

LABORATOIRE



INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 6070

EXTRACTION DE RÈGLES SELON DES CRITÈRES MULTIPLES : L'ART DU COMPROMIS

Dominique FRANCISCI, Laurent BRISSON, Martine COLLARD

Projet MECOSI

Rapport de recherche
ISRN I3S/RR-2003-11-FR

Mai 2003

RÉSUMÉ :

Cet article s'intéresse à l'utilité des règles découvertes en fouille de données. Notre approche est expérimentale et basée sur la remarque suivante : la qualité des connaissances extraites dans un processus de fouille de données doit être évaluée selon plusieurs critères comme la précision, la nouveauté, la compréhensibilité ; cependant les travaux sur ce thème ne prennent pas en compte l'intégration de plusieurs facteurs. Nous présentons quelques résultats de l'étude comparative que nous avons menée sur le comportement relatif de différentes mesures de qualité. Nous traçons des nuages de points représentant des règles caractérisées par la donnée de deux mesures. Nous observons certains phénomènes et en particulier des antagonismes entre mesures. Cette situation suggère de parcourir l'espace de recherche des règles pour découvrir des compromis plutôt que des règles optimales. Nous étudions différents cas et montrons l'intérêt d'une approche scalaire pour découvrir les règles optimisant plusieurs critères. Pour implémenter cette solution, nous avons choisi un algorithme génétique apte à parcourir des espaces de recherche très important et permettant de combiner aisément plusieurs critères par l'intermédiaire de sa fonction de fitness.

MOTS CLÉS :

fouille de données, extraction de règles, critères de qualité, approche scalaire, algorithme génétique

ABSTRACT:

In this paper, we are interested in the interestingness of rules extracted from data. Our approach is experimental and based on the following remark : the quality of discovered rules in data mining process has to be measured according several criteria such as accuracy, interestingness and comprehensibility, but most work on rule interestingness do not take into account the integration of multiple factors. We present a comparative study which highlights the relative behaviour of different quality measures. We plot sets of rules according a pair of criteria. We obtain graphical representations which show several antagonisms among quality measures. This situation suggests to look in the rule search space for compromise rather than for best rules. We study different solutions and show the advantage of a scalar method for multi-criteria optimisation. The solution is implemented via a genetic algorithm. We have run the algorithm several times in order to propose a set of compromise to the user.

KEY WORDS :

data mining, rules extraction, quality criteria, scalar approach, genetic algorithm

Extraction de règles selon des critères multiples : l'art du compromis

Dominique Francisci, Laurent Brisson, Martine Collard

Laboratoire I3S - Université de Nice-Sophia Antipolis

Les Algorithmes - Bât. Euclide B

2000 route des Lucioles, B.P. 121

06903 Sophia Antipolis Cedex

francis@i3s.unice.fr, brisson@essi.fr, mcollard@unice.fr

MOTS-CLÉS : fouille de données, extraction de règles, critères de qualité, approche scalaire, algorithme génétique

RESUME

Cet article s'intéresse à l'utilité des règles découvertes en fouille de données. Notre approche est expérimentale et basée sur la remarque suivante : la qualité des connaissances extraites dans un processus de fouille de données doit être évaluée selon plusieurs critères comme la précision, la nouveauté, la compréhensibilité ; cependant les travaux sur ce thème ne prennent pas en compte l'intégration de plusieurs facteurs. Nous présentons quelques résultats de l'étude comparative que nous avons menée sur le comportement relatif de différentes mesures de qualité. Nous traçons des nuages de points représentant des règles caractérisées par la donnée de deux mesures. Nous observons certains phénomènes et en particulier des antagonismes entre mesures. Cette situation suggère de parcourir l'espace de recherche des règles pour découvrir des compromis plutôt que des règles optimales. Nous étudions différents cas et montrons l'intérêt d'une approche scalaire pour découvrir les règles optimisant plusieurs critères. Pour implémenter cette solution, nous avons choisi un algorithme génétique apte à parcourir des espaces de recherche très importants et permettant de combiner aisément plusieurs critères par l'intermédiaire de sa fonction de fitness.

1. Introduction

La fouille de données se définit comme une activité de découverte de connaissances nouvelles, enfouies dans des volumes de données d'une dimension nécessitant des traitements automatiques. Dans les algorithmes mis en œuvre pour l'extraction des modèles de connaissances, le critère de sélection du meilleur modèle est en général relatif au degré d'exactitude des modèles. Cependant, il s'est établi un consensus sur le fait que ces mesures standards sont insuffisantes pour extraire des connaissances réellement pertinentes, c'est à dire intéressantes, surprenantes et utiles. Aussi, un certain nombre de mesures a été défini pour pallier ce problème. La définition de chaque facteur d'intérêt comme le coefficient *RI* de Piatesky-Shapiro [PIAT91], le *lift* [IBM96], la *J-Measure* de Goodman et Smith [SMIT91], le coefficient de Sebag et Schoenauer [SEBA88] est justifiée par le fait qu'ils vérifient des propriétés posées axiomatiquement comme propriétés d'une théorie de l'*intéressabilité* des règles. Quelques synthèses ont été réalisées ; certaines se proposent d'évaluer, de manière la plus exhaustive possible, ces différentes mesures par rapport à un certain nombre de meta-critères.

Dans cet article, nous nous intéressons également à l'extraction de règles potentiellement utiles. Notre approche est expérimentale et se fonde sur la remarque suivante : la qualité d'une règle se mesure non seulement à son intérêt, mais également à sa précision et sa compréhensibilité et éventuellement selon un ensemble de critères dépendant du contexte et de l'objectif fixé ; cependant, les travaux sur le caractère intéressant des modèles n'évoquent pas cette question qui impose d'intégrer plusieurs mesures dans l'évaluation.

Nous nous focalisons sur les modèles à forme de règles $A \rightarrow B$ où A et B sont des conjonctions de termes attribut-valeur ; notre objectif est de mesurer la qualité de la règle c'est à dire la pertinence du lien de causalité selon différents critères.

En fouille de données, règles de classification et règles d'association sont les types de modèles les plus fréquemment recherchés dans les processus de fouille de données. Elles constituent en effet, les modèles les plus compréhensibles et les plus aisément exploitables par l'utilisateur. Les algorithmes standards mis en œuvre pour découvrir chacun de ces modèles sont radicalement différents les uns des autres ; ils sélectionnent les règles les meilleures règles selon des critères différents. Les systèmes de classifieur constitués de règles de classification ont été étudiés intensivement dans le domaine de l'apprentissage automatique à partir d'exemples. Les algorithmes de recherche visent à obtenir un ensemble de règles le plus exact possible, c'est à dire couvrant un maximum d'exemples. L'algorithme *C4.5* [QUIN93] est un exemple, parmi les plus cités, d'algorithme de ce type. Il est mis en œuvre en fouille de données également. Pour les règles d'association, définies spécifiquement dans le cadre de la fouille de données sont définis des algorithmes pour la plupart extensions de l'algorithme fondateur *Apriori* [AGRA94]. Cet algorithme vise à sélectionner les règles les plus générales et, parmi celles-ci les règles les plus pertinentes selon le critère statistique de confiance qui représente la proportion d'exemples vérifiant la conclusion de la règle parmi ceux qui vérifient ses prémisses. Dans cet article, nous utilisons ces deux algorithmes comme éléments de comparaison dans les résultats des tests.

Cependant, le but poursuivi lors de la mise en œuvre d'actions de fouille de données est de découvrir non pas nécessairement des modèles généraux, mais plutôt des informations, liens, dépendances, motifs utiles, intéressants, surprenants, éventuellement vérifiés sur une petite portion de données. Par contre, les algorithmes standards ne sélectionnent pas les règles sur des critères d'utilité potentielle pour l'utilisateur et génèrent ainsi souvent de grandes quantités de règles sans réelle pertinence.

