouvent des *fonctions d'évaluation faibles*, c'est-à-dire positivement corrélés, même faiblement, avec la fonction de tri idéale qui place tous les exemples de la classe '+' avant ceux de la classe '−'. Pour mesurer cette corrélation entre la méthode étudiée et la méthode idéale inconnue, on pourrait utiliser le *coefficient de corrélation de Pearson* s'appliquant directement aux scores :

$$S(f_i, f_j) \;=\; \frac{\sum_l (f_i^l - \mu_{f_i})(f_j^l - \mu_{f_j})}{\sqrt{\sum_l (f_i^l - \mu_{f_i})^2 (f_j^l - \mu_{f_j})^2}}$$

où f_i^l est le score associé à l'objet l par la fonction d'évaluation f_i, et μ_{f_i} est la moyenne des évaluations par la fonction f_i (voir figure 1).

ou bien utiliser le *coefficient de corrélation de rang de Spearman* s'appliquant sur les tris :

$$S(f_i, f_j) \;=\; 1 - 6 \sum_l \frac{(f_i^l - f_j^l)^2}{m(m^2-1)}$$

ici f_i^l désigne le rang assigné à l'objet l dans $\mathcal{S}$ par la fonction d'évaluation f_i.

Il est clair que pour être sûr d'avoir des fonctions d'évaluation faibles dans $\mathcal{F}$, il suffit de prendre les fonctions initiales f_i et d'ajouter dans $\mathcal{F}$ leur opposée, qui induit donc l'ordre inverse sur les objets de $\mathcal{S}$.

Il reste alors à identifier des fonctions d'évaluation faibles présentes dans $\mathcal{F}$ et à combiner leurs résultats pour obtenir un classement final sur les objets de $\mathcal{S}$.

2.2 Comment trouver des fonctions d'évaluation faibles

Les fonctions d'évaluation que nous cherchons doivent avoir tendance à placer les exemples de la classe '+' en tête du classement qu'elles induisent sur $\mathcal{S}$. Ainsi, ces fonctions d'évaluation sont-elles positivement corrélées entre elles sur le haut de leur classement des objets de $\mathcal{S}$. Il ne suffit cependant pas de sélectionner toutes les fonctions ainsi corrélées, car elles pourraient être corrélées indépendamment de l'échantillon d'objets, simplement car elles mesurent le même type de régularités dans les objets et

ont donc tendance à être d'accord entre elles. À la limite, une fonction d'évaluation f_i et sa copie f'_i seront parfaitement corrélées sur $\mathcal{S}$, sans que cela ne révèle quelque chose sur $\mathcal{S}$ et la présence ou non d'objets de la classe '+' dans $\mathcal{S}$.

On ne retiendra donc que les fonctions d'évaluation positivement corrélées sur l'échantillon $\mathcal{S}$, et particulièrement sur le haut du classement, et décorrélées en général, c'est-à-dire sur des échantillons $\mathcal{S}_0$ quelconques (nous utilisons la notation $\mathcal{S}_0$ en référence à l'hypothèse nulle en statistique).

2.3 Combiner les résultats des fonctions d'évaluation faibles

Chaque fonction d'évaluation retenue produit un vecteur de scores associés aux objets $\mathbf{x}$ de $\mathcal{S}$ et induit un tri de $\mathcal{S}$. Plusieurs méthodes peuvent être utilisées pour combiner ces "sorties" et obtenir un classement final, voire une partition entre objets supposés de la classe '+' et ceux supposés de la classe '−'.

Parce que la répartition des scores dépend de paramètres propres aux fonctions d'évaluation qui ne sont pas nécessairement significatifs, l'approche la plus usitée pour combiner consiste à considérer les tris et non les scores. Dans ce cadre, la manière la plus directe d'opérer est d'utiliser le rang moyen obtenu par les objets pour calculer leur rang dans le tri final.

On pourrait cependant attribuer un poids à chaque fonction d'évaluation retenue et tenir compte de ce poids dans le calcul du tri final. Nous revenons sur cette question dans la section 6.

Dans la suite, nous détaillons la détection de fonctions d'évaluation faible en section 4 et les approches pour combiner leurs résultats en section 6. Nous commençons pas décrire la mesure de corrélation utilisée pour comparer les fonctions d'évaluation.

3 Mesurer les corrélations

Une mesure de corrélation entre fonctions d'évaluation mesure à quel point une information sur la valeur ou sur le classement d'un exemple par l'une des fonctions d'évaluation fournit une information sur la valeur ou sur le classement de cet exemple par l'autre fonction d'évaluation. Dans le cas des classements, les mesures les plus utilisées sont le coefficient de corrélation de rang de Kendall et le coefficient de corrélation de rang de Spearman. Dans le contexte de la recherche d'information, le *Discounted cumulative gain* (DCG) et sa version normalisée (NDCG) sont aussi très employés (Järvelin et Kekäläinen, 2002) (voir (Wang et al., 2013) pour une étude théorique de ses propriétés). L'avantage de NDCG est de pondérer la mesure de corrélation en prenant en compte le rang des objets triés et en favorisant les objets placés en haut du classement.

