Proposition d'une méthode de classification associative adaptative

Emna Bahri*, Stéphane Lallich*

*Laboratoire ERIC
5, avenue Pierre Mendès-France, 69500 Bron
emna.bahri | stéphane.lallich@univ-lyon2.fr,
http://eric.univ-lyon2.fr

Résumé. La classification associative est une méthode de prédiction à base de règles issue de la fouille de règles d'association. Cette méthode est particulièrement intéressante car elle recherche de façon exhaustive les règles d'association pertinentes qu'elle filtre pour ne garder que les règles d'association de classe (celles admettant pour conséquent une modalité de classe), qui sont utilisées comme classifieur. Les connaissances produites sont ainsi directement interprétables. Des études antérieures montrent les inconvénients de cette approche, qu'il s'agisse de la génération massive de règles non utilisées ou de la mauvaise prédiction de la classe minoritaire lorsque les classes sont déséquilibrées. Nous proposons une approche originale du type boosting de règles d'association de classes qui utilise comme classifieur faible une base de règles significatives construites par un algorithme de génération d'itemsets fréquents qui se limite à l'extraction des seules règles de classe significatives et qui prend en compte le déséquilibre des données. Des comparaisons avec d'autres méthodes de classification associative montrent que notre approche améliore la précision et le rappel.

1 Introduction

La classification associative est une méthode d'apprentissage supervisé à base de règles introduite par (Liu et al. (1998)), puis améliorée par différents auteurs, ainsi (Yin et Han (2003)) et (Li et al. (2001)). De façon générique, cette méthode utilise pour la prédiction un ensemble de règles d'association dites règles d'association de classe (au sens où leur conséquent est une modalité de classe) qui sont produites à partir des algorithmes de recherche d'itemsets fréquents. Le principal intérêt de la classification associative est qu'il s'agit d'une méthode dont le classifieur est un ensemble de règles, ce qui permet à l'expert de comprendre le processus de classification et de justifier auprès d'un tiers les résultats de la prédiction (par exemple dans le cas du *scoring* bancaire ou dans le domaine médical). Par rapport aux arbres de décision (Quinlan (1993)), qui constituent l'une des principales méthodes d'apprentissage supervisé à base de règles, la classification associative présente deux avantages. En premier lieu, elle procède à une exploration exhaustive des règles, et non pas à une exploration gloutonne. En second lieu, elle

obtient généralement de meilleurs résultats que C4.5, l'algorithme d'arbre de décision le plus couramment utilisé (voir les comparaisons sur 26 bases effectuées par (Yin et Han (2003)), (Li et al. (2001)) ou (Bahri et Lallich (2009b))).

Cependant, différents travaux (Li et al. (2001)) ont souligné les faiblesses de la classification associative. On retiendra principalement le fait que le temps d'exécution de ces méthodes et l'espace de stockage nécessaire se trouvent pénalisés par la génération massive d'itemsets qui ne sont pas tous utilisés car ils ne débouchent pas nécessairement sur une règle de classe. En outre, face à des données déséquilibrées, les performances de la classification associative sont amoindries, particulièrement celles relatives à la classe minoritaire.

Dans des travaux antérieurs, nous avons pris en compte ces problèmes pour élaborer W-CARP, une méthode de classification associative qui a été conçue pour fournir le classifieur de base d'une procédure de classification associative adaptative. W-CARP tient compte du déséquilibre de classe en recourant à un seuil de support local et ne génère que les itemsets indispensables, à savoir les itemsets de classe, ceux qui contiennent une modalité de classe. A partir de ces itemsets de classe, W-CARP construit les règles d'association de classe et filtre les règles significatives qui sont rassemblées dans la SRB, ou base de règles significatives, qui sert de classifieur.

C'est ainsi que dans ce papier, nous proposons CARBoost une procédure de classification associative adaptative qui utilise itérativement la SRB produite par W-CARP pour se concentrer tout à la fois sur les cas difficiles à prédire et sur les règles qui prédisent correctement les exemples à prédire.

Cet article sera structuré comme suit. La section 2 introduit la classification associative, ses méthodes avant de présenter W-CARP. La section 3 présente les méthodes ensemblistes. Notre approche CARBoost qui se base sur le boosting de règles d'association de classe, est présentée en section 4. Puis la section 5 rend compte des expériences pratiquées et analyse les résultats obtenus. Finalement, nous concluons en section 6.

2 La classification associative

La classification associative est une méthode d'apprentissage supervisé à base de règles introduite par (Liu et al. (1998)). Depuis lors, différents travaux ont mis en évidence les différents avantages de la classification associative par rapport aux arbres de décision, notamment la diminution du taux d'erreur et l'exploration exhaustive des règles, tout en conservant la lisibilité des résultats.

