# Nouvelle méthode d'extraction de règles de classification multi-labels

Abdelhamid Zemirline, Laurent Lecornu, Basel Solaiman

# Nouvelle méthode d'extraction de règles de classification multi-labels

Abdelhamid Zemirline, Laurent Lecornu, Basel Solaiman

Departement ITI, ENST Bretagne
abdelhamid.zemirline@enst-bretagne.fr,
laurent.lecornu@enst-bretagne.fr,
basel.solaiman@enst-bretagne.fr

**Résumé.** Les études concernant les méthodes d'extraction de règle de classification proposent de classer une instance, le plus souvent en se basant sur une seule règle, sans prendre en considération les règles du même type ayant un label de classe différent (par ex. : $X \rightarrow a$, $X \rightarrow b$). Dans certains domaines, comme par exemple le domaine médical, il est important de connaître toutes les conséquences induites par une règle donnée. Dans cette étude, nous proposons une nouvelle méthode d'extractions de règle de classification multi-labels. Elle se caractérise par sa rapidité, ceci est dû à la procédure utilisée pour l'extraction des motifs fréquents par le biais des ensembles flous. Elle se caractérise également par l'extraction des règles en fonction de leur label, les règles sont extraites à partir des ensembles de données appartenant à une même classe et non à partir de l'ensemble global des données. Cette approche permet d'extraire des règles spécifiques à chaque classe et d'extraire de nouvelles catégories de règle d'exception. La méthode développée présente une grande précision et ses performances sont comparables à celles d'autres méthodes de classification de références.

## 1 Introduction

Dans le domaine médical, les systèmes d'aide au diagnostic ont pour objectif de donner un ou plusieurs diagnostics au praticien, ceci induit l'utilisation de règle multi-label, c-à-d des règles à plusieurs conséquences. Il existe peu d'études sur les méthodes qui génèrent des règles multi-label (Thabtah et al. (2004, 2006)). Par contre, il y en a un grand nombre d'études qui portent sur des méthodes de génération de règle de classification avec un seul label (Li et al. (2001); Liu et al. (1998)). Dans cette étude, nous proposons une méthode d'extraction de règles de classification multi-labels, qui s'appuie sur une nouvelle approche d'extraction. Cette approche consiste à extraire les règles à partir d'un ensemble d'instances appartenant à une même classe. Une instance est un ensemble d'attribut de type alpha-numérique, un de ces attributs indique l'appartenance de l'instance à une classe. Cette approche permet d'obtenir des règles spécifiques pour chaque classe que contient la base d'apprentissage. Contrairement aux méthodes classiques qui génèrent des règles globales, extraites à partir de l'ensemble de la base d'apprentissage. De plus, notre approche permet d'extraire des règles dites d'exception.

Par exemple dans le domaine médical, ce type de règle permet aux systèmes d'aide au diagnostic d'identifier des pathologies qui sont difficiles à diagnostiquer par le praticien en raison de leur faible fréquence d'apparition et de permettre également de distinguer une pathologie qui présente un symptôme similaire à une pathologie courante. Dans cette étude, nous commençons par définir les règles de classification. Par la suite, nous présentons notre algorithme et sa description. Avant de conclure, notre méthode est comparée à d'autres algorithmes de classification.