Cependant, lorsque l'on suppose l'existence de deux classes d'objets, le classement relatif des objets d'une classe n'a pas d'importance : tous les objets '+' (resp. '−') se valent. C'est pourquoi nous introduisons une autre mesure de corrélation proche de l'*indice de Jaccard* pour la comparaison de sous-ensembles d'éléments.

3.1 La mesure de corrélation utilisée

Dans la suite, nous appellerons top_n^i les n exemples de $\mathcal{S}$ les mieux classés par la fonction d'évaluation f_i.

Nous noterons $\cap_n^{i,j}$ l'intersection des top_n obtenus par deux fonctions d'évaluation f_i et f_j : $\cap_n^{i,j} = top_n^i \cap top_n^j$. Ainsi si $top_5^i = \{a, b, c, d, e\}$ et $top_5^j = \{g, a, f, e, d\}$, alors $top_5^{i,j} = \{a, d, e\}$.

Nous proposons ici une nouvelle mesure de corrélation s'appuyant sur la comparaison des top_n successifs des deux classements considérés. Précisément, la corrélation entre deux classements d'un ensemble par deux fonctions d'évaluation f_i et f_j sera caractérisée par la courbe $| \cap_n^{i,j} |$ pour $1 \leq n \leq m$ si $\text{Card}(\mathcal{S}) = m$.

Cette mesure est inspirée de la *loi hypergéométrique* qui donne la loi probabiliste de la taille de l'intersection de deux tirages indépendants sans remise. On suppose qu'un tirage sans remise dans une urne de taille m tire n boules et les marque par une étiquette, par exemple « gagnant ». Les boules sont remises dans l'urne et un deuxième tirage de n boules a lieu. La taille de l'intersection des deux tirages est alors le nombre de boules marquées « gagnant » qui ont été tirées dans le second tirage. La loi hypergéométrique prédit :

$$\mathbf{p}(| \cap_n^{i,j} | = k) \;=\; \frac{\binom{n}{k} \cdot \binom{m-n}{n-k}}{\binom{m}{n}}$$

Par exemple, deux tirages aléatoires et indépendants de 500 éléments parmi 6000 ont une probabilité maximale d'avoir 42 éléments en commun. On remarquera que la valeur la plus probable vérifie $k/n = n/m$ (e.g. $42/500 \approx 500/6000$).

Si les résultats observés s'éloignent de ceux prédits par la loi hypergéométrique, les deux tirages ne sont probablement pas indépendants. Dans un cas extrême, l'un des tirages est une copie de l'autre, et la taille de l'intersection est donnée par : $| \cap_n^{i,j} | = n, (\forall n \leq m)$. Dans le cas extrême inverse, le deuxième tirage tire autant que possible des boules non tirées par le premier. Ce cas est analogue à celui de deux méthodes de tri qui rangent les éléments en sens inverse. La loi de la taille d'intersection est alors donnée par une taille nulle jusqu'à $n = m/2$ puis par une croissance selon $\frac{2(n-(m/2))}{n}$ (voir figure 2). Il existe ainsi tout un spectre de comportements possibles de la loi de l'intersection selon le degré de corrélation des tirages.

3.2 Utilisation de la mesure de corrélation

Le principe de la méthode proposée consiste à mesurer la différence de comportement de la taille de l'intersection $| \cap_n^{i,j} |$ pour toute paire de fonctions d'évaluation sur l'échantillon d'intérêt $\mathcal{S}$ et sur des échantillons supposés quelconques $\mathcal{S}_0$. En effet, si les deux fonctions f_i et f_j sont nettement plus corrélées lorsqu'elles trient $\mathcal{S}$ que lorsqu'elles trient les $\mathcal{S}_0$, c'est que d'une certaine manière, elles détectent un « signal » dans $\mathcal{S}$ qui n'est pas généralement présent. En particulier, elles ont tendance à mettre les mêmes exemples en tête de leurs classements.

La figure 3 est typique d'une *surcorrélation* mesurée entre deux fonctions d'évaluation, ici ANOVA et RELIEF, sur des données correspondant à 6400 gènes dont l'activité

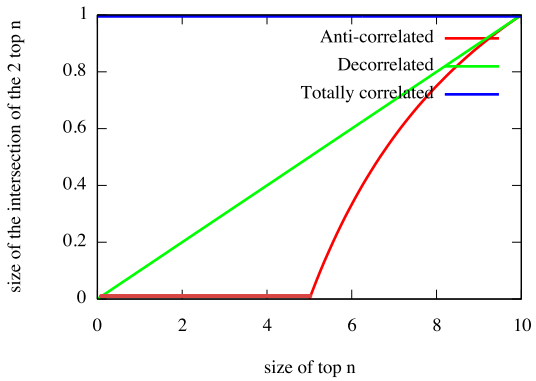

FIG. 2 – *Courbe* $|\cap_n^{i,j}|/n$ *en fonction de* n*. Deux tirages indépendants produisent en probabilité la diagonale. Deux tirages parfaitement corrélés sont tels que* $|\cap_n^{i,j}|/n = 1 \ \ (\forall n)$*. Deux tirages anticorrélés produisent la courbe basse (en rouge). Tous les comportements possibles s'inscrivent entre ces deux courbes extrêmes.*