2.1 Notions de bases

La classification associative est fondée sur la prédiction de la classe à partir de règles d'association particulières, dites règles d'association de classe ou règles d'association prédictives. Une règle d'association de classe est une règle dont le conséquent doit être la variable indicatrice de l'une des modalités de la classe. Une telle règle s'écrit donc $A \rightarrow c_i$, où A est une conjonction de descripteurs booléens et c_i est la variable indicatrice de la i_e modalité de classe. L'intérêt des règles de classe est de permettre la focalisation sur des groupes d'individus, éventuellement très petits, homogènes du point de vue des descripteurs et présentant la même

classe. Par rapport aux arbres de décision, la classification associative a l'intérêt d'explorer les règles de façon exhaustive et non pas de pratiquer une stratégie gloutonne.

Les méthodes de classification associative procèdent en deux phases. La phase 1 correspond à la construction des règles d'association de classe. Elle se décompose habituellement en 2 étapes, l'une consacrée à l'extraction des itemsets fréquents à l'aide d'algorithmes tels que Apriori (Agrawal et Srikant (1994)) ou FP-Growth (Han et al. (2000)), l'autre filtrant les itemsets fréquents obtenus pour ne conserver que les itemsets comportant un item de classe (itemset de classe) et en déduire les règles d'association de classe satisfaisant le seuil de confiance préfixé. La phase 2 est dévolue à la prédiction à partir des règles de classe sélectionnées.

2.2 Principales méthodes de classification associative

Parmi les méthodes utilisées pour la classification associative, CBA (*Classification Based on Association*, (Liu et al. (1998))) est le premier algorithme proposé. Les deux phases précitées sont accomplies de la façon suivante :

- CBA-RG utilise Apriori pour générer toutes les règles d'association qui satisfont les seuils de support et de confiance choisis au départ ;
- CBA-CB est un algorithme heuristique qui assure la prédiction : pour chaque cas d'apprentissage on stocke la règle de plus forte confiance parmi les règles qui le couvrent. Les règles stockées sont classées par ordre de confiance décroissant et on applique à un nouveau cas la règle qui le couvre de plus forte confiance.

Yin et Han (2003) ont proposé une version améliorée de CBA appelée CPAR (*Classification based on Predictive Association Rules*) qui améliore le temps d'exécution tout en assurant une qualité de classification similaire à celle des méthodes existantes. Cette approche génère les règles à l'aide de l'algorithme PRM (*Predictive Rule Mining*) qui se base sur le principe de FOIL (*First Order Inductive Learner*,(Quinlan et Cameron-Jones (1995))). Pour la prédiction, CPAR évalue d'abord les règles de classe en estimant leur *expected accuracy* au moyen de la mesure de Laplace qui pénalise la confiance en fonction du nombre de classe. Ensuite, pour chaque cas à prédire, CPAR détermine les règles qui couvrent ce cas, choisit les K meilleures régles débouchant sur chaque classe, calcule *l'expected accuracy* moyenne et retient la classe ayant le meilleur résultat.

Une autre amélioration proposée est CMAR (*Classification based on Multiple Association Rules*,(Li et al. (2001))). CMAR utilise FP-Growth au lieu de Apriori. Il sélectionne les règles avec une confiance élevée et analyse la corrélation entre l'antécédent et le conséquent de ces règles. La mesure utilisée est le *weighted Chi-square* qui exprime la force d'une règle à partir de la condition de support et de la distribution de la classe. Dans un souci d'efficacité, CMAR emploie une nouvelle structuration de données, *CR-Tree*, pour enregistrer et sélectionner les règles et pour chercher rapidement l'antécédent de la règle.

Parmi les reproches le plus couramment faits aux méthodes usuelles de classification associative, on retiendra la génération massive d'itemsets ou de règles non utilisées et la difficulté à assurer une bonne prédiction de la classe minoritaire en cas de déséquilibre des classes. Pour répondre à ces reproches, nous avons proposé W-CARP (Bahri et Lallich (2009b)) qui améliore la phase d'extraction par l'extraction directe des seuls itemsets de classe et des règles de classe correspondantes grâce à FCP-Growth-P, puis construit une base de règles significatives. En outre, W-CARP améliore la phase de prédiction grâce à une pondération adéquate des règles significatives. Nous détaillons ci-dessous ces 3 améliorations.