Ceci a motivé les recherches pour proposer des mesures d'intérêt qui tendent à mesurer la différence entre la situation de dépendance entre prémisses et conclusion par rapport au cas d'indépendance ; cependant, les propositions de nouveaux facteurs suggèrent implicitement de procéder en deux phases : extraction des règles selon un algorithme standard, puis filtrage par rapport à une mesure d'intérêt. Or, cette approche peut passer à côté de motifs pertinents. Par exemple *C4.5* optimise le gain d'information résultant de l'occurrence d'un attribut dans la règle et *Apriori* maximise support et confiance. L'approche en deux étapes ne permet de filtrer les motifs intéressants qu'à partir des règles sélectionnées selon ces critères.

Nous nous intéressons ici à tout motif sous forme de règle et ne faisons pas de distinction entre les différents types de règles ; notre objectif est de mesurer l'intérêt du lien de causalité entre prémisses et conclusions.

Les travaux que nous présentons, se caractérisent par une approche expérimentale ; en effet, l'observation de quelques mesures statistiques sur les règles extraites par *C4.5* et par *Apriori* nous a amené à réaliser une étude comparative avec l'objectif de mettre en évidence le comportement relatif des facteurs de qualité pouvant intervenir dans l'évaluation d'une règle. Aussi, pour mettre en évidence la corrélation ou non-corrélation entre deux mesures données, nous traçons des nuages de points représentant des ensembles de règles caractérisées par leurs valeurs pour ces mesures. Les ensembles de règles représentées sont de trois types : il s'agit des règles extraites à l'aide des algorithmes standards *C4.5* et *Apriori* et d'un ensemble de règles générées aléatoirement. Les graphes obtenus pour différents couples de critères et sur différents jeux de données font apparaître les comportements sélectifs des critères dans l'espace de recherche ainsi que le comportement des algorithmes. Nous vérifions en particulier le fait qu'il existe des règles ignorées par les algorithmes standards, et cependant intéressantes au vu des mesures choisies. Nous observons également des antagonismes apparents entre mesures qui suggèrent, s'ils ont confirmés, la recherche de compromis en l'absence de solution optimale. Pour la

recherche de ces règles, nous étudions diverses possibilités et montrons l'intérêt d'une approche scalaire d'optimisation multi-critères. Pour implémenter cette solution, nous avons fait le choix d'un algorithme génétique qui a l'avantage d'effectuer une recherche globale manipulant une population de solutions potentielles par opposition aux méthodes heuristiques de recherche locale (hill-climbing, recherche Tabou...) qui risquent de n'identifier que des optima locaux. L'algorithme génétique permet également de définir aisément une fonction d'évaluation intégrant les différents critères. Nous réalisons plusieurs exécutions de l'algorithme en faisant varier les poids représentant l'importance relative de chaque critère ce qui nous permet de proposer un ensemble de compromis dans lequel l'utilisateur peut faire un choix.

Cet article est organisé de la manière suivante : la section II présente un rappel sur le fonctionnement des méthodes standards en se focalisant sur le problème des modèles intéressants ; dans la section III, nous présentons une synthèse de l'étude comparative menée sur différents couples de mesures ; la section IV est consacrée à la mise en œuvre de la solution scalaire par algorithme évolutionnaire et aux résultats obtenus ; enfin la section V établit un bilan des travaux menés et donne un aperçu des travaux futurs.

2. Extraction de règles et mesures de qualité

Règles de classification et règles d'association sont les types de modèles les plus fréquemment recherchés dans les processus de fouille de données. Elles constituent en effet, les modèles les plus compréhensibles et les plus aisément exploitables par l'utilisateur.

L'extraction de règles de classification tend à découvrir un petit ensemble de règles qui constituent un classifieur le plus exact possible alors que l'extraction de règles d'association consiste en la recherche des règles de meilleur support et confiance conformément à l'algorithme fondateur *Apriori* défini par R. Agrawal & al. [AGRA94]. A. Freitas [FREI00] discute des particularités de chacun des types de règles et insiste sur leurs différences ; il montre que le principe de fonctionnement de l'algorithme *Apriori* implique que les règles découvertes jouent un rôle descriptif alors que les règles de classification, apprises sur un ensemble d'exemples et testées sur un autre ensemble entrent dans le cadre de la prédiction.

Un classifieur est évalué par sa qualité de prédiction sur l'ensemble de test, c'est à dire par le taux de succès, pourcentage d'exemples de l'ensemble de test qui sont bien classés par le modèle. Pour une règle $A \rightarrow C$, ce taux peut s'écrire en termes de probabilité : $p(A \cap C) + p(\bar{A} \cap \bar{C})$. Pour mesurer la qualité de classement, la plupart des algorithmes utilisent cette mesure qui reste globale.

La mesure de taux d'erreur, bien que significative, ne rend pas compte des performances réelles du classifieur, car il ne prend pas en compte le déséquilibre qui peut exister dans la distribution des classes dans l'ensemble d'exemples. Aussi, les performances des modèles prédictifs font souvent état de mesures sur les "vrais positifs" et les "vrais négatifs". Il est courant d'utiliser les notations suivantes pour une classe C et un classifieur :

TP (ou true positives) pour le nombre d'exemples appartenant à C et prédits dans C par le modèle

FP (ou false positives) pour le nombre d'exemples n'appartenant pas à C et prédits dans C par le modèle

TN (ou true negatives) pour le nombre d'exemples n'appartenant pas à C et non prédits dans C par le modèle

FN (ou false negatives) pour le nombre d'exemples appartenant à C et non prédits dans C par le modèle

On définit le taux de vrais positifs, encore appelé *sensibilité* ou *recall* par $Se = TP / (TP+FN)$ et le taux de vrais négatifs, ou *spécificité* par $Sp = TN / (TN+TP)$. Ces mesures permettent d'évaluer l'intérêt du modèle de manière plus précise en évaluant les taux de prédiction dans chaque classe. Pour une règle $r : A \rightarrow C$, on peut noter que $Se(r)$ représente la probabilité conditionnelle $p(A/C)$ ou encore la couverture de C par A . De la même façon, $Sp(r)$ représente la probabilité conditionnelle $p(\bar{A}/\bar{C})$ ou la couverture de \bar{C} par \bar{A} . Un classifieur peut présenter un taux de succès général élevé alors que la sensibilité ou la spécificité sont faibles. On définit également la *précision* par la formule $TP / (TP+FP)$; cette mesure représente donc la probabilité conditionnelle $p(C/A)$ et donc la couverture de C par A , ce qui correspond à la confiance dans le contexte des associations. On peut remarquer que $|C|$ étant fixé, précision et spécificité sont fortement et positivement corrélées, alors que Se et Sp ne sont pas corrélées.

Selon l'objectif poursuivi par la tâche d'extraction, il est nécessaire d'extraire des règles les meilleures selon un critère particulier. Par exemple, s'il s'agit d'évaluer le classement dans la classe "malade" selon le résultat d'un test de diagnostique médical, on considère plus essentiel de minimiser le taux des erreurs consistant à prédire qu'un sujet est sain alors qu'il est en fait malade plutôt que de minimiser l'erreur contraire. On va donc dans ce cas, rechercher des règles qui minimisent le taux de faux négatifs et qui sont donc les meilleures en termes de sensibilité. Supposons maintenant qu'il s'agisse de prédire le profil (bon ou mauvais) d'un client pour une demande crédit ; supposons également que les objectifs soient doubles pour l'organisme de crédit : ne pas prendre le risque de perdre de l'argent, mais également ne pas risquer de perdre un bon client. Dans ce cas, ces contraintes se traduisent par la nécessité de trouver les règles les meilleures selon Se et également selon Sp . On peut d'ailleurs citer la proposition de [FIDE00] de rechercher les règles qui optimisent le produit $Se \times Sp$.

Les considérations présentées ici sur les mesures d'évaluation des modèles prédictifs sont relatives à son pouvoir prédictif. Cependant, si l'on considère un point de vue plus orienté "fouille de données", la qualité d'un classifieur ne réside pas seulement dans sa précision, mais également dans son utilité pour les experts du domaine, dans la mesure où il leur révèle des informations insoupçonnées et intéressantes. L'évaluation devient alors plus délicate et plus complexe. Mais cet aspect concerne également les modèles descriptifs.