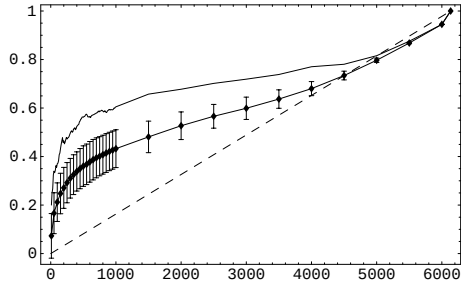

FIG. 3 – *Courbes de corrélation mesurées sur l'échantillon d'intérêt (courbe supérieure) et sur des échantillons quelconques (courbe avec l'écart-type indiqué par des barres).*

a été mesurée selon deux conditions. La courbe supérieure correspond à $|\cap_n^{i,j}|$ sur les données réelles, tandis que la courbe avec des intervalles de confiance correspondant à un écart-type, correspond à la moyenne de $|\cap_n^{i,j}|$ lorsque des échantillons $\mathcal{S}_0$ sont considérés (ici 100 échantillons $\mathcal{S}_0$).

En fonction de la force du signal détecté, la surcorrélation peut être plus ou moins marquée. De plus, la courbe de surcorrélation peut présenter un pic ou un maximum relatif qui peut être indicatif du nombre d'exemples '+' dans l'échantillon d'intérêt $\mathcal{S}$. C'est ce que montre la figure 4.

3.3 Étude théorique pour la combinaison de deux fonctions d'évaluation

On cherche ici à voir si un modèle théorique simple permet de rendre compte qualitativement des comportements mesurés empiriquement.

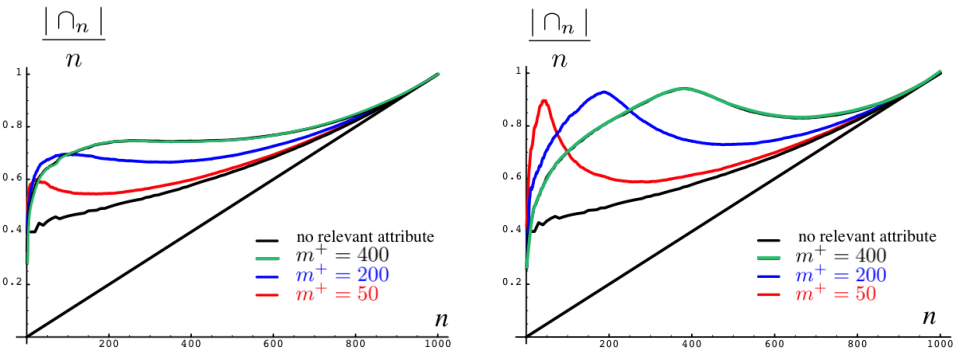

FIG. 4 – *Courbes de corrélation entre paires de tris obtenues pour des données artificielles pour diverses valeurs du nombre d'exemples '+' (50, 200 et 400 sur 1000 exemples au total), avec surcorrélation plus forte et des pics plus marqués à droite.*

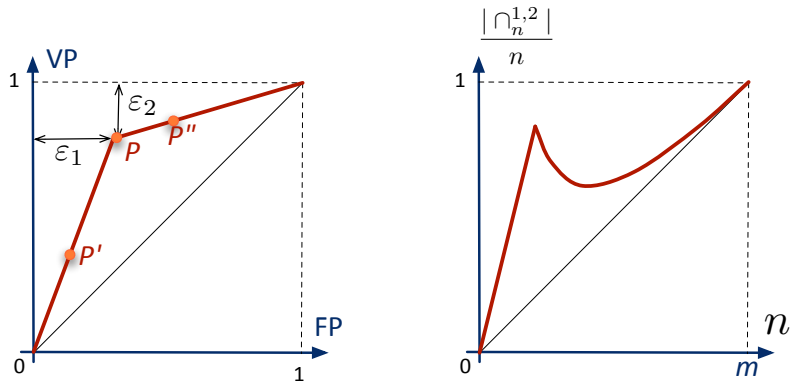

FIG. 5 – *(À gauche) Modèle de courbe ROC des fonctions d'évaluation utilisées dans l'étude théorique. (À droite) Courbe résultante de corrélation la plus probable.*

On suppose deux fonctions d'évaluation f_1 et f_2 définies de $\mathcal{X} \to \mathbb{R}$ caractérisées par une courbe ROC rendant compte de leur propension à ranger les éléments '+' avant les éléments '−'.

La courbe ROC considérée (voir figure 5) est la plus simple qui puisse rendre compte d'un comportement fréquent avec une assez bonne sélection des exemples '+' en haut du classement suivie d'un régime dégradé (voir Flach (2012)).

On suppose de plus ici que *les deux fonctions* considérées ont une *courbe ROC de même caractéristique* et que les deux fonctions sont *décorrélées dans leur classement étant donnée une classe d'exemples* ('+' ou '−'). Une étude plus poussée avec des fonctions de caractéristiques un peu différentes ne modifie pas qualitativement les résultats.

Afin de calculer la courbe de corrélation des fonctions f_1 et f_2, il faut considérer trois régimes, correspondant respectivement à (i) un point P' se déplaçant sur la droite

pour $x < \varepsilon_1$, (ii) au point P pour lequel le taux de faux positifs vaut ε_1 et le taux de vrais positifs vaut $1-\varepsilon_2$, et (iii) à un point P'' se déplaçant sur la droite quand $x > \varepsilon_1$.