- **Phase d'extraction, génération des seuls itemsets et règles de classe.** Pour générer les règles de classe, W-CARP utilise FCP-Growth-P, défini dans (Bahri et Lallich (2009a)). FCP-Growth-P est une adaptation de FP-Growth destinée à extraire seulement les itemsets de classe. Son principe de fonctionnement est similaire à FP-Growth mais avec des modifications de construction du FP-Tree et une procédure d'élagage différente. Pour améliorer la prédiction de la classe minoritaire, au lieu d'utiliser un seuil de support fixe comme dans la version de base de FP-Growth, FCP-Growth-P utilise un seuil de support local, adapté à chaque modalité c_i, qui dépend de n_i, le nombre de transactions validant la modalité de classe considérée. Pour la modalité c_i, on aura $\sigma_i = \sigma \times n_i/n$. En outre, nous avons introduit dans FCP-Growth-P la méthode d'élagage proposée par((Li, 2006)) dans le cadre d'Apriori. Celle-ci repose sur le fait que pour toute mesure m qui vérifie la condition "si $supp(Ax\overline{c}) = Supp(A\overline{c})$, alors $m(Ax \rightarrow c) \leq m(A \rightarrow c)$" (mesure optimale au sens de Li), toute spécialisation d'une règle qui ne diminue pas le nombre de contre-exemples est moins intéressante que la règle de départ.
- **Phase d'extraction, filtrage des règles significatives.** Pour assurer la prédiction, W-CARP constitue une base de règles significatives, notée SRB, qui rassemble les règles de classe dont la confiance est significativement supérieure à la probabilité *a priori* de la classe. Pour éviter l'inflation de fausses découvertes, on peut éventuellement ajuster les p-value (Lallich et al. (2006)) produites par le test de signification. Les règles retenues dans la SRB étant significatives, celle-ci reste pertinente pour les différents échantillons bootstrap issus de l'ensemble d'apprentissage. On remarquera que les règles obtenues sont de complexité très modérée. Par exemple, avec la base Breast, on trouve 34 règles de classe significatives, 2,1 items en moyenne dans l'antécédent et 85% des règles ayant au plus 2 items. Avec la base Waves, on obtient 342 règles de classe significatives, 2,2 items en moyenne et 70% des règles ayant au plus 2 items.
- **Phase de prédiction.** La procédure de prédiction de W-CARP est effectuée à partir des règles de la SRB en pondérant celles-ci par leur mesure de Loevinger, qui a l'intérêt de prendre en compte la distribution des classes :
  - étant donné un cas à prédire, on détermine d'abord les règles de la SRB qui couvrent ce cas ;
  - on pondère ces règles par la valeur de leur mesure de Loevinger, notée $\frac{P(c_j/A)-P(c_j)}{P(\overline{c}_j)}$ pour une règle $A \rightarrow c_j$. L'intérêt de cette mesure, est qu'elle avantage la classe minoritaire. En outre, elle est compatible avec la stratégie d'élagage de (Li (2006)), ce qui permet d'élaguer et de travailler avec des seuils de support très bas ;
  - pour chaque cas à prédire, on calcule le score de Loevinger de chaque classe, qui correspond à la somme des valeurs de Loevinger de toutes les règles couvrant le cas considéré qui débouchent sur la classe considérée ;
  - la classe prédite est celle qui maximise le score de Loevinger.

Pour générer les règles de classe, FCP-Growth-P a une complexité en O(nKR), sachant que n est le nombre de cas, K le nombre d'attributs et R le nombre de règles générées. La complexité de la phase de prédiction qui suit est de l'ordre de $0(R)$ puisque pour chaque cas à prédire on accède à la SRB et on parcourt au maximum toutes les règles. Comme montré dans Bahri et Lallich (2009b), W-CARP est plus rapide que les approches traditionnelles, tout en assurant une précision au moins égale à la précision de celles-ci, ce qui en fait un bon candidat pour servir de classifieur faible dans une procédure adaptative de type boosting.

3 Intérêt des méthodes ensemblistes

En contrepoint de leur intelligibilité, les méthodes à base de règles présentent souvent des performances quelque peu inférieures à celles des autres algorithmes. Afin d'augmenter les performances de la classification associative, tout en veillant à conserver une bonne part de son intelligibilité, nous nous sommes tournés vers les méthodes ensemblistes.

Le théorème du Jury de Condorcet (1785) a montré que le vote à la majorité de plusieurs juges qui se prononcent indépendamment sur une alternative avec un même risque de se tromper inférieur à 0,5 permettait de réduire considérablement le risque d'erreur du jury, jusqu'à être asymptotiquement nul.

Transposé en fouille des données, ce théorème signifie que l'agrégation de classifieurs (ou méthode ensembliste) permet de réduire considérablement le risque d'erreur à condition que les classifieurs soient suffisamment bons (risque d'erreur inférieur à 0.5) et suffisamment divers (au sens où leurs erreurs sont indépendantes). Toute la difficulté des approches ensemblistes est d'assurer cette diversité. Celle-ci peut provenir d'abord de l'utilisation de classifieurs hétérogènes. Sinon, en cas de relance d'un même algorithme, la diversité peut provenir notamment de la modification des paramètres de l'algorithme et/ou d'un aléa sur les individus par le biais d'échantillons bootstrap et/ou d'un aléa sur les attributs.