## 2 Règles d'association de classes

La méthode des règles d'association de classes (Li et al. (2001); Liu et al. (1998)) comprend l'extraction des règles et la classification à l'aide de ces règles. Cette méthode définit un classifieur du type C4.5 de Quinlan (1993). Ce type de méthodes dérive des méthodes d'extraction de règle d'association de Agrawal et Srikant (1994). Le problème de règle d'association de classes se définit de façon suivante : soit $D$ l'ensemble de données qui est composé de $N$ cas décrits par un ensemble d'items. Soit $I$ l'ensemble des items qui décrivent les cas de l'ensemble $D$. Chaque élément de $D$ est associé à un élément de l'ensemble $C$ des classes. Les règles d'association de classe sont implémentées de la façon suivante $X \rightarrow c$ où $X \subseteq I$ et $c \in C$. $X$ est un itemset (c-à-d ensemble d'items). La règle $X \rightarrow c$ est extraite de l'ensemble $D$ avec une confiance ($Conf$). $Conf$ est le rapport entre le nombre de cas de $D$ contenant $X$ et appartenant à $c$ et le nombre de cas de $D$ contenant $X$. La règle $X \rightarrow c$ a un support ($Supp$) dans $D$. $Supp$ est le rapport entre le nombre de cas de $D$ contenant $X$ et appartenant à $c$ et le nombre de l'ensemble des cas de $D$. L'objectif de cette approche est de générer des règles d'association de classes qui ont un support et une confiance supérieurs à un certain seuil ($minSupp$, $minConf$) et de mettre en place un classificateur efficace à partir de ces règles.

Dans un problème multi-labels, on retrouve la même définition que dans le problème de règles d'association de classe. Cependant, pour une instance $d \in D$, il peut être assigné un ensemble de classes $c^1, c^2, \cdots, c^k$ pour $c^i \in C$. Ceci est représenté de la façon suivante : $(d, (c^1, c^2, \cdots, c^k))$ où $(c^1, c^2, \cdots, c^k)$ est une liste ordonnée de classes pour l'instance $d$. L'ordonnancement de la liste se fait à l'aide de la base de connaissance déduite du classificateur.

## 3 Algorithme MCCAR

Dans cette partie, nous présentons un nouvel algorithme d'extraction de règles d'association de classes que nous avons appelé l'algorithme MCCAR (Multi-label Classification based on Class Association Rule). Les particularités de notre algorithme sont les suivantes :

- Extraction de règles à partir d'un sous-ensemble d'instances appartenant à une même classe et non pas sur l'ensemble des instances,
- classification des instances en se basant sur plusieurs règles,
- extraction de règles d'exception par deux procédures (la procédure "multi-seuil" et la procédure "générer-supprimer") §3.2,
- génération de règles multi-labels.

Hussain, F., H. Liu, E. Suzuki, et H. Lu (2000). Exception rule mining with a relative interestingness measure. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 86–97.

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 *KDD*, pp. 80–86.

Liu, B., W. Hsu, et Y. Ma (1999a). Mining association rules with multiple minimum supports. In *Knowledge Discovery and Data Mining*, pp. 337–341.

Liu, H., H. Lu, L. Feng, et F. Hussain (1999b). Efficient search of reliable exceptions. In *PAKDD*, pp. 194–203.

Merz, C. et P. Murphy (1996). *UCI repository of machine learning databases*. Department of Information and Computer Science, University of California, Irvine,.

Quinlan, J.-R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

Quinlan, J. R. et R. M. Cameron-Jones (1993). FOIL : A midterm report. In *Machine Learning : ECML-93, European Conference on Machine Learning, Proceedings*, Volume 667, pp. 3–20. Springer-Verlag.

Thabtah, F. A., P. I. Cowling, et Y. Peng (2004). Mmac : A new multi-class, multi-label associative classification approach. In *ICDM*, pp. 217–224.

Thabtah, F. A., P. I. Cowling, et Y. Peng (2006). Multiple labels associative classification. *Knowl. Inf. Syst. 9*(1), 109–129.

Weka (2006). *weka : data mining software in java*. http ://www.cs.waikato.ac.nz/ml/weka.

Yin, X. et J. Han (2003). CPAR : Classification based on predictive association rules. In *Proceedings of 2003 SIAM International Conference on Data Mining*, San Fransisco, CA.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning - ii. *Information Sciences 8*, 301–357.

Zaki, M. J., S. Parthasarathy, M. Ogihara, et W. Li (1997). New algorithms for fast discovery of association rules. Technical report, Rochester, NY, USA.