On note m^- le nombre d'exemples de la classe '$-$' et m^+ le nombre d'exemples de la classe '$+$' contenus dans l'échantillon $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$

Pour ces trois régimes, on obtient les parties de courbes de corrélation les plus probables [1] suivantes (x désigne le taux de faux positifs et varie de 0 à 1) :

1. Pour $x < \varepsilon_1$.

   Après calculs, on trouve le nombre n d'éléments classés en haut (par f_1 et par f_2) faisant ressortir la proportion d'exemples '$+$' et d'exemples '$-$', et la taille de l'intersection des deux classements de f_1 et f_2 en fonction de x comme :

$$\left\{ \begin{array}{rcl} n & = & x\, m^- + \frac{1-\varepsilon_2}{\varepsilon_1}\, x\, m^+ \\ | \cap_n^{1,2} | & = & x^2\, m^- + \left(\frac{1-\varepsilon_2}{\varepsilon_1}\right)^2 x^2\, m^+ \end{array} \right. \tag{1}$$

   (La deuxième égalité utilise l'hypothèse que les fonctions f_1 et f_2 sont décorrélées entre elles étant donnée la classe. La taille de l'intersection dans les '$+$' et les '$-$' obéit donc à la loi hypergéométrique : par exemple, $|\cap_{n^-}{}^{1,2}| = \frac{(n^-)^2}{m^-}$ pour deux tirages de n^- exemples négatifs parmi m^-. D'où ici, pour les exemples négatifs : $|\cap_{n^-}{}^{1,2}| = (\frac{n^-}{m^-})^2 m^- = x^2\, m^-$. Le même raisonnement vaut pour les exemples positifs.)

   d'où la portion de courbe de corrélation d'équation :

$$\frac{| \cap_n^{1,2} |}{n} = \frac{x^2\, m^- + \left(\frac{1-\varepsilon_2}{\varepsilon_1}\right)^2 x^2\, m^+}{x\, m^- + \frac{1-\varepsilon_2}{\varepsilon_1}\, x\, m^+} = x\, \frac{m^- + \left(\frac{1-\varepsilon_2}{\varepsilon_1}\right)^2 m^+}{m^- + \frac{1-\varepsilon_2}{\varepsilon_1}\, m^+} \tag{2}$$

2. Pour $x = \varepsilon_1$ (point P).

$$\left\{ \begin{array}{rcl} n & = & \varepsilon_1\, m^- + (1-\varepsilon_2)\, m^+ \\ | \cap_n^{1,2} | & = & \varepsilon_1^2\, m^- + (1-\varepsilon_2)^2\, m^+ \end{array} \right. \tag{3}$$

   qui donne le point :

$$\frac{| \cap_n^{1,2} |}{n} = \frac{\varepsilon_1^2\, m^- + (1-\varepsilon_2)^2\, m^+}{\varepsilon_1\, m^- + (1-\varepsilon_2)\, m^+} \tag{4}$$

3. Pour $\varepsilon_1 < x$.

$$\left\{ \begin{array}{rcl} n & = & x\, m^- + \left[(1-\varepsilon_2) + \frac{\varepsilon_2}{1-\varepsilon_1}(x-\varepsilon_1)\right] m^+ \\ | \cap_n^{1,2} | & = & x^2\, m^- + \left[(1-\varepsilon_2) + \frac{\varepsilon_2}{1-\varepsilon_1}(x-\varepsilon_1)\right]^2 m^+ \end{array} \right. \tag{5}$$

   d'où l'équation de la courbe de corrélation pour cet intervalle :

$$\frac{| \cap_n^{1,2} |}{n} = \frac{x^2\, m^- + \left[(1-\varepsilon_2) + \frac{\varepsilon_2}{1-\varepsilon_1}(x-\varepsilon_1)\right]^2 m^+}{x\, m^- + \left[(1-\varepsilon_2) + \frac{\varepsilon_2}{1-\varepsilon_1}(x-\varepsilon_1)\right] m^+} \tag{6}$$

1. Les calculs font appel à la loi hypergéométrique qui donne la taille de l'intersection la plus probable.

Ces équations donnent la courbe de corrélation $\frac{|\cap_n^{1,2}|}{n}$ la plus probable telle qu'apparaissant dans la figure 5 à droite. On constate le bon accord avec les courbes obtenues expérimentalement, malgré la simplicité du modèle.

4 La détection de fonctions utiles

La sélection de fonctions de base utiles s'opère selon l'algorithme 1. On trie d'abord les fonctions de $\mathcal{F}$ par surcorrélation décroissante, celle-ci étant la différence de corrélation mesurée sur l'échantillon de données $\mathcal{S}$ et la corrélation moyenne mesurée sur les échantillons aléatoires $\mathcal{S}_0$. On retient dans $\mathcal{F}'$ les fonctions dépassant un certain seuil de surrcorrélation avec au moins une autre fonction.