Plus formellement, les méthodes ensemblistes s'interprètent dans le cadre du compromis entre le biais des algorithmes d'apprentissage (erreur systématique qui ne dépend pas de l'échantillon d'apprentissage) et leur variance (qui provient de la variabilité des résultats issus de l'échantillon d'apprentissage). Le Stacking (Wolpert (1992)) construit un méta-modèle de décision qui a pour but de minimiser le biais, alors que le Bagging (Breiman (1996)) opère sur des échantillons boostrap de l'ensemble d'apprentissage pour réduire la variance sans trop augmenter le biais. (Freund et Schapire (1996)) s'efforcent de réduire simultanément le biais et la variance en travaillant sur des échantillons bootstrap de l'ensemble d'apprentissage et en forçant le classifieur à se concentrer sur la prédiction des cas difficiles à prédire grâce à une repondération adaptative des cas, au risque de sur-apprendre en cas de données bruitées. Les forêts aléatoires (Breiman (2001)) combinent la construction d'arbres non élagués sur des échantillons bootstrap de l'ensemble d'apprentissage, qui diminue le biais, et la sélection au hasard des attributs qui participent à l'éclatement de chaque noeud de chaque arbre de la forêt, ce qui améliore la diversité des arbres de la forêt. Les forêts aléatoires sont ainsi une méthode de référence, rapide, très compétitive et robuste face au bruit.

Pour notre part, nous avons choisi de mettre en oeuvre une procédure adaptative inspirée du boosting qui nous paraît bien correspondre à la classification associative, laquelle est capable de fournir des règles couvrant peu d'exemples mais avec une très grande confiance. La procédure adaptive doit permettre à ces règles d'être valorisées.

A notre connaissance, peu de travaux ont appliqué les méthodes ensemblistes à la classification associative. On citera d'abord (Sun et al. (2006)) qui en appliquant Adaboost à des règles prédictives simples obtiennent de meilleurs résultats qu'en utilisant des règles complexes. (Yoon et Lee (2008)) utilisent une approche équivalente au boosting pour filtrer les règles d'association et s'adapter ainsi à la catégorisation de textes à grande échelle. Enfin, dans le cadre des concepts formels, (Meddouri et Maddouri (2009)) construit à chaque itération un concept pertinent pour les exemples de l'itération, qui servira de classifieur faible lors du boosting.

4 Pour une classification associative adaptative : CARBoost

4.1 Principes de CARBoost

Le but poursuivi est la conception d'une méthode de classification associative qui bénéficie de l'apport des méthodes ensemblistes pour améliorer ses performances de classification tout en gardant au mieux la lisibilité de son principe de prédiction à base de règles. Cette méthode itérative, nommée CARBoost repose sur les principes suivants :

- Le classifieur faible retenu est la base de règles significatives SRB produite par W-CARP (section 2.2). W-CARP produit une base de règles significatives qui permet de classer les cas avec des performances au moins aussi bonnes que CBA, CMAR et CPAR pour un temps d'exécution réduit.

- Nous avons choisi une procédure adaptative inspirée du boosting qui tient compte des erreurs de chaque itération pour forcer l'algorithme à se concentrer sur les cas difficiles à prédire. A chaque itération, on travaille sur un échantillon bootstrap des cas où le poids des cas mal classés à l'itération précédente est accru.

- La SRB est construite une fois pour toutes, ce qui contribue à réduire la complexité de la méthode. Cependant, pour introduire de la diversité et favoriser la prédiction des exemples difficiles, à chaque itération, les règles sont repondérées de telle sorte que l'on favorise les règles qui ont prédit correctement au moins un exemple mal classé de l'itération courante.

- La prédiction finale qui résulte du vote pondéré des classifieurs de base, permet d'échapper au sur-ajustement.

Les modifications d'une relance à l'autre sont donc dues à l'aléa sur les cas qui est issu de l'échantillon bootstrap, à la repondération des cas mal classés et à la repondération des règles efficaces sur au moins un exemple mal classé. Pour un nouvel exemple, d'une itération à l'autre, les règles qui le couvrent restent stables, de même que les classes qui sont les conséquents de ces règles, ce sont les pondérations des règles qui changent ainsi que le score de Loevinger associé à chaque classe pour l'exemple considéré. En effet, le bootstrap sur les individus, associé à l'évolution adaptative du poids de ceux-ci, modifie la mesure de Loevinger des règles. D'autre part, le poids des règles est modifié de façon adaptative pour favoriser les règles qui prédisent correctement au moins un exemple mal classé. La prédiction finale d'un nouvel exemple qui est obtenue par un vote pondéré des résultats des différentes itérations, se présente comme l'ensemble des liste de règles de la SRB couvrant l'exemple qui débouchent sur chacune des classes, chaque liste étant accompagnée d'un ensemble de poids optimisés qui résulte des repondérations précitées.