## Summary

In this paper, we propose a new algorithm of multi-labels classification MCCAR (Multi-labels Classification based on Class Association Rule). Our approach is based on producing association rules and knowledge base informing about the accuracy of rules for the classification. Our algorithm finds reliable exceptions with a simple and efficient approach. A rule has a multi-label class. To rank class labels for a rule, we use the degree of membership of rule to a class. This new approach enables us to evaluate the importance of rules among themselves.

La majorité des méthodes d'extraction de règles d'association de classes se décomposent en trois phases. La première phase consiste à générer les règles, la seconde à les évaluer et la dernière phase à utiliser les règles générées sur les instances de la base de test. Dans notre approche, nous avons quatre phases. La première phase regroupe les instances dans des sous-ensembles en fonction de leur classe et définit le support minimum pour chaque sous-ensemble. Cela permet par la suite de découvrir les items fréquents et de générer des règles dites communes et d'exception. Dans notre approche, les règles sont extraites à partir des ensembles de données appartenant à une même classe et non à partir de l'ensemble global des données. Ceci nous amène à redéfinir le calcul du support, $Supp$ : le rapport entre le nombre de cas de $D$ contenant $X$ et appartenant à $c$ et le nombre de l'ensemble des cas de $D$ contenant $c$. La procédure permettant de générer les seuils est expliquée dans le §3.1. La génération des règles d'exception est abordée dans le §3.2. La seconde phase consiste à supprimer les instances de l'ensemble de la base d'apprentissage, qui sont couvertes par les règles générées précédemment et à générer les règles d'exception à l'aide de la procédure "générer-supprimer" (generate-and-remove). La procédure "générer-supprimer" est décrite dans le §3.2. Dans la troisième phase, les règles générées dans les phases précédentes sont évaluées afin de permettre leur exploitation par la procédure de classification. La procédure de cette phase est décrite dans §3.3. Dans la dernière phase, toutes les règles générées précédemment sont regroupées afin de produire des règles multi-labels (décrite dans §3.4). Par la suite, une procédure de classification des instances de la base de test est réalisée afin d'estimer les performances de l'algorithme. Pour réaliser cette procédure, un classificateur est mis en place. Ce classificateur manipule les règles afin de prédire la classe de l'instance à prédire (§3.5).

### 3.1 Sélection des itemsets fréquents et génération de règles

Plusieurs approches de sélection des itemsets fréquents et de génération de règles ont été proposées par Agrawal et Srikant (1994); Liu et al. (1998); Li et al. (2001); Yin et Han (2003); Thabtah et al. (2004); Zaki et al. (1997); Han et al. (2000); Quinlan et Cameron-Jones (1993).

Dans ces travaux, nous proposons une nouvelle approche afin de découvrir les itemsets fréquents et de générer des règles. Cette approche génère des règles dites locales, c-à-d que ces règles sont générées à partir d'un sous-ensemble d'instances qui appartiennent à une même classe. Cela permet d'avoir des règles spécifiques pour chaque classe, contrairement aux autres méthodes qui génèrent leurs règles à partir de l'ensemble des instances. Dans notre méthode, nous nous inspirons de la méthode d'intersection proposée par Zaki et al. (1997) afin de parcourir une seule fois l'ensemble des instances pour calculer les supports des règles générées.

Notre approche divise la base d'apprentissage $T$ en plusieurs sous-ensembles qui regroupent les instances appartenant à une seule et même classe. Chacun d'entre eux est parcouru une seule fois (grâce à la méthode d'intersection) afin de calculer la fréquence d'apparition des items. Pour chaque classe, il est défini une fonction d'appartenance en fonction des fréquences des items.