Il est ensuite important de ne conserver autant que possible que des fonctions d'évaluation décorrélées entre elles. Cela peut se mesurer en examinant la surcorrélation de chaque fonction avec toutes les autres fonctions de $\mathcal{F}'$. Il faut que cette surcorrélation dépasse un seuil minimal. On devrait ne sélectionner que les fonctions de surcorrélation minimale avec chacune des autres fonctions qui seront finalement dans $\mathcal{F}''$. Cela nécessiterait cependant un algorithme assez lourd de satisfaction de contraintes. L'implémentation actuelle simplifie le problème en retenant les fonctions $f_i \in \mathcal{F}'$ dont la *somme* des surcorrélations avec les autres fonctions de $\mathcal{F}'$ dépasse un seuil minimal.

Algorithme 1: Sélection de fonctions de base pertinentes

Entrées : La base d'exemples $\mathcal{S}$
L'ensemble $\mathcal{F}$ de fonctions d'évaluation de base
Sorties : Un sous-ensemble $\mathcal{F}'' \in \mathcal{F}$ de fonctions de base

Génération de N échantillons "aléatoires" $\mathcal{S}_0$;

pour tous les *couples de fonctions d'évaluation* $(f_i, f_j)_{(i \neq j)} \in \mathcal{F}$ **faire**
  Calculer la surcorrélation de (f_i, f_j) sur $\mathcal{S}$ par rapport à la corrélation moyenne sur les échantillons $\mathcal{S}_0$
fin pour tous

Sélectionner les fonctions d'évaluation $f_i \in \mathcal{F}$ de surcorrélation $\geq$ seuil_min_surcor : soit $\mathcal{F}'$

Initialisation : $\mathcal{F}'' = \emptyset$

pour tous les $f_i \in \mathcal{F}'$ **faire**
  si $\sum_{j \neq i} surcorr(f_i, f_j) \geq seuil$ **alors**
    Mettre f_i dans $\mathcal{F}''$
fin pour tous

5 Expériences

Les expériences réalisées visent à tester la capacité de la méthode à sélectionner, dans l'ensemble $\mathcal{F}$, des fonctions d'évaluation positivement corrélées avec la fonction cible inconnue.

Pour ce faire, nous avons utilisé des données générées à l'aide d'un modèle génératif à deux distributions sur l'espace d'entrée $\mathcal{X} \in \mathbb{R}^d$: la distribution $\mathbf{P}^-_{\mathcal{X}}$ des exemples '$-$' et la distribution $\mathbf{P}^+_{\mathcal{X}}$ des exemples '$+$'. Un échantillon de données non supervisé $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ comprenant m^+ exemples '$+$' et m^- exemples '$-$' est ainsi généré.

Puisque nous connaissons la classe de chaque exemple généré (mais pas la méthode), nous pouvons calculer la courbe ROC de chaque fonction d'évaluation $f_i \in \mathcal{F}$. Parmi ces fonctions, la moitié ont une corrélation positive avec la fonction cible et l'autre moitié une corrélation négative puisque nous considérons pour chaque fonction d'évaluation f_i son opposée $-f_i$. Une fonction d'évaluation aléatoire est ajoutée à $\mathcal{F}$. La corrélation d'une fonction f_i avec la fonction cible est mesurée par son aire sous la courbe ROC (AUC).

Pour chaque expérience, un échantillon $\mathcal{S}$ est engendré, ainsi que 100 échantillons aléatoires $\mathcal{S}_0$. Cer derniers sont obtenus par permutations dans les valeurs des colonnes de la matrice $\mathcal{S}$ afin de conserver la même distribution de valeurs pour chaque descripteur.

Les statistiques des résultats sont obtenues par répétition des expériences pour chaque jeu de paramètres étudié. Nous retenons ainsi :

1. l'AUC de la fonction $f_i \in \mathcal{F}$ d'AUC maximale, l'AUC de la fonction $f_j \in \mathcal{F}$ d'AUC minimale et l'AUC moyenne des fonctions de $\mathcal{F}$. Nous donnons également les écart-types mesurés.
2. les mêmes AUC dans l'ensemble $\mathcal{F}''$ des fonctions sélectionnées par la méthode. Nous pouvons ainsi contrôler que l'AUC minimale est > 0.5, c'est-à-dire que seules des fonctions positivement corrélées à la fonction cible sont sélectionnées. L'AUC moyenne permet de mesurer le gain d'AUC par rapport à l'AUC moyenne avant sélection (qui est quasiment égale à 0.5 (mais pas tout à fait en raison de la présence de la fonction d'évaluation aléatoire)).

Nous avons fait varier la difficulté de la tâche en bruitant plus ou moins les distributions $\mathbf{P}^-_{\mathcal{X}}$ et $\mathbf{P}^+_{\mathcal{X}}$. Concrètement, nous avons fait varier la variance de ces distributions gaussiennes.

L'importance du rapport $\frac{m^+}{m}$ des exemples '$+$' dans l'échantillon $\mathcal{S}$ a été testé avec les valeurs $40/320 = 1/8 \approx 12\%$, $80/320 = 1/4 = 25\%$, $120/320 = 3/8 \approx 37\%$ et $160/320 = 50\%$.

Les résultats rapportés dans le tableau 1 sont issues d'expériences répétées 10 fois pour chaque jeu de paramètres. Les écart-types étant très réduits, il n'était pas nécessaire de multiplier les expériences pour chaque jeu de valeurs. Ici, $m = 320$, le nombre de fonctions d'évaluation considéré est $|\mathcal{F}| = 49$ (24 fonctions d'évaluation et leurs opposées et une fonction d'évaluation aléatoire), la dimension de l'espace d'entrée $\mathcal{X}$ est $d = 20$.