4.2 Explication détaillée de CARBoost

1. Inputs
   - Base de cas
   - Base de règles significatives obtenue à l'aide W-CARP. On génère les règles de classe à l'aide de FCP-Growth-P, ce qui permet de ne générer que les règles de classe tout en utilisant la condition d'élagage de Li. On filtre les règles obtenues pour ne garder que celles dont la confiance est significativement supérieure à la probabilité a priori de la classe qui figure en conséquent de la règle, constituant ainsi la SRB. Pour assurer la prédiction d'un nouveau cas, on pondère les règles qui couvrent ce cas avec la valeur

de la mesure de Loevinger et on calcule le score de chaque classe. La classe prédite pour le nouveau cas est celle qui a le meilleur score de Loevinger.

2. Initialisation : on considère m cas qui ont tous le même poids $W_0 = 1/m$
3. Itération
   Soit T le nombre d'itérations
   Pour $t = 1, 2, ...T$
   - Construction d'un échantillon bootstrap des cas
   - Appel à SRB et application des règles pertinentes pour l'échantillon
   - Calcul du taux d'erreur ϵ_t en divisant le nombre d'erreurs de prédiction obtenues à l'aide du score de Loevinger par le nombre de cas.
   - Calcul de l'exactitude $\alpha_t = 0.5 * Ln((1 - \epsilon_t)/\epsilon_t)$
   - Repondération des cas : on augmente le poids des cas mal prédits à l'itération t par le score de Loevinger ;
     - si le cas est mal prédit par le score de Loevinger, alors $W_{t+1} = W_t \exp^{\alpha_t}$.
     - si le cas est bien prédit par le score de Loevinger, alors $W_{t+1} = W_t \exp^{-\alpha_t}$.
   - Repondération des règles : on augmente le poids des règles qui classifient bien au moins un exemple mal classé à l'itération t. Si l'on désigne par $\epsilon(r, t)$ le taux d'erreur de la règle r à l'itération t, le poids de la règle r à l'itération t devient alors $\alpha_{r,t} = (1 - \epsilon(r, t))/\epsilon(r, t)$.
4. Prédiction : La classe d'un nouveau cas est prédite en faisant voter les classifieurs faibles pondérés par les α_t

5 Experiences et analyse des résultats

Pour évaluer l'efficacité de CARBoost, la nouvelle approche de classification associative que nous proposons, nous avons comparé expérimentalement sa précision en généralisation avec celle issue des principales méthodes de classification associative, à savoir CMAR, CPAR et W-CARP dont on sait qu'elles ont de meilleurs résultats que CBA. En outre, pour servir de référence, nous avons inclus dans la comparaison deux autres méthodes à base de règles, à savoir C4.5, l'algorithme à base d'arbres de décision le plus populaire et RF, les forêts aléatoires (Breiman (2001)), qui sont considérées comme l'une des méthodes ensemblistes à base d'arbres de décision les plus performantes.

Cette comparaison porte sur les mêmes 26 bases de données de l'UCI Machine Learning Repository (Asuncion et Newman (2007)) que celles utilisées dans (Li et al. (2001)). Toutes les expériences sont exécutées sur une machine de 2.8 GHz Pentium-4 avec 1GO de mémoire. Nous avons choisi pour ces experiences de pratiquer T= 10 itérations. En effet, c'est la valeur la plus couramment utilisée dans les procédures de type *Boosting-like*. C'est aussi celle choisie par Sun et al. (2006). La comparaison est fondée sur le calcul de la précision et du rappel de chaque méthode sur chaque base. Pour une méthode donnée sur une base donnée, rappelons que la précision d'une classe est la proportion de prédictions correctes parmi les cas prédits comme étant de cette classe. Le rappel d'une classe est la proportion parmi les cas de la classe de ceux qui sont correctement prédits. La précision et le rappel sur la base considérée sont alors les moyennes arithmétiques respectives des précisions et des rappels des différentes classes. Nous avons choisi de calculer la précision plutôt que l'*accuracy* afin de mieux prendre en

compte de mauvais résultats éventuels sur les classes minoritaires. En effet, la précision étant la moyenne arithmétique des précisions des différentes classes, elle fait intervenir les résultats des classes minoritaires au même titre que les résultats des classes plus nombreuses, ce qui n'est pas le cas de l'*accuracy*.

Le tableau 1 indique d'abord les caractéristiques des différentes bases : nombre de cas (NCas), nombre d'attributs (NAtt), nombre de classes (NCl). Il se poursuit par les précisions moyennes des 6 méthodes expérimentées, sur chacune des 26 bases. Les taux de rappel moyens des différentes méthodes sur chacune des 26 bases figurent dans le tableau 2.