Les associations représentent des sortes de régularités observées dans les données qui mettent en évidence des corrélations fortes entre les attributs. Les algorithmes d'extraction de règles d'association basé sur *Apriori* procèdent en deux phases : extraction des itemsets $A \cap B$ les plus fréquents puis filtrage des règles induites $A \rightarrow B$ selon la confiance de la règle.

Dans ce contexte, la fréquence des itemsets ou fréquence des exemples qui vérifient A et B est appelée support et noté $supp(A \cap B)$. La confiance $conf$ d'une règle $A \rightarrow B$ est définie par la fréquence des exemples vérifiant A qui vérifient B également. Ainsi, en termes de probabilité, on peut écrire $supp(A \cap B) = p(A \cap B)$ et $conf(A \rightarrow B) = p(B/A)$. Les règles considérées comme les meilleures et donc extraites satisfont simultanément un seuil de support et un seuil de confiance minimum. Cependant, support et confiance s'avèrent être insuffisants pour sélectionner un ensemble minimal de règles "intéressantes". Les algorithmes de recherche d'associations peuvent générer des milliers de règles alors que peu d'entre elles sont réellement utiles. Une première sélection consiste à supprimer les règles redondantes qui peuvent être déduites

à partir des autres. Mais ceci ne permet pas d'éviter des règles sans intérêt. Par exemple, si la règle *ordinateur* \rightarrow *jeux_vidéo* présente une confiance de 40%, elle n'apporte cependant aucune information utile si, par ailleurs, la fréquence des ventes d'*ordinateur* est de 50%.

Cette observation a amené à définir pour une règle $A \rightarrow B$, des mesures dont les valeurs sont d'autant plus grandes que l'on s'éloigne (avec une corrélation positive) de la situation d'indépendance. En premier lieu, G. Piatetsky-Shapiro [PIAT91] a défini trois principes pour une mesure d'intérêt et a proposé la mesure *RI* définie par le produit $|A| \times (p(B/A) - p(B))$. La mesure de *lift* [IBM96] définit par l'expression $\frac{p(A \cap B)}{p(A) \times p(B)}$ peut s'interpréter comme le coefficient multiplicateur qui est appliqué à la probabilité d'avoir B a priori lorsque A est réalisée. La mesure *RI* et le *lift* sont symétriques en A et B alors que la mesure de Sebag et Schoenauer [SEBA88] défini par $\frac{p(A \cap B)}{p(A \cap \bar{B})}$ ou la *J-Mesure*

[SMIT91] prennent en compte les contre-exemples. La *J-Mesure* s'inspire de la théorie de l'information en calculant une entropie croisée qui correspond à une distance entre les probabilités de B a priori et a posteriori. Elle est de la forme :

$$J\text{-Mesure}(A \rightarrow B) = p(A) \times \left(p(B/A) \times \log \left(\frac{p(B/A)}{p(A)} \right) + p(\bar{B}/A) \times \log \left(\frac{p(\bar{B}/A)}{p(A)} \right) \right)$$

Des travaux de synthèses sur les différentes mesures ont été menées [FREI99, HILD00, LALL02, LENC02]. [FREI99] énonce un certain nombre de principes pour sélectionner des règles intéressantes et propose d'intégrer différents facteurs d'intérêt répondant à ces principes dans une fonction d'évaluation, [HILD00] définit une théorie de l'intéressabilité pour des agrégats et des généralisations, [LENC02] propose d'appliquer une stratégie d'aide à la décision multi-critères pour faire un choix parmi les nombreux critères d'intérêt, [LALL02] compare chaque critère d'une liste très complète à un ensemble de principes énoncés définissant une "bonne" mesure d'intérêt. Il montre également que certaines mesures comme le *lift* et *RI* classent les règles de la même façon.

Dans ce travail, nous nous intéressons non seulement aux mesures d'intérêt, mais également aux mesures de précision vus ci-dessus. Un test rapide, comme l'illustrent les résultats affichés dans les tableaux n°1 et n°2, montrent que ces mesures ne classent pas de la même manière et ne paraissent pas corrélées. Les tests ont été réalisés sur les bases de l'UCI [UCI] Mushroom et Dermatology. Le tableau n°1 montrent les mesures de confiance et de *RI* de règles d'association obtenues sur Mushroom et le tableau n°2 montrent les mesures de taux de succès et de *J-Mesure* de règles de classification obtenues sur Dermatology ; on peut observer sur chaque exemple, l'absence de corrélation positive ou négative entre la mesure de précision et la mesure d'intérêt.

On comprend ainsi que des règles intéressantes au sens de la mesure *RI* soient ignorées par une algorithmes qui effectue une première sélection basée sur la mesure de confiance, par exemple.

Comment rechercher des règles qui satisfassent plusieurs critères : confiance, couverture, intérêt au sens de la *J-Mesure*, de *RI*, du *lift*, sensibilité, spécificité ? Est-il justifié d'utiliser un produit comme l'expression de *RI* ou $Se \times Sp$?

Dans la section suivante, nous étudions expérimentalement, les éventuelles corrélations entre critères qui pourraient justifier de se ramener à un seul critère.

<i>Confiance</i>	<i>RI</i>
0,5584	520,4215
0,6286	504,1221
0,6985	1013,1443
0,7723	1223,7676
0,7839	1004,1398
0,9037	1470,6844
0,9295	1260,9355

Tableau n°1 :

Confiance versus RI sur la base Mushroom

<i>Taux de succès</i>	<i>J-Mesure</i>
0,724	0,0356
0,8634	0,0237
0,8743	0,1153
0,9126	0,0934
0,9208	0,191
0,9563	0,0318
0,9809	0,1033

Tableau n°2 :

Taux de succès versus J-Mesure sur la base Dermatology

3. Etude comparative de mesures

Généralement, afin de déterminer les règles qui peuvent être intéressantes parmi toutes celles générées par l'algorithme utilisé, nous n'utilisons pas un unique critère mais un ensemble de critères, que nous souhaitons optimiser. Dans le cadre de notre étude, nous nous sommes restreints à des comparaisons sur des couples de mesures uniquement. Nous avons réalisé un ensemble de tests sur les jeux de données Dermatology, Vote, Adult, Mushroom, Soybean, Nursery, Zoo et Sonar issus de la bibliothèque de l'UCI. Pour les algorithmes standards, nous avons utilisé les implémentations du système *Weka* [WEKA, WITT99] développé à l'université de Waikato. Pour chacun de ces couples nous avons comparé les règles obtenues de trois manières différentes : en exécutant les implémentations de *Weka* de *C4.5 (J4.8)* et d'*Apriori* et en générant des règles aléatoirement. Bien entendu, puisque que nous désirions effectuer une tâche de classement, seules les règles générées par *Apriori* sous la forme d'une règle de classification ont été conservées. Cela explique la baisse de performances d'*Apriori* dans certaines conditions. Il est également important de noter que les règles générées par *J4.8* sont évaluées individuellement alors que dans un arbre de décision on évalue leurs performances en tant que classifieur. Cela n'est toutefois pas gênant dans le cadre que nous nous sommes fixés car nous cherchons plutôt à observer le comportement des différentes mesures pour des règles uniques intéressantes que pour des classifieurs. Le générateur de règles aléatoires, quant à lui, nous permet de mettre en évidence certaines règles mais il est important de garder à l'esprit qu'il n'est toutefois pas capable de donner une vision globale de l'espace de recherche des règles. Pour chaque jeu de données jusqu'à 4000 règles ayant de 2 à 5 attributs en partie gauche ont été générées.

Nous avons conduit une étude comparative sur différents types de mesures. Dans ce qui suit, nous présentons des résultats caractéristiques obtenus sur trois couples différents de mesures : le premier illustre les comportements respectifs d'une mesure de précision et d'une mesure d'intérêt, le deuxième démontre que les mesures de sensibilité et spécificité ne sont pas optimisées simultanément par les algorithmes standards et le dernier compare le facteur de généralité et le facteur d'intérêt intégré dans la mesure *RI* de Piatetsky-Shapiro [PIAT91].