La première chose à noter est que la pire fonction d'évaluation sélectionnée dans $\mathcal{F}''$ a toujours une AUC > 0.5. La méthode est donc capable d'éliminer les fonctions d'éva-

σ	$\frac{m^+}{m}$	Avant sélection		Après sélection			
		auc_m	auc^M	auc_m	auc^M	$\overline{auc}$	AUC comb
1.5	$\frac{40}{320}$	0 ± 0	1 ± 0	$\mathbf{0.92} \pm 0.03$	1 ± 0	0.98 ± 0.01	$\mathbf{1} \pm 0$
	$\frac{80}{320}$	0 ± 0	1 ± 0	$\mathbf{0.87} \pm 0.06$	1 ± 0	0.97 ± 0.01	$\mathbf{1} \pm 0$
	$\frac{120}{320}$	0 ± 0	1 ± 0	$\mathbf{0.84} \pm 0.07$	1 ± 0	0.95 ± 0.01	$\mathbf{1} \pm 0$
2.5	$\frac{40}{320}$	0.02 ± 0.01	0.98 ± 0.01	$\mathbf{0.94} \pm 0.03$	0.98 ± 0.00	0.96 ± 0.02	$\mathbf{0.98} \pm 0.01$
	$\frac{80}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.85} \pm 0.05$	0.98 ± 0.01	0.91 ± 0.02	$\mathbf{0.97} \pm 0.01$
	$\frac{120}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.76} \pm 0.03$	0.98 ± 0.01	0.88 ± 0.02	$\mathbf{0.97} \pm 0.01$
	$\frac{160}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.73} \pm 0.04$	0.97 ± 0.01	0.85 ± 0.02	$\mathbf{0.95} \pm 0.01$
3.5	$\frac{40}{320}$	0.09 ± 0.02	0.91 ± 0.02	$\mathbf{0.75} \pm 0.06$	0.90 ± 0.03	0.83 ± 0.01	$\mathbf{0.90} \pm 0.03$
	$\frac{80}{320}$	0.09 ± 0.02	0.92 ± 0.02	$\mathbf{0.65} \pm 0.05$	0.92 ± 0.02	0.79 ± 0.02	$\mathbf{0.90} \pm 0.02$
	$\frac{120}{320}$	0.09 ± 0.02	0.91 ± 0.01	$\mathbf{0.64} \pm 0.04$	0.91 ± 0.01	0.77 ± 0.02	$\mathbf{0.89} \pm 0.02$
	$\frac{160}{320}$	0.10 ± 0.01	0.91 ± 0.02	$\mathbf{0.63} \pm 0.03$	0.91 ± 0.02	0.76 ± 0.02	$\mathbf{0.88} \pm 0.02$
4.5	$\frac{40}{320}$	0.13 ± 0.02	0.86 ± 0.02	$\mathbf{0.67} \pm 0.03$	0.86 ± 0.02	0.76 ± 0.02	$\mathbf{0.86} \pm 0.02$
	$\frac{80}{320}$	0.15 ± 0.02	0.85 ± 0.02	$\mathbf{0.65} \pm 0.03$	0.84 ± 0.03	0.75 ± 0.02	$\mathbf{0.84} \pm 0.03$
	$\frac{120}{320}$	0.15 ± 0.02	0.84 ± 0.02	$\mathbf{0.62} \pm 0.06$	0.84 ± 0.02	0.73 ± 0.03	$\mathbf{0.84} \pm 0.02$
	$\frac{160}{320}$	0.15 ± 0.01	0.85 ± 0.01	$\mathbf{0.61} \pm 0.03$	0.85 ± 0.01	0.72 ± 0.02	$\mathbf{0.83} \pm 0.03$

TAB. 1 – *Résultats expérimentaux en fonction des paramètres σ et proportion de la classe '+'. Notations : auc_m est l'AUC minimale, auc^M est l'AUC maximale, $\overline{auc}$ est l'AUC moyenne, AUC comb est l'AUC obtenue après combinaison des fonctions sélectionnées (voir section 6).*

luation non positivement corrélées avec la fonction cible inconnue. Il arrive également que la meilleure fonction sélectionnée ne soit pas la meilleure fonction disponible dans l'ensemble initial $\mathcal{F}$. Cela signifie que cette meilleure fonction n'a pas été surcorrélée suffisamment avec une autre fonction de $\mathcal{F}$.

Quand la variance des distributions $\mathbf{P}^-_{\mathcal{X}}$ et $\mathbf{P}^+_{\mathcal{X}}$ augmente, et donc qu'il devient de plus en plus difficile de distinguer les exemples d'une classe par rapport à l'autre, l'AUC moyenne des fonctions sélectionnées dans $\mathcal{F}''$ diminue, mais nous verrons en section 6 que l'on peut en partie compenser cette perte par combinaison des évaluations des fonctions de $\mathcal{F}''$.

On observe que la proportion d'exemples de la classe '+' dans $\mathcal{S}$ modifie peu les résultats avec cependant des résultats un peu meilleurs pour des proportions plus faibles.