Bases	**NCas**	**NAtt**	**NCl**	**C4.5**	**CPAR**	**CMAR**	**RF**	**WCARP**	**CARBoost**
ANNEAL	898	38	6	92.8	95.2	96.1	98.8	98.5	99.5
AUSTRAL	690	14	2	81.5	85.6	85.1	86.7	86.3	88.2
AUTO	205	25	7	72.1	77.4	74.3	82.1	79.1	82.1
BREAST	699	10	2	89.4	94.8	95.1	97.3	96.5	98.2
CLEVE	303	13	2	74.7	80.1	80.9	83.5	82.2	86.7
CRX	690	15	2	80.4	85.1	83.9	85.9	85.2	86.8
DIABETES	768	8	2	69.2	74.9	75.3	75.7	75.4	81.9
GERMAN	1000	20	2	67.3	72.4	72.7	73.8	74.6	83.9
GLASS	214	9	7	63.7	71.6	68.9	78.8	74.0	80.2
HEART	270	13	2	71.8	77.9	77.8	82.9	82.3	83.9
HEPATIC	155	19	2	71.6	75.5	78.9	83.7	79.6	85.9
HORSE	386	22	2	73.2	80.1	79.1	84.5	83.5	85.7
HYPO	3163	25	2	95.2	97.3	97.6	98.2	98.5	95.6
IONO	351	34	2	87.2	90.4	89.9	92.5	92.5	94.6
IRIS	150	4	3	90.3	94.7	92.3	95.5	94.6	95.4
LABO	57	16	2	79.5	80.2	85.2	87.5	85.8	88.9
LED7	3200	7	10	68.5	71.2	70.2	74.5	72.6	73.9
LYMPH	148	18	4	73.5	69.8	80.2	83.8	81.5	85.4
PIMA	768	8	2	68.5	72.1	71.6	75.8	74.1	75.8
SICK	2800	29	2	90.5	92.4	94.1	98.2	97.8	97.8
SONAR	208	60	2	73.2	75.8	76.1	82.0	79.3	83.6
TIC-TAC	958	9	2	92.4	95.4	96.3	99.3	98.4	99.8
VEHICLE	846	18	4	65.6	66.7	66.5	73.6	68.3	80.3
WAVEFORM	5000	21	3	78.1	80.1	80.9	82.7	83.3	82.5
WINE	178	13	3	92.1	92.5	92.4	95.9	95.4	96.6
ZOO	101	16	7	92.2	95.1	96.1	96.8	96.5	98.2
Moyenne				79.0	82.5	83.0	86.5	85.2	88.1

TAB. 1 – *Precision de C4.5, CBA, CPAR, CMAR, W-CARP, CARBoost, Random Forest*

Pour s'assurer que les différences observées sont significatives, c'est-à-dire qu'elle ne sont pas le simple fruit du hasard, nous avons d'abord utilisé le test apparié de Student pour comparer les précisions moyennes de deux méthodes sur les 26 mêmes bases. Nous avons redoublé ce test par un test non paramétrique pour échantillons appariés, le test du signe, qui teste la signification du résultat du match entre deux méthodes sur les 26 bases par comparaison avec la loi binomiale $B(26; 0.5)$ attendue sous l'hypothèse nulle (méthodes équivalentes). Les résultats de ces différents tests sont rapportés dans les tableaux 3 et 4 où le niveau de signification des comparaisons est indiqué par *(significatif), ** (très significatif) ou *** (très hautement significatif) suivant que 0.01 < p-value < 0.05, 0.001 < p-value < 0.01 ou p-value < 0.001. On rappelle que la p-value d'un test est la probabilité d'obtenir une valeur de la statistique

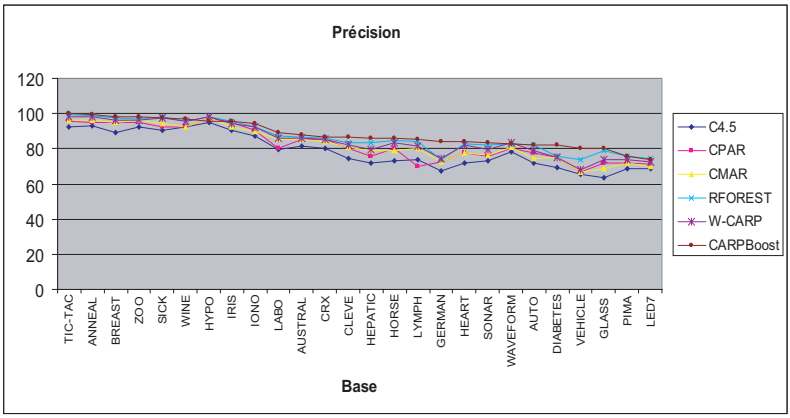

FIG. 1 – *Précision*

de test au moins aussi extrême, dans le sens de l'hypothèse alternative, que celle qui a été effectivement respectée, en supposant que l'hypothèse nulle est vraie.