Dans l'optique d'une comparaison de deux mesures, ce premier choix se portant sur un couple précision/intérêt nous a semblé intéressant car il permet d'une part de juger des performances d'une règle sous un angle statistique et d'autre part de juger de sa faculté à nous intéresser. Sur la figure n°1 on peut observer les performances des différents algorithmes pour le couple Se/RI sur la base Dermatology. La sensibilité a été choisie en tant que mesure de précision pour sa capacité à mesurer, pour une règle $A \rightarrow B$, la couverture de B par A . La mesure RI complète le critère de précision par un critère d'intérêt ; elle permet de mesurer une corrélation positive ou négative entre A et B . Nous pouvons remarquer un antagonisme apparent entre les deux mesures, illustré sur la figure n°1 par la ligne tracée en frontière du nuage : en effet, parmi les règles extraites par les algorithmes, la mesure de RI ne donne pas de les meilleurs résultats pour une sensibilité élevée. Il est impossible, avec les algorithmes standards, de trouver des règles obtenant les meilleurs résultats sur les deux mesures simultanément. Seul un ensemble de compromis optimisant au mieux le couple de mesures pourra être sélectionné.

La résolution de nombreux problèmes réels nécessitent que plusieurs critères soient optimisés. Ces critères sont souvent irrationnels et dépendants les uns des autres. Bien que lors de l'optimisation d'un unique critère la solution soit généralement clairement définie, ce n'est pas le cas dans le cadre d'une optimisation multi-critères. Dans ce domaine il n'existe pas une unique « meilleure » solution mais un ensemble de compromis ; cet ensemble est souvent appelé « ensemble des optima de Pareto ». Ces solutions sont dites optimales car aucune solution de l'espace de recherche n'est

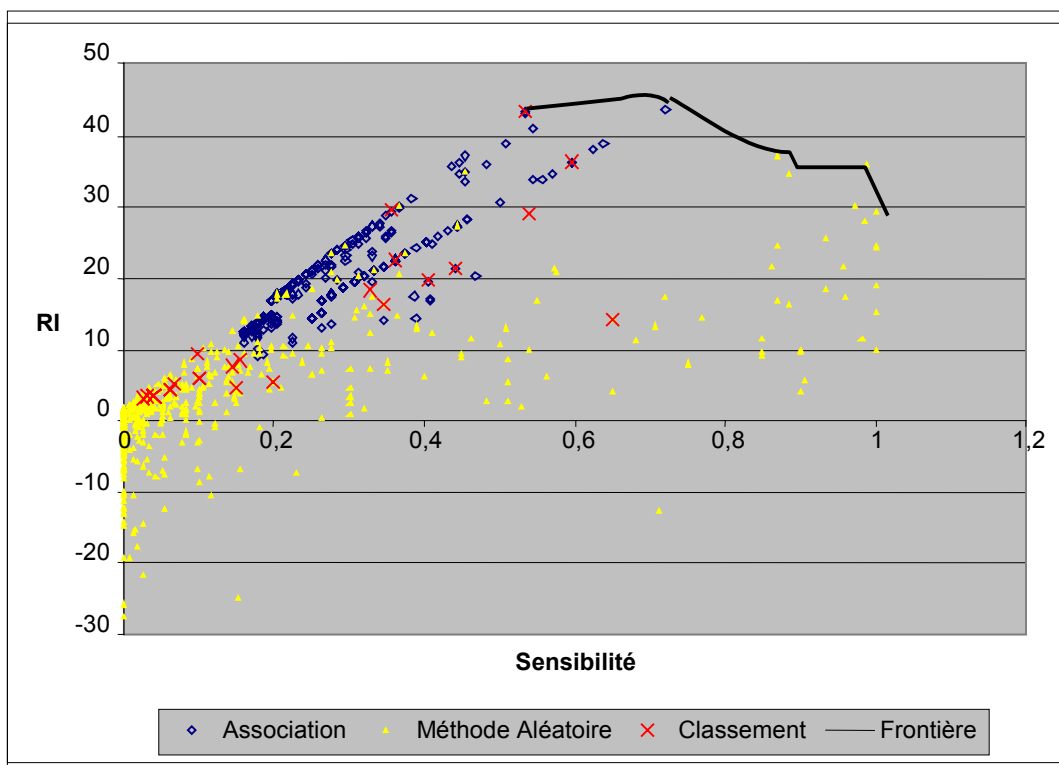


Figure n° 1 : Se versus RI sur la base Dermatology

meilleure si l'on considère tous les critères simultanément. Ce phénomène est plus ou moins avéré selon les différentes bases et les couples de critères que nous avons observés ; on peut remarquer souvent la présence de « niches de règles ».

Par exemple sur la figure n°1 nous pouvons observer que les règles obtenues par les algorithmes de recherche de règles ont, sur le jeu de données Dermatology, une assez bonne valeur de RI , cependant une niche de bonnes règles (ou bons compromis) qui émerge grâce au générateur de règles aléatoires est ignorée. Cet exemple a ainsi mis en évidence le fait que pour certaines mesures, la méthode qui consisterait à filtrer les règles fournies par un algorithme standard selon les mesures Se et RI ignorerait de bonnes solutions. Toutefois, le choix d'un couple précision/intérêt n'est pas pertinent dans toutes les situations. Les critères considérés doivent s'adapter aux objectifs propres au domaine d'application. Dans certains secteurs d'activité, la qualité d'une règle se réfère à d'autres qualités que la précision ou l'intérêt au sens précédemment exposé. Prenons par exemple le domaine médical où la mesure de sensibilité est complétée par la mesure spécificité qui représente la couverture des contre-exemples par \bar{A} . Afin d'obtenir les meilleurs compromis possibles il peut être intéressant d'optimiser le produit $Se \times Sp$ sur l'ensemble des règles découvertes par *Apriori* ou *C4.5*.

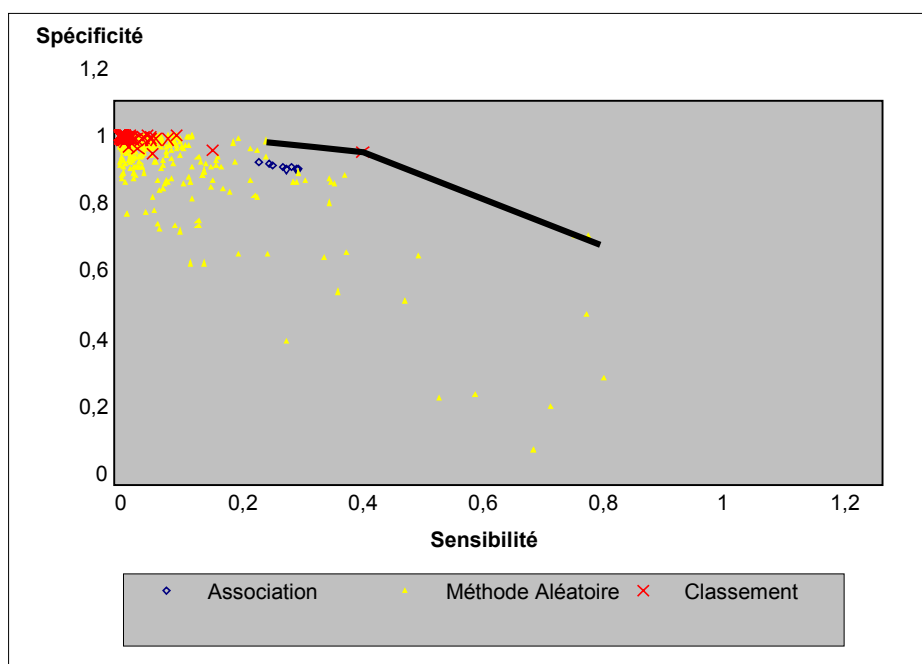


Figure n°2 : Se versus Sp sur la base Adult

Il est intéressant de remarquer sur la figure n°2, qu'une nouvelle fois certaines des règles générées aléatoirement sont meilleures (au sens des deux critères) que celles obtenues par les algorithmes standards. Comme le montre le nuage, ces algorithmes n'optimisent pas le produit considéré : nous observons donc à nouveau les défauts d'une technique qui sélectionne des règles générées selon d'autres critères. Des solutions alternatives ont été explorées : par exemple M. Fidelis [FIDE00] propose de rechercher les solutions qui maximisent le produit $Se \times Sp$ à l'aide d'un algorithme génétique dont la fonction d'évaluation est égale au produit. Toutefois, bien que le produit permette d'accéder à certaines règles négligées par les algorithmes traditionnels nous pouvons objecter que cette approche manque de précision. En effet il peut exister de bons compromis que l'optimisation du simple produit de deux mesures ne permet pas de découvrir.