Ces fonctions d'appartenance classent les items en deux sous-ensembles : les items qui appartiennent au sous-ensemble *rare* et les autres au sous-ensemble *fréquent*. Nous faisons appel à la théorie des ensembles flous de Zadeh (1975) afin de définir les fonctions d'appartenance aux sous-ensembles *rare* et *fréquent*. La plupart des méthodes de règles d'association ont leur support seuil qui est fixé au préalable, ceci peut présenter certains désavantages tels que :

– Aucun item n'a son support qui dépasse le support minimum.

– Certains items ne sont pas pris en compte malgré que leur support sont voisins au support minimal.

Les termes "rares" et "fréquents" définissent une variable linguistique, celle-ci est caractérisée par le quintuplet $(x, T(x), U, G, M)$ (Zadeh (1975)) où :

– $x$ est le nom de la variable linguistique, ici c'est la fréquence ;
– $T(x)$ est l'ensemble des termes associés à la valeur linguistique, ici cet ensemble est : *{ Rare, Fréquent}* ;
– $U$ est l'espace de définition, notre domaine de définition $U = [f_{min}, f_{max}]$ ( $f_{min}$ correspond à la fréquence minimale d'apparition d'un item qui est normalisée par la fréquence maximale ($max_{freq}$) d'apparition d'un item pour la même classe, $f_{max}$ correspond à la fréquence maximale ($max_{freq}$) normalisée, elle équivaut à 1) ;
– $G$ est une règle syntaxique pour la génération des noms de valeur de x ;
– $M$ est une règle sémantique pour associer à chaque valeur sa signification.

Les termes de $T(x)$ sont caractérisés par des ensembles flous définis par les fonctions d'appartenance suivantes :

– $F$ est l'ensemble des centroïdes des ensembles flous obtenus par l'algorithme de fuzzy c-means (FCM) Bezdek (1981), tel que $F$ se présente de la façon suivante $\{f_{rare}, f_{fréquent}\}$.
– $\mu_{i,rare}$ correspond à la fonction d'appartenance au terme linguistique *rare* pour la classe $i$ et prend comme argument la fréquence normalisée $f$ d'un item donné.

$$\mu_{i,rare}(f) = \begin{cases} 1 & \text{si } f \leq f_{rare} \\ 1 - (f - f_{rare})/(f_{fréquent - f_{rare}}) & \text{si } f_{rare} < f \leq f_{fréquent} \\ 0 & \text{sinon.} \end{cases}$$

– $\mu_{i,fréquent}$ correspond à la fonction d'appartenance au terme linguistique *fréquent* pour la classe $i$ et prend comme argument la fréquence normalisée $f$ d'un item donné.

$$\mu_{i,fréquent}(f) = \begin{cases} 0 & \text{si } f \leq f_{rare} \\ 1 - (f - f_{fréquent})/(f_{fréquent - f_{rare}}) & \text{si } f_{rare} < f \leq f_{fréquent} \\ 1 & \text{sinon.} \end{cases}$$

Pour notre méthode, les items sont considérés comme *fréquent* si le degré d'appartenance à l'ensemble *fréquent* est supérieur au degré d'appartenance au sous-ensemble *rare*. Deux seuils subjectifs sont déduits : premier support $sup_1$ et le second support $sup_2$. La valeur de $sup_1$ correspond à l'abscisse de l'intersection des fonctions $\mu_{rare}$ et $\mu_{fréquent}$ pour une classe donnée, multipliée par $max_{freq}$. La valeur de $sup_1$ correspond à la valeur du centroïde $f_{rare}$ multiplié par $max_{freq}$. Grâce à cette méthode, des seuils supports subjectifs spécifiques à une classe donnée ont été définis. Tous les items qui ont leur degré d'appartenance à la valeur linguistique *fréquent* égal à 1 sont considérés comme des items fréquents. Les items qui ont leur degré d'appartenance à la valeur linguistique *fréquent* inférieur à 1 mais supérieur au degré d'appartenance à la valeur linguistique *rare* sont considérés comme items d'exception. La combinaison des items fréquents et d'exception va générer des règles dites communes ou d'exception selon les supports de ces règles, ceci sera précisé par la suite.