Par ailleurs, quand $\mathcal{F}$ contient 45 fonctions de base, comme c'est le cas dans les expériences rapportées ici, le nombre d'experts sélectionnés dans $\mathcal{F}''$ décroît de ≈ 10 pour les tâches faciles ($\sigma = 1.5$) à environ 5.5 pour les tâches plus difficiles ($\sigma = 4.5$). Le bruit dans les données réduit en effet le niveau de surcorrélation entre les fonctions de base, même si elles sont *a priori* bonnes pour la tâche cible.

Finalement, comme dernier test, nous n'avons mis dans $\mathcal{F}$ que des fonctions d'éva-

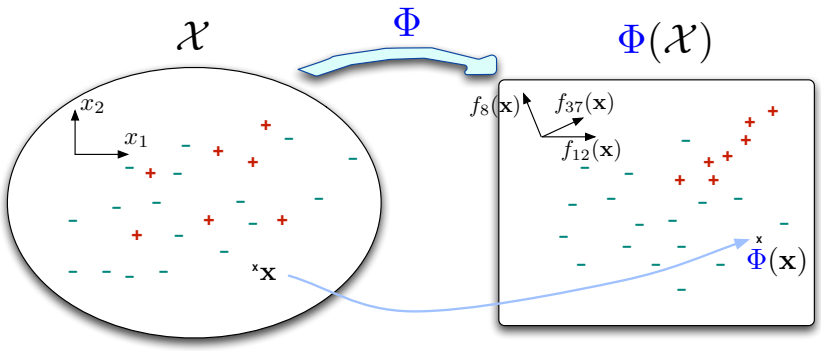

FIG. 6 – *Les points de $\mathcal{S}$ décrits dans l'espace d'entrée $\mathcal{X}$ sont projetés dans un espace $\Phi(\mathcal{X})$ dont les axes, les évaluations selon les fonctions d'évaluation, dépendent des caractéristiques de l'échantillon étudié $\mathcal{S}$.*

luation d'AUC < 0.5, c'est-à-dire négativement corrélées avec la fonction cible. Dans ce cas, dans environ 60% des expériences, le système sélectionne entre 2 et 4 fonctions de base, et la combinaison obtenue a alors une AUC < 0.5, c'est-à-dire que les exemples de la classe '$-$' ont tendance à être placés avant les exemples de la classe '$+$'. Il suffit cependant que 3 ou 4 fonctions de base soient positivement corrélées à la fonction cible pour que la sélection de fonctions négativement corrélées ne se produise plus.

6 La combinaison des résultats

Chacune des fonctions d'évaluation sélectionnées f_i dans $\mathcal{F}''$ produit une liste des exemples $\mathbf{x}$ de $\mathcal{S}$ triée par valeur décroissante de $f_i(\mathbf{x})$. Il est alors possible de décrire les exemples $\mathbf{x}$ dans un nouvel espace de description dont les axes sont les évaluations en fonction de chaque fonction d'évaluation $f_i \in \mathcal{F}''$ (voir figure 6). Dans ce nouvel espace, les coordonnées de chaque point $\mathbf{x}$ sont $(f_i(\mathbf{x}))_{f_i \in \mathcal{F}''}$.

Dans cet espace $\Phi(\mathcal{X})$, les points de la classe '$+$' ont tendance à être proches de la diagonale (car leurs classements par les f_i sont davantage corrélés) et à être éloignés de l'origine puisqu'ils ont des scores élevés.

Une méthode pour combiner les résultats des fonctions d'évaluation sélectionnées est donc de trier les exemples en mesurant la distance de leur projection sur la diagonale principale de l'espace $\Phi(\mathcal{X})$.

Cette méthode donne la même importance à toutes les fonctions f_i sélectionnées. Nous avons voulu donner plus de poids aux fonctions les plus surcorrélées aux autres fonctions de $\mathcal{F}''$ en pondérant les coordonnées par ce poids.

Nous utilisons donc une projection $\Phi(\mathcal{X})$ dans laquelle les coordonnées des points sont w_i $(f_i(\mathbf{x}))$, w_i dépendant de la surcorrélation de f_i avec les autres fonctions de $\mathcal{F}''$. Dans les expériences rapportées, w_i est une exponentielle de cette surcorrélation.

Les résultats obtenus avec cette stratégie de combinaison sont reportées dans la colonne de droite du tableau 1. Il est remarquable que malgré la difficulté croissante de distinguer les deux classes d'exemples, la méthode combinée reste très performante.

7 Conclusion et perspectives

À notre connaissance, c'est la première fois qu'une méthode d'ensemble en apprentissage non supervisée est présentée capable de fonctionner en partant d'une collection de fonctions d'évaluation de performances complètement inconnues.

Jusque là, les méthodes d'ensemble proposées dans le contexte du clustering présupposaient l'emploi de méthodes appropriées pour la tâche visée et cherchaient à en atténuer le biais et surtout la variance par rapport aux variations de l'échantillon.

Dans le cadre des expériences réalisées, notre méthode a montré sa capacité à identifier automatiquement des fonctions utiles dans un ensemble de fonctions quelconques. Par ailleurs, la combinaison des résultats de ces fonctions permet d'atteindre des performances proches de celle de la meilleure fonction d'évaluation inconnue *a priori*, et toujours bien meilleures que la moyenne des fonctions retenues, cela étant mesuré par les AUC des fonctions.