Afin d'introduire une technique plus précise, nous avons choisi les deux facteurs de la mesure RI de Piattetsky-Shapiro définie précisément comme un produit. Pour une règle r de la forme $A \rightarrow B$, les facteurs $|A|$ et $conf(r) - p(C)$ évaluent respectivement la taille de l'antécédent et la corrélation (positive ou négative) entre antécédent et conclusion, le

produit des deux facteurs étant la mesure RI . Sur la figure n°3 obtenue sur le jeu de données Zoo nous pouvons observer l'antagonisme apparent des deux critères, sur les règles générées aléatoirement. Le tracé de l'hyperbole en pointillés qui coupe le nuage de points extraits par *Apriori*, montre que la recherche des règles maximisant le produit des facteurs produirait une solution commune avec *Apriori*. En définissant une certaine marge, il est possible de sélectionner d'autres règles (générées par *Apriori* pour la plupart) optimisant ce produit d'une façon quelque peu différente (hyperbole en trait plein). Cependant le nuage de règles montre que cette technique ne permet pas de sélectionner certaines règles se situant sur la frontière de Pareto. Les règles ignorées représentées sur cette figure sont aussi bien issues de l'algorithme *J4.8* que du générateur de règles aléatoires.

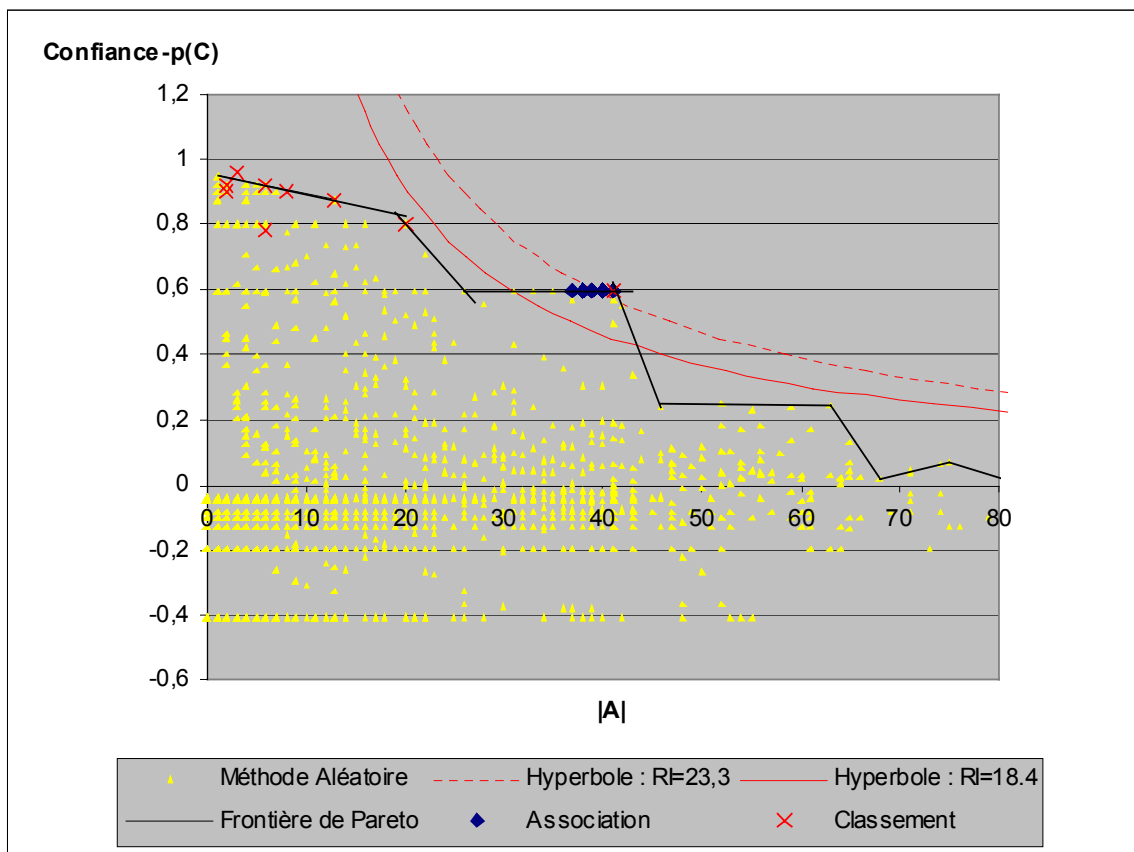


Figure n°3 : $|A|$ versus $conf(r) - p(C)$ sur la base Zoo

Ainsi, puisque la sélection par optimisation du produit n'est pas suffisante il est nécessaire de s'orienter vers une méthode permettant de déterminer au mieux les couples se situant sur la frontière de Pareto, afin d'obtenir une sélection plus précise de règles répondant à nos attentes. Comme nous pouvons le voir sur la figure n°3 une bonne approximation de la frontière peut être obtenue au moyen d'un ensemble de segments de droites. Optimiser des combinaisons linéaires de mesures pourrait donc permettre de découvrir les règles intéressantes se situant sur cette frontière. Sous cet aspect, cette étude s'identifie à une évaluation multi-critères des règles pour laquelle nous nous orientons vers une technique scalaire. L'utilisation d'un algorithme génétique où la fonction de fitness prend la forme de différentes combinaisons linéaires de

mesures s'avère assez souple ; de plus, la mise en œuvre d'un algorithme de ce type parcourant plus largement l'espace de recherche permet de s'assurer de l'antagonisme de certains critères comme observé sur les figures n°1 et n°2.

3. Mise en œuvre d'une solution scalaire

Dans cette section, nous présentons les résultats obtenus en mettant en œuvre une solution scalaire pour l'évaluation de critères antagonistes. Nous avons mis en œuvre un algorithme génétique (AG) dont le rôle est d'optimiser une fonction définie comme moyenne pondérée de deux mesures de qualité. Notre choix s'est porté sur un algorithme évolutionnaire (AE) en général et un AG en particulier pour trois raisons principales :

- la fonction d'évaluation, permet d'exprimer de façon aisée le concept de combinaisons de critères pondérés,
- d'autre part, compte tenu des vastes espaces de recherche associés aux bases que nous utilisons et d'une manière générale, aux bases exploitées en data mining, les AG offrent des perspectives intéressantes compte tenu de leur opérateurs génétiques (croisement, mutation, ...) permettant d'explorer de façon très large ces espaces,
- enfin, ils sont « potentiellement » aptes à ne pas converger vers des optima locaux.

Les AG ont été introduits par John Holland en 1975, avec la présentation d'une méthode inspirée de l'observation des capacités d'adaptation et d'évolution des espèces. Il construit un système artificiel qui s'appuyait sur les principes de sélection de Darwin et sur les méthodes de combinaison des gènes de Mendel. Les AG renvoient aux principaux mécanismes de l'évolution naturelle, c'est à dire essentiellement la sélection, la reproduction et la mutation. Ils décrivent l'évolution, au cours de générations successives, d'une population d'individus en réponse à leur environnement. Ils sélectionnent les individus, en accord avec le principe de la survie du plus adapté. Comme leurs équivalents biologiques, les individus sont constitués de gènes qui ont chacun un rôle propre. Dans une simulation génétique, les individus les plus adaptés ont une probabilité plus élevée d'être sélectionnés et reproduits, donc d'être présents à la génération suivante. L'opération de mutation d'un gène permet de maintenir une certaine diversité dans la population.

Les AG travaillent sur une population de solutions potentielles. Le processus, schématisé dans la figure n°4, conduit à l'élimination des éléments les plus faibles pour favoriser la conservation et la reproduction des individus les plus performants. La recombinaison (reproduction par hybridation génétique ou croisement) des individus les plus forts peut donner naissance à des individus encore meilleurs à la génération suivante.