### 3.2 Extraction des règles dites d'exception

Tout d'abord, les règles d'exception sont définies comme étant des règles qui contredisent la croyance commune. Elles sont souvent inconnues ou négligées. Cependant, ce type de règles Certaines classes ne peuvent pas déduire les items qui leur sont réellement fréquents. Il y a une autre cause à cet écart important, elle est du même type que celle décrite pour la base d'apprentissage *tic-tac*.

## 5 Conclusion

Dans cette étude, nous avons proposé une nouvelle méthode d'extraction de règles multi-labels. Cette approche a vocation à être intégrée dans un système d'aide au diagnostic. Elle se compose de plusieurs caractéristiques qui ne sont pas présentes dans les méthodes d'extraction de règles de classification traditionnelles telles que : (1) couverture de l'ensemble des instances, (2) un seul parcours de la base d'apprentissage, (3) extraction de règles d'exception et (4) utilisation des degrés d'appartenance pour l'évaluation des règles. Ces caractéristiques associées au fait que les performances obtenues sont comparables aux autres méthodes de règles de classification de références, montrent la puissance de l'approche proposée.
Dans la méthode présentée, nous faisons appel aux fonctions d'appartenance afin d'évaluer les règles de classification. Par ce biais, nous pouvons facilement intégrer un opérateur de fusion qui peut par la suite être introduit dans des systèmes d'aide au diagnostic. Ces opérateurs de fusion auront la tâche d'intégrer dans un seul système des règles provenant de différentes sources afin de générer un classificateur ayant des performances très élevées.
L'algorithme MCCAR présente une nouvelle approche d'extraction de règles multi-labels et en particulier l'extraction de règles d'exception. Pour les règles d'exception, il serait intéressant par la suite de mettre une procédure de filtrage, afin de mettre à l'écart certaines règles qui sont considérées comme du bruit pour un domaine donné et mettre en avant celles qui peuvent avoir un grand intérêt dans la prise de décision pour ce même domaine.

## Références


Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499. Morgan Kaufmann.

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. *Plenum Press*.

CBA (1998). *CBA : The DM-II system*. http ://www.comp.nus.edu.sg/ dm2/.

Chauvin, S. (1995). *Thèse : Evaluation des théories de la décision appliquées à la fusion de capteurs en image satellitaire*. Ph. D. thesis, Thèse de Doctorat d'Université, Nantes.

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 115–123. Morgan Kaufmann.

Frank, E. et I. H. Witten (1998). Generating accurate rule sets without global optimization. In *Proc. 15th International Conf. on Machine Learning*, pp. 144–151. Morgan Kaufmann, San Francisco, CA.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In W. Chen, J. Naughton, et P. A. Bernstein (Eds.), *2000 ACM SIGMOD Intl. Conference on Management of Data*, pp. 1–12. ACM Press.

| Dataset | RIPPER | PART | CBA | MCCAR |
|---|---|---|---|---|
| anneal | 98.1 | 98.32 | 93.43 | 96.5 |
| austra | 86.81 | 85.79 | 84.78 | 83.15 |
| autos | 73.17 | 75.6 | 64.25 | 86.3 |
| breast | 95.42 | 96.28 | 96.86 | 97.14 |
| cleve | 82.17 | 81.18 | 75.86 | 79.64 |
| crx | 86.08 | 85.5 | 85.07 | 83.5 |
| diabetes | 76.82 | 77.21 | 74.42 | 74.60 |
| german | 72.4 | 73.4 | 70.97 | 71.2 |
| glass | 72.42 | 73.81 | 65.76 | 64.21 |
| heart | 84.07 | 82.22 | 81.8 | 81.11 |
| hepati | 79.35 | 81.93 | 84.97 | 85.06 |
| horse | 84.78 | 84.51 | 82.86 | 80.95 |
| hypo | 99.08 | 99.24 | 97.43 | 95.25 |
| iono | 90.88 | 91.73 | 95.73 | 91.15 |
| iris | 94 | 95.33 | 93.25 | 96 |
| labor | 84.12 | 84.21 | 94.99 | 94.66 |
| led7 | 69.7 | 73.56 | 67.01 | 72.93 |
| lymph | 77.7 | 80.4 | 83.4 | 71.78 |
| pima | 66.14 | 65.1 | 76.65 | 77.59 |
| sick | 93.77 | 93.74 | 96.07 | 94.87 |
| sonar | 80.76 | 75.96 | 76.88 | 76.91 |
| tic-tac | 97.8 | 94.25 | 99.6 | 82.24 |
| vehicle | 68.32 | 70.56 | 60.22 | 65.1 |
| waveform | 76 | 76,66 | 73.03 | 73,93 |
| wine | 95.5 | 94.38 | 96.66 | 96.63 |
| zoo | 86.13 | 92.07 | 92.34 | 91.36 |