Si cette réalisation ouvre la possibilité d'utiliser des collections de fonctions d'évaluation sans chercher à les régler finement pour la tâche à résoudre, il reste cependant de nombreux points sur lesquels progresser.

D'abord, si le gain en AUC permis par la méthode est important, il ne suffit pas d'avoir une fonction de tri avec une bonne AUC pour avoir résolu le problème d'identifier deux classes d'objets. Il faut également déterminer un seuil de décision grâce auquel il devient possible de déterminer la classe d'un objet. L'observation des pics de surcorrélation (voir figures 3 et 4) offre une piste sérieuse pour la détermination de ce seuil, mais il reste cependant à préciser comment en déduire une valeur de seuil.

Pour cela, une étude théorique utilisant des modèles de fonctions d'évaluation (tel que le modèle à courbe ROC simplifié de la figure 5) sera certainement utile. Ce même type d'étude devrait également permettre de fonder rigoureusement une méthode de combinaison des résultats des fonctions d'évaluation en lieu et place de la méthode actuelle qui résulte d'une étude empirique et dont les résultats peuvent sans doute être encore améliorés.

La constitution et le rôle des échantillons aléatoires $\mathcal{S}_0$ restent aussi un point à clarifier. Ces échantillons servent à mesurer la corrélation *a priori* des fonctions de base, mais ils servent également de référence par rapport à laquelle sont contrastés les classes d'objets '+' et '−'. Les conditions permettant de garantir que c'est bien la classe '+', celle qui est recherchée, qui sera identifiée, et non la classe '−' doivent être précisées.

Remerciements

Nous tenons à signaler que ce travail a connu ses premiers développements lors d'une étude préalable avec Romaric Gaudel en 2007 avant d'entrer en hibernation pour plusieurs années. Merci à lui.

Références

Alizadeh, H., B. Minael-Bigdoli, et H. Parvin (2014). To improve the quality of cluster ensembles by selecting a subset of base clusters. *Journal of Experimental & Theoretical Artificial Intelligence 26*(1), 127–150.

Blum, A. et P. Langley (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence journal* (97), 245–271.

Cesa-Bianchi, N. et G. Lugosi (2006). *Prediction, learning and games*. Cambridge University Press.

Cornuéjols, A. et L. Miclet (2010). *Apprentissage Artificiel. Concepts et algorithmes (2nd Ed.)*. Eyrolles.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems*, pp. 1–15. Springer.

Dimitriadou, E., A. Weingessel, et K. Hornik (2001). Voting-merging : An ensemble method for clustering. In *Artificial Neural Networks—ICANN 2001*, pp. 217–224. Springer.

Domeniconi, C. et M. Al-Razgan (2009). Weighted cluster ensembles : Methods and analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD) 2*(4), 17.

Fern, X. Z. et W. Lin (2008). Cluster ensemble selection. *Statistical Analysis and Data Mining 1*(3), 128–141.

Flach, P. (2012). *Machine learning : the art and science of algorithms that make sense of data*. Cambridge University Press.

Freund, Y. et R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*(1), 119–139.

Guyon, I. et A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157–1182.

Järvelin, K. et J. Kekäläinen (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS) 20*(4), 422–446.

Kohavi, R. et G. John (1997). Wrappers for feature subset selection. *Artificial Intelligence journal*, 273–324.

Kononenko, I. (1994). Estimating attributes : analysis and extensions of relief. In *Machine Learning : ECML-94*, pp. 171–182. Springer.

Littlestone, N. et M. K. Warmuth (1994). The weighted majority algorithm. *Information and computation 108*(2), 212–261.

Press, W. H., S. Teukolsky, W. Vetterling, et B. Flannery (2007). *Numerical recipes 3rd edition : The art of scientific computing*. Cambridge university press.

Saeys, Y., T. Abeel, et Y. Van de Peer (2008). Robust feature selection using ensemble feature selection techniques. In *Machine learning and knowledge discovery in databases (ECML-PKDD-2008)*, pp. 313–325. Springer.

Schapire, R. E. et Y. Freund (2012). *Boosting : Foundations and Algorithms*. MIT Press.

Wang, Y., W. Liwei, Y. Li, D. He, W. Chen, et T.-Y. Liu (2013). A theoretical analysis of ndcg ranking measures. In *26th Annual Conference on Learning Theory*.

Zhou, Z.-H. (2012). *Ensemble methods : foundations and algorithms*. CRC Press.

Summary

In Machine Learning, ensemble or collaborative methods are based on the assumption that the performance of each expert or weak learner is measurable (in supervised learning) or can be estimated beforehand (unsupervised learning). This is deemed necessary in order to weight the expert's advices according to some confidence level.

In this paper, we present an unsupervised learning method applicable in the case of two unknown classes, which makes use of an ensemble of "experts" of which the performance on the task at hand is unknown. We show how to select, in the absence of any prior information, experts that are positively correlated with the unknown target decision function and how to combine their results in order to get a final decision function that is generally at least as good as the best unknown expert in the ensemble. Our empirical results on controlled experiments confirm that the method performs well.