Dans l'AG que nous mettons en œuvre dans le cadre de cet article, chaque individu représente l'antécédent d'une règle de la forme *SI condition ALORS prédiction*. La partie "prédiction" de la règle étant fixée au début de l'exécution de l'algorithme et correspond à un attribut cible. Ainsi, pour découvrir plusieurs règles prédisant des attributs cibles différents, il est nécessaire d'effectuer plusieurs exécutions de l'AG, une pour chacune des valeurs de l'attribut cible. Les sections suivantes présentent respectivement la représentation des individus dans l'AG, la fonction d'évaluation, la sélection ainsi que les opérateurs génétiques mis en œuvre et enfin les résultats expérimentaux obtenus par application de cet algorithme sur plusieurs bases. Le schéma de la figure 1 synthétise le fonctionnement d'un AG et les principales entités qui le composent.

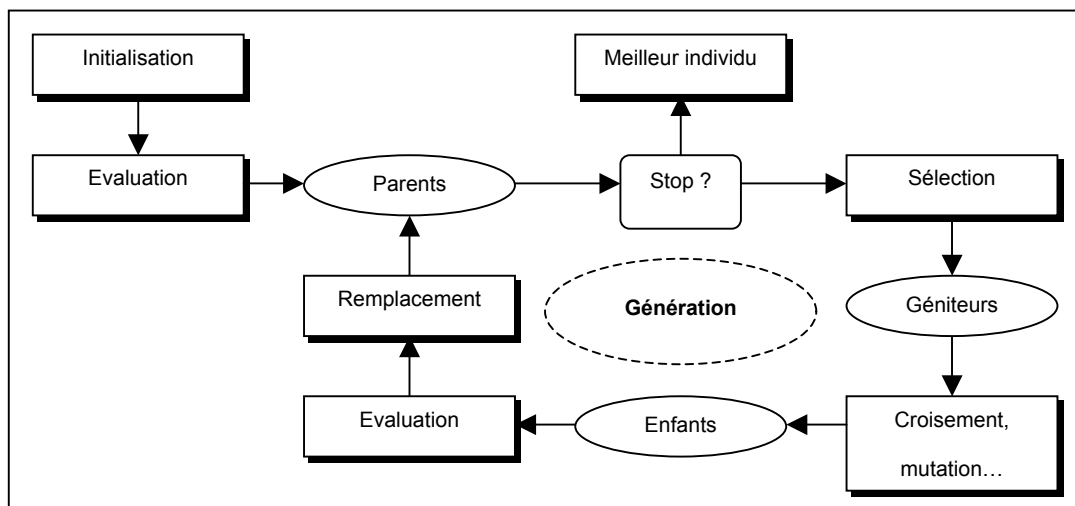


Figure 4 : Fonctionnement d'un algorithme génétique

Représentation des individus

Il existe deux approches différentes pour extraire des règles en utilisant un AG : l'approche de Pittsburgh [GOLD89] et l'approche de Michigan [GOLD89]. La première consiste à coder plusieurs règles au sein d'un individu tandis que dans la seconde, une règle ne code qu'un seul individu. Le génome d'un individu, représente la conjonction de termes attribut-valeur qui correspond à l'antécédent d'une règle. Le codage est un codage de position qui consiste en une suite de gènes rangés dans le même ordre que les attributs du jeu de données. Chaque condition est codée par un gène et consiste en un triplet de la forme $(A_i \text{ op } V_{ij})$ où A_i est le $i^{\text{ème}}$ attribut de la table sur laquelle l'algorithme est appliqué. Le terme op est un des opérateurs "=", "<=" ou ">=". V_{ij} est une valeur du domaine de l'attribut A_i . Chaque gène comporte un champ booléen B_i qui indique si le $i^{\text{ème}}$ gène est actif ou pas, c'est à dire si la $i^{\text{ème}}$ condition est présente dans la règle. Ainsi, bien que les individus aient la même longueur de génotype, les règles associées à ceux-ci (leur phénotype) sont de longueur variable. La structure d'un génome est indiquée sur la figure 5, où m est le nombre total d'attributs de la table.

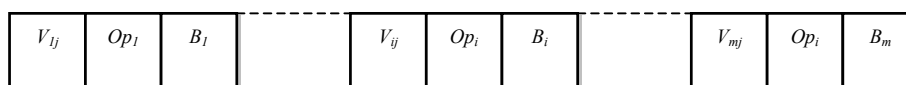


Figure 5 : génome d'un individu représentant l'antécédent d'une règle.

L'algorithme peut manipuler des variables entières, réelles ou encore catégorielles. Pour un attribut catégoriel, l'opérateur utilisé est soit l'opérateur "=" soit un des opérateurs "<=" et ">=" pour des variables entières ou réelles.

Fonction d'évaluation

Comme mentionné dans la section 2, notre but est de prendre en compte de façon simultanée d'une part le pouvoir prédictif de la règle et d'autre part son intérêt. Ceci est implémenté de façon simple dans l'AG par le biais de la fonction

d'évaluation. Les AG sont des algorithmes robustes à l'origine conçus pour faire de l'optimisation ; ceci est en partie réalisé au moyen de la fonction d'évaluation qui permet l'expression aisée de plusieurs critères de qualité. Dans notre cas, cette fonction doit être maximisée et est de façon générale de la forme suivante :

$$fitness(r) = \sum_{i=1}^k w_i \times mesure_i(r)$$

avec r une règle de la forme $r : A \rightarrow C$ et $\sum_{i=1}^k w_i = 1$

Nous nous limitons dans cet article au cas de deux mesures/critères qui peuvent être une mesure de précision et une mesure d'intérêt de la règle r ou deux mesures complémentaires comme la sensibilité et la spécificité.

La fonction d'évaluation est appliquée à un individu lors de l'initialisation de la population initiale ainsi que chaque fois qu'un individu subit une modification soit de la part de l'opérateur de croisement, soit de la part de celui de mutation. L'évaluation d'un individu est réalisée sur l'ensemble d'apprentissage.

Sélection et opérateurs génétiques

L'algorithme utilise la sélection par tournoi à deux individus qui consiste à choisir deux individus de façon aléatoire dans la population courante puis à comparer leurs valeurs de fitness. De façon imagée, celui qui a la meilleure fitness remporte ce tournoi et pourra éventuellement transmettre le matériel génétique à la génération suivante. Ceci est répété P fois, P étant la taille de la population. Les individus ainsi sélectionnés seront ensuite soumis aux opérateurs génétiques de croisement et de mutation.

L'algorithme utilise un croisement sur un seul site. Celui-ci consiste d'abord en la sélection d'un site unique de croisement et en l'échange entre les deux individus, du matériel génétique situé après ce site. Le site est choisi de façon à éviter d'engendrer des individus vides. La probabilité de croisement utilisée est 0.8 et est déterminée de façon empirique.

L'opérateur de mutation consiste à modifier principalement la longueur des règles. Soit il spécialise une règle en ajoutant une condition si tous les gènes ne sont pas actifs, soit il la généralise en supprimant une condition, si toutefois la longueur de l'antécédent de cette règle est strictement supérieure à un. L'ajout d'une condition est réalisée par génération aléatoire d'un opérateur de comparaison ainsi que d'une valeur appartenant au domaine de l'attribut associé au gène muté. Cet opérateur ainsi utilisé permet de généraliser une règle par la suppression d'une condition ou de spécialiser une règle par l'ajout d'une condition générée de façon aléatoire. L'opérateur de mutation est appliqué avec une probabilité de $1/m$, m étant la longueur du génotype.

Résultats expérimentaux

Nous avons expérimenté l'AG sur les jeux de données citées en section 3. Nous présentons quelques résultats significatifs obtenus sur les critères de *sensibilité*, *spécificité* et *RI*. Nous présentons ici des résultats obtenus sur les bases Dermatology et Vote. La base Dermatology contient 366 descriptions de six maladies de peau selon 33 attributs

catégoriels et I attribut entier. La base Vote contient 232 descriptions de candidats appartenant à deux classes politiques selon 17 attributs catégoriels.