Globalement, les résultats des expériences concernant la précision (tableaux 1 et 3) conduisent à des conclusions très claires :

- par rapport à CPAR et CMAR, les méthodes usuelles de classification associative, CARBoost enregistre un gain supérieur à 5 points de précision qui est très hautement significatif (p-values du test apparié de Student de l'ordre de 10^{-7}). On notera que CARBoost l'emporte presque systématiquement sur les méthodes précitées (25 fois sur 26, ce qui est très hautement significatif selon la p-value du test du signe qui est de l'ordre 10^{-7}). Par rapport à W-CARP la méthode que nous avons élaborée pour être le classifieur faible de CARBoost, le gain est à peine moins marqué (près de 3 points) tout en restant très hautement significatif (p-value de l'ordre de 10^{-5}) et en étant presque systématique (23 victoires et une égalité sur 26 bases, p-value de l'ordre de 10^{-5}).

- par rapport aux méthodes à base de règles prises comme référence, CARBoost améliore la précision de C4.5 de 9.1 points en moyenne (p-value très hautement significative, de l'ordre de 10^{-11}) et ne connaît aucune défaite sur les 26 bases testées (p-value de l'ordre de 10^{-11}). Comme attendu, le meilleur concurrent de CARBoost est l'algorithme des forêts aléatoires, mais CARBoost l'emporte quand même 19 fois sur 26 plus 2 égalités (p-value du test du signe égale à 0.0066), pour un gain moyen de 1.6 point, qui reste très significatif (p-value du test de Student égale à 0.0041).

En ce qui concerne le rappel, on trouve des résultats à peine moins marqués, très hautement significatifs dans l'ensemble :

- CARBoost enregistre un gain de rappel de 4.8 points par rapport à CPAR avec 24 victoires sur 26 contre un gain de 2.7 points par rapport à CMAR avec 21 victoires sur 26 plus 2 égalités. Face à W-CARP, CARBoost améliore systématiquement le rappel qui est augmenté de 4.1 points en moyenne.

Bases	C4.5	CPAR	CMAR	RF	WCARP	CARBoost
ANNEAL	82	82	84	87	85	89
AUSTRAL	76	77	78	79	78	81
AUTO	62	65	66	68	67	69
BREAST	73	79	78	83	78	85
CLEVE	75	77	80	82	79	83
CRX	81	81	83	85	83	86
DIABETES	62	67	78	71	68	75
GERMAN	65	68	67	70	67	71
GLASS	61	69	70	75	68	77
HEART	82	85	89	87	87	89
HEPATIC	65	68	71	77	69	81
HORSE	82	86	88	86	86	89
HYPO	80	82	85	92	84	91
IONO	84	87	89	91	88	92
IRIS	91	93	96	97	95	97
LABO	67	74	77	75	75	79
LED7	72	72	75	77	71	74
LYMPH	54	59	62	62	63	69
PIMA	65	65	64	65	64	66
SICK	85	84	85	89	81	84
SONAR	59	65	69	65	67	69
TIC-TAC	89	90	92	91	91	93
VEHICLE	69	71	74	74	72	75
WAVEFORM	73	74	71	75	70	72
WINE	85	86	88	89	87	91
ZOO	80	85	87	88	87	90
Moyenne	73.4	76.6	78.7	80.0	77.3	81.4

TAB. 2 – *Rappel de C4.5, CBA, CPAR, CMAR, W-CARP, CARBoost, Random Forest*

- de la même façon, CARBoost gagne 7.6 points de rappel sur C4.5 en moyenne et l'emporte 24 fois sur 26. Face aux forêts aléatoires, le gain n'est que de 1.4 points mais il reste très significatif ($p - value = 0.0070$) et il se retrouve sur la plupart des bases, avec 21 victoires plus 1 égalité sur 26 ($p - value = 0.0009$)

6 Conclusion et perspectives

La classification associative a l'intérêt d'être une méthode de prédiction à base de règles qui a de bonnes performances. Dans ce papier, nous proposons CARBoost, une version adaptative de la classification associative, inspirée du boosting, qui permet de faire émerger des règles capables de prédire correctement les exemples difficiles. Les performances de CARBoost, tant en termes de précision moyenne que de rappel moyen sont très améliorées aussi bien par rapport à C4.5, que par rapport à CPAR et CMAR les méthodes de classification associative les plus couramment utilisées, tout en gardant en grande partie l'intelligibilité du processus de classification d'un nouvel individu. La supériorité de CARBoost sur les forêts aléatoires est moins marquée, mais elle demeure très significative, de l'ordre de 1.5 point, aussi bien en rap-