La population est de taille fixe et contient 100 individus. Une exécution de l'algorithme consiste en une centaine de générations. L'AG est exécuté 10 fois pour optimiser la même fonction de fitness définie par un couple de poids. Les résultats obtenus à partir de l'algorithme génétique sont produits selon l'algorithme suivant :

```

algorithme test
  COUPLES = { (se, lift), (se, RI), (se, JMeasure), (se, sp), (se, conf), (Ts, lift), (Ts, RI), (Ts, JMeasure) }

début
  pour chaque base  $B_i$  faire
    pour chaque couple  $c_j \in$  COUPLES faire
      pour chaque valeur  $v_k$  de la classe de  $B_i$  faire
        pour chaque couple  $(w_1, w_2) \in \{0.1, 0.2, \dots, 0.9\}^2 / w_1 + w_2 = 1$  faire
          {
            exécuter 10 fois l'AG avec  $fitness = w_1 \times c_j.x + w_2 \times c_j.y$  et  $v_k$  à prédire
            //  $x$  et  $y$  étant respectivement le premier et le second critère dans  $c_j$ 
            garder le/les meilleur(s) individu(s) de la population finale
          }

```

Remarque sur l'algorithme ci-dessus : nous avons choisi de façon empirique 9 couples de poids différents afin de réduire les durées d'exécution.

Nous avons comparé les résultats obtenus par l'algorithme ci-dessus avec ceux obtenus grâce aux deux algorithmes utilisés dans la section précédente : *J4.8* qui est une extension de l'algorithme *C4.5* et une extension de l'algorithme *Apriori* de *Weka* [WEKA, WITT99]. L'AG utilise le même formalisme que celui utilisé par *Weka* pour décrire les attributs d'une base (réduite en fait à une table) pour renseigner les algorithmes sur les types des attributs en présence. Nous étendons ce formalisme afin de permettre l'emploi des opérateurs " \leq " et " \geq " avec les attributs numériques.

Nous avons représenté graphiquement sur les figures 6, 7 et 8, les règles obtenues par les algorithmes *J4.8* et *Apriori* de *Weka* de la même manière que dans la section 3 et d'autre part par l'AG pour les différents couples de critères. Afin de comparer les règles extraites par la méthode scalaire avec celles trouvées par les méthodes standards. Les règles aléatoirement générées sont également représentées. La figure 6 met en évidence l'émergence de bonnes règles extraites par l'AG ignorées par les algorithmes standards. La comparaison avec la figure 1 montre non seulement l'efficacité de la méthode évolutionnaire dans le parcours l'espace de recherche et l'existence de solutions parmi les meilleures simultanément sur les deux critères *Se* et *RI*.

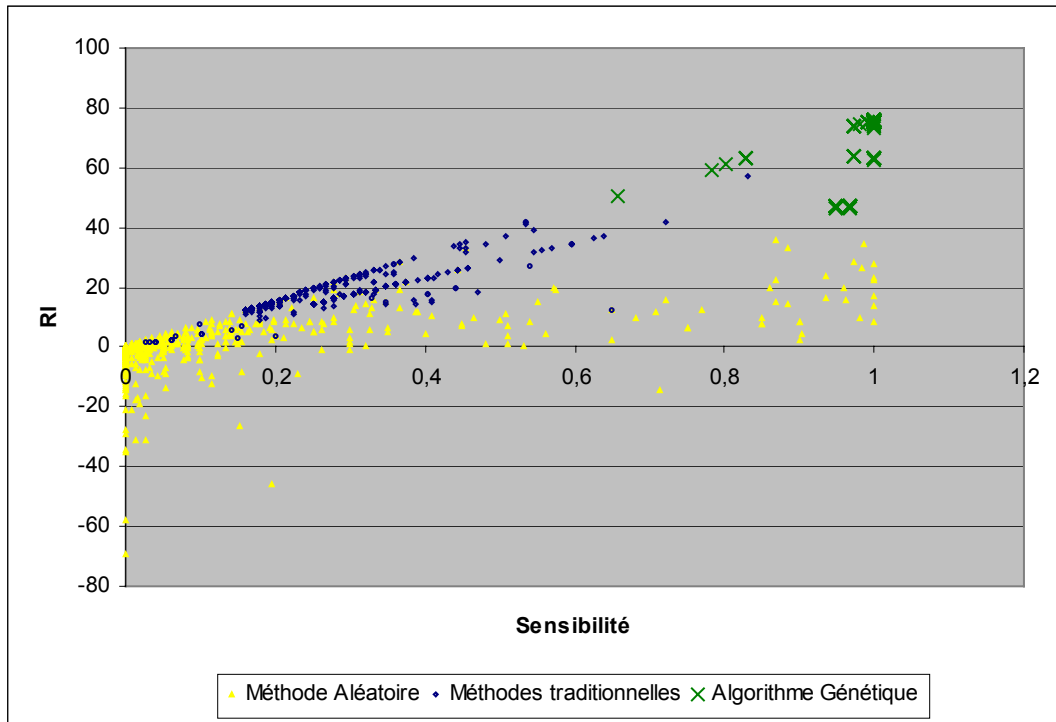


Figure 6 : Se versus RI sur la base Dermatology

Sur ce graphes, les triangles représentent les règles générées de façon aléatoire. Les losanges correspondent aux règles trouvées par les méthodes standards .

La figure 7 illustre également la supériorité de l'AG qui est capable d'extraire des règles optimisant simultanément les deux critères Se et Sp et montrant que sur cet exemple, les deux critères ne sont pas antagonistes.

La figure 8 illustre l'intérêt de la méthode scalaire qui permet d'approcher les solutions représentant des compromis sur l'ensemble des deux critères. En effet, les différentes exécutions de l'AG optimisant différentes sommes pondérées des deux facteurs découvrent des règles (représentées par des cercles rouges) qui se placent au voisinage de la frontière de Pareto. Ici, l'antagonisme des deux facteurs semble se confirmer ; l'AG ne permet pas d'extraire de meilleures solutions, mais il permet la mise en œuvre de la méthode scalaire qui approche la frontière de Pareto.