Précision	C4.5	CPAR	CMAR	RF	W-CARP
Avantage CARBoost	9,1	5,7	5,2	1,6	2,9
ratio de Student	11,46	7,29	7,91	3,16	4,73
p-value test Student	2E-11	1E-07	3E-08	0,0041	8E-05
signification	***	***	***	**	***
nb succès CARBoost	26	25	25	19	23
nb échecs CARBoost	0	1	1	5	2
nbex aequo	0	0	0	2	1
p-value test signe	3E-08	8E-07	8E-07	0,0066	2E-05
signification	***	***	***	**	***

TAB. 3 – *Signification de la précision*

Rappel	C4.5	CPAR	CMAR	RF	W-CARP
Avantage CARBoost	7,6	4,8	2,7	1,4	4,1
ratio de Student	8,29	7,89	4,72	2,94	8,44
p-value test de Student	1E-08	3E-08	0,0001	0,0070	9E-09
signification	***	***	***	**	***
nb succès CARBoost	24	24	21	21	26
nb échecs CARBoost	2	1	3	4	0
nb égalités	0	1	2	1	0
p-value Signe	1E-05	2E-06	0,0003	0,0009	3E-08
signification	***	***	***	***	***

TAB. 4 – *Signification du rappel*

pel qu'en précision. En outre, les améliorations indiquées ont un aspect systématique, car elles se retrouvent avec plus ou moins d'intensité sur la quasi totalité des bases testées.

Nous envisageons de poursuivre ce travail en testant diverses repondérations des règles lors des différentes itérations de la procédure et en intégrant des critères d'évaluation supplémentaires. Nous souhaitons aussi examiner l'impact du nombre d'itérations T sur les résultats de notre méthode. Une première analyse sur Breast et Waves montre que l'erreur d'apprentissage est voisine de 0 pour $T = 10$ et que l'erreur en généralisation recommence à croître à partir de $T = 50$. Des expériences sur données réelles compléteront utilement celles déjà réalisées.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499.

Asuncion, A. et D. Newman (2007). UCI machine learning repository.

Bahri, E. et S. Lallich (2009a). Introduction de l'élagage pour l'extraction de règles d'association de classe sans génération de candidats. *Atelier QDC, EGC 2009, Strasbourg A6*, 29–36.

Bahri, E. et S. Lallich (Mai 2009b). Pour une classification associative plus efficace. *9ème Conférence francophone sur lŠapprentissage artificiel (CAP 09)*, 148–152.

Breiman, L. (1996). Bagging predictors. In *Machine Learning*, pp. 123–140.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Freund, Y. et R. Schapire (1996). Experiments with a new boosting algorithm. *In Machine Learning : Proceedings of the Thirteenth National Conference*, 148–156.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In *2000 ACM SIGMOD Intl. Conf. on Management of Data*, pp. 1–12. ACM Press.

Lallich, S., O. Teytaud, et E. Prudhomme (2006). *Association rules interestingness : measure and validation*, pp. 251–275. Quality Measures in Data Mining. Heidelberg, Germany : Springer.

Li, J. (2006). On optimal rule discovery. *IEEE Transformation on Knowldge and Data Engeneering. 18*(4), 460–471.

Li, W., J. Han, et J. Pei (2001). CMAR : Accurate and efficient classification based on multiple class-association rules. In *ICDM*, pp. 369–376.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pp. 80–86.

Meddouri, N. et M. Maddouri (2009). Générer des règles de classification par dopage de concepts formels. In *EGC*, pp. 181–186.

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Quinlan, J. R. et R. M. Cameron-Jones (1995). Induction of logic programs : FOIL and related systems. *New Generation Computing 13*, 287–312.

Sun, Y., Y. Wang, et A. K. Wong (2006). Boosting an associative classifier. *IEEE Transactions on Knowledge and Data Engineering 18*(7), 988–992.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*, 241–259.

Yin, X. et J. Han (2003). CPAR : Classification based on predictive association rules. In *3rd SIAM International Conference on Data Mining*, pp. 331–335.

Yoon, Y. et G. G. Lee (2008). Text categorization based on boosting association rules. *International Conference on Semantic Computing 0*, 136–143.

Summary

Associative classification is a method of effective prediction resulting from the mining association rules. This method is particularly interesting because it search in an exhaustive way the relevant association rules which it filters to keep only the class association rules (those having for consequent a class attribute). Those association rules are used as classifier. Discovery knowledge is directly interpretable. Former studies show the disadvantages of this approach, which it is of the massive generation of not useless rules or about the bad prediction of the minority class when the classes are unbalanced. One proposes an original approach of the boosting class association rules which uses as weak learner a base of significant rules built by an supervised frequent itemsets generation's algorithm that is limited to the extraction of the only significant class rules and that takes into account the imbalance of the data. Comparisons with other methods of associative classification show that our approach improves the precision and the recall.