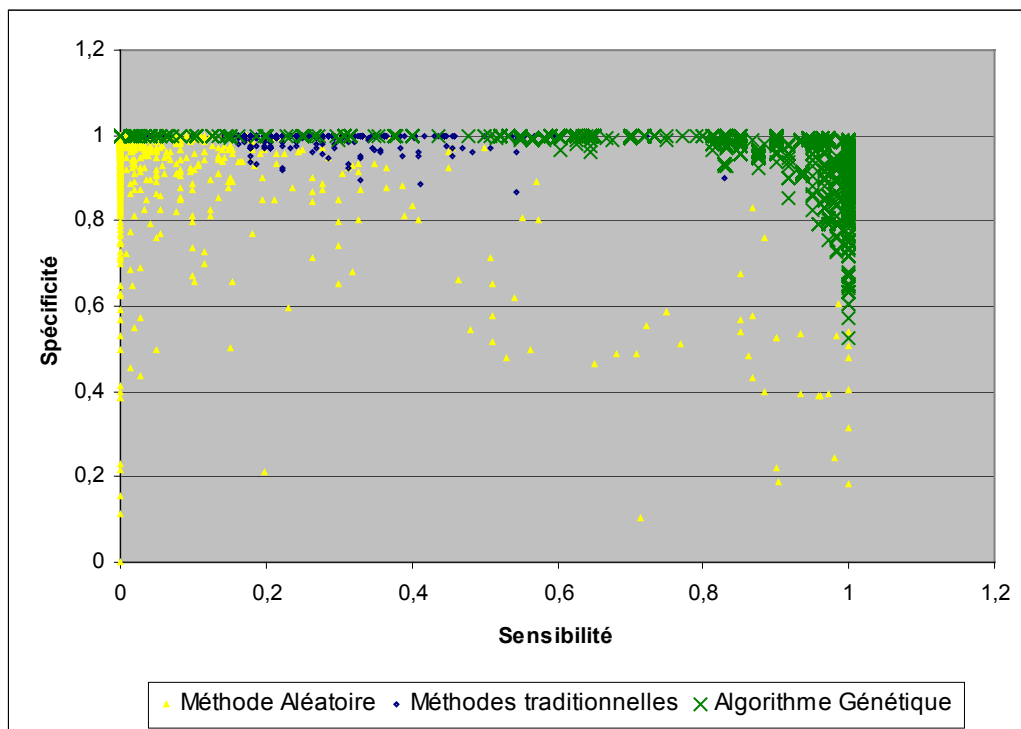


Figure 7 : Se versus Sp sur la base Dermatology

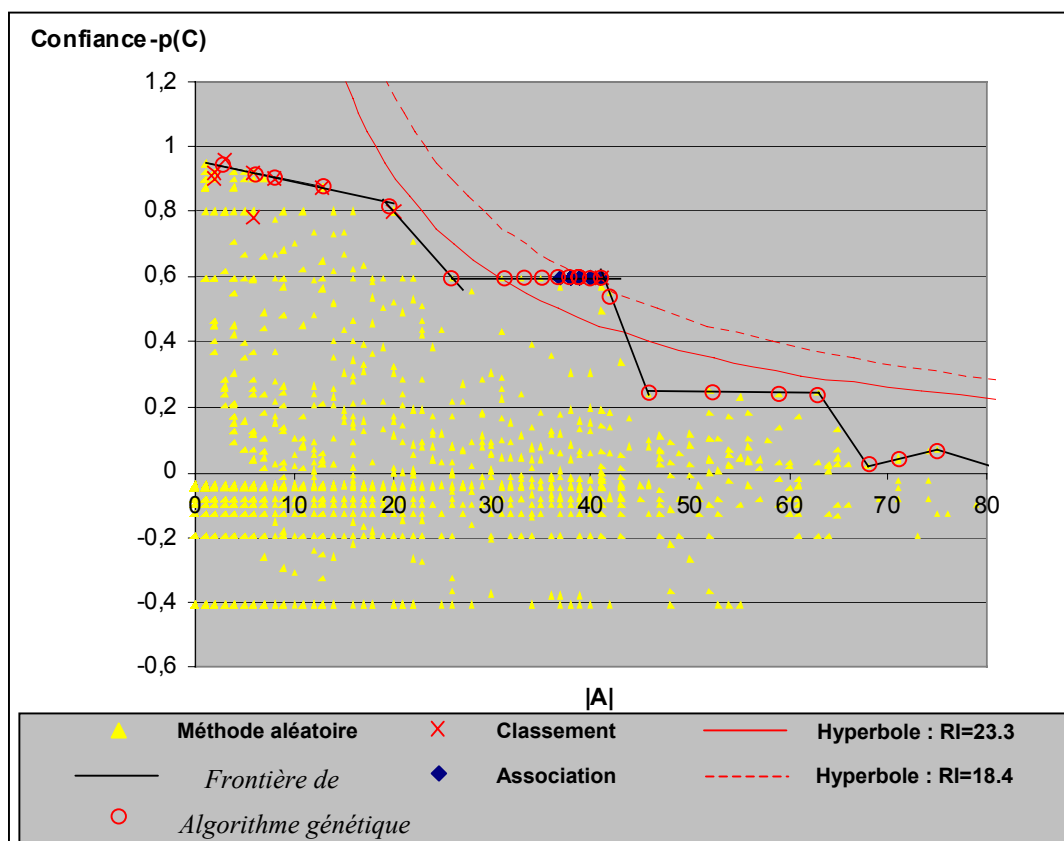


Figure 8 : $|A|$ versus $conf(r) - p(C)$ sur la base Zoo

4. Conclusion

Dans cet article, nous avons proposé une approche multi-critères permettant d'extraire des règles répondant à plusieurs critères de qualité. En effet, pour répondre à l'objectif central de la fouille de données, il est important de disposer de moyen pour sélectionner des règles utiles, c'est à dire intéressantes, surprenantes, exprimant des connaissances insoupçonnées, mais également des règles précises, compréhensibles ou vérifiant d'autres contraintes particulières. Aussi, nous nous sommes intéressés, dans ce travail, à la façon de sélectionner les règles simultanément sur plusieurs critères. Une étude comparative de différentes mesures et des algorithmes d'extraction de règles nous a permis de mettre en évidence d'une part, les limites des algorithmes standards dans ce domaine et d'autre part, l'existence d'antagonismes entre certaines mesures. Ceci nous a conduit à rechercher une méthode d'optimisation multi-critères permettant de découvrir un ensemble de compromis en l'absence de solution optimale. Notre choix s'est porté sur un algorithme évolutionnaire pour ses capacités à parcourir des espaces de recherche importants et également pour la souplesse apportée par la fonction d'évaluation. Nous avons ainsi mis en oeuvre une méthode scalaire qui a fait émerger des solutions ignorées par les algorithmes standards. Cependant ce type de méthode ne permet pas toujours d'explorer de manière exhaustive les points de la frontière de Pareto. Nos travaux futurs s'orientent vers une méthode multi-critères plus robuste.

5. Références

- [AGRA93] R. Agrawal, T. Imielinski, A. Swani. Mining Association Rules between sets of items in large databases. *Proc. Int.Conf. on Management of Data, SIGMOD*, 1993.
- [AGRA94] R. Agrawal et A. Srikant. *Fast algorithms for mining association rules*. Proc. VLDB'94, pp487-499.
- [FIDE00] M.V. Fidelis, H.S. Lopes et A.A. Freitas. Discovering comprehensible classification rules with a genetic algorithm. *Proc. Congress on Evolutionary Computation (CEC-2000)*, pp 805-810. La Jolla, CA, USA. Juillet 2000.
- [FREI99] A. A. Freitas. On rule interestingness measures. *Knowledge-Based Systems journal* 12 (5-6), pages 309-315. Octobre 1999.
- [FREI00] A. A. Freitas. Understanding the crucial differences between classification and discovery of association rules - a position paper. *ACM SIGKDD Explorations (ACM 2000)*, 2(1), pp 65-69. 2000.
- [GOLD89] D. E. Goldberg. Genetic algorithms in search, optimization and machine learning. *Addison Wesley*. 1989.
- [HILD00] R. J. Hilderman et H. J. Hamilton. Principles for Mining Summaries Using Objective Measures of Interestingness. *Proc. of the 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00)*, Vancouver, Canada, pages 72-81. Novembre 2000.
- [IBM96] International Business Machines, IBM intelligent Miner, User's guide, version 1, release 1, 1996.

- [LALL02] S. Lallich et O. Teytaud. Evaluation et validation de l'intérêt des règles d'association. *Rapport de recherche pour le groupe de travail GafoQualité de l'action spécifique STIC fouille de bases de données, ERIC*, Université Lyon 2. 2002.
- [LENC02] P. Lenca, P. Meyer, B. Vaillant et P. Picouet. Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – Modélisation des préférences de l'utilisateur. *Rapport de recherche pour le groupe de travail GafoQualité de l'action spécifique STIC fouille de bases de données, département IASC, ENST Bretagne*. 2002.
- [NODA99] E. Noda, A.A. Freitas et H.S. Lopes. Discovering interesting prediction rules with a genetic algorithm. *Proc. Congress on Evolutionary Computation (CEC-99)*, pages 1322-1329. Washington D.C., USA. Juillet 1999.
- [PIAT91] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. *G. Piatetsky – Shapiro et W. J. Frawley, editors, Knowledge Discovery in Databases*. MIT Press. 1991.
- [QUIN93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [SEBA88] M. Sebag, M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases.. *Proc. of the European Knowledge Acquisition Workshop, EKAW'88*. 1988.
- [SMIT91] P. Smyth et H. M. Goodman. Rule induction using information theory. *Knowledge Discovery in Databases. G. Piatetsky – Shapiro et W. J. Frawley, editors, MIT Press*. 1991.
- [UCI] UCI Machine Learning, <ftp.ics.uci.edu/pub/machine-learning-databases>.
- [WEKA] Weka Software, Université de Waikato, Nouvelle Zélande, <http://www.cs.waikato.ac.nz/ml/weka/>
- [WITT99] I. H. Witten et E. Frank. Data Mining – Practical machine learning tools and techniques with Java implementations. *Morgan Kaufmann Publishers*. 1999.