TAB. 1 – *Précision de la classification de différentes méthodes*

PART. L'algorithme CBA a été exécuté à l'aide d'une application qui a été fournie par CBA (1998). Pour cette expérience, nous fixons la valeur de la confiance minimale (minConf) à 0,3 et la valeur du support minimale (minSup) à 0,03 pour CBA, RIPPER et PART.

Notre système présente des résultats de taux de fiabilité très intéressants. Les résultats obtenus dans le tableau 1 montrent que notre méthode présente le même niveau de précision que les méthodes de classification de référence. Les cas où notre méthode présente des résultats inférieurs aux méthodes de références, l'écart est en moyenne pas plus de 2 points, sauf pour les bases d'apprentissage *lymph* et *tic-tac*. Pour la base d'apprentissage *tic-tac*, la mauvaise estimation que notre méthode nous renvoie est due au fait que les valeurs des attributs ont une répartition équiprobable dans l'ensemble des cas des différentes classes. Ce qui implique que notre méthode génère les même règles pour les différentes classes et ceci induit des ambiguïtés dans la classification de certaines instances. Pour la base d'apprentissage *lymph*, l'écart important de mauvaise estimation s'explique en partie par la petite taille de la base d'apprentissage.

permet de couvrir des cas rares. Une règle commune présente un phénomène commun avec un support et une confiance élevés. La règle d'exception contredit certaines règles dites communes et a un faible support. Cependant, elle possède une valeur de confiance aussi élevée que pour les règles dites communes.

Il existe deux types de méthodes qui permettent la découverte des règles d'exception à partir d'ensemble de règles d'association (Hussain et al. (2000); Liu et al. (1999b,a)) : la méthode à multi-supports et la méthode générer-supprimer. La méthode à multi-supports consiste à introduire deux supports minimaux. Les règles ayant leurs items entre ces deux seuils sont les règles d'exception. La méthode générer-supprimer consiste à générer un ensemble de règles $R$ à partir d'une base d'apprentissage $T$ à l'aide d'un algorithme d'induction, puis à supprimer les objets $T'$ qui sont couverts par les règles $R$ de $T$. Un nouvel ensemble de règles $R'$ est généré à partir de $T - T'$ à l'aide du même algorithme d'induction. Ce processus est répété jusqu'à qu'il n'y ait plus d'élément dans la base d'apprentissage.

Les règles de classification s'écrivent de la façon suivante : $X \rightarrow c$ ; ceci implique une légère différence dans la définition des règles d'exception, d'où nous proposons une nouvelle définition de règles d'exception. Rappelons rapidement que l'objectif des règles de classification est de définir un classificateur afin de prédire l'appartenance à une classe de certains nouveaux objets. Les règles d'exception (pour les règles de classification) sont des règles qui indiquent l'appartenance d'un item à une classe en contradiction avec une règle qui a un poids plus fort. Elle couvre des instances qui sont inhabituelles ou ambiguës. Dans notre étude, nous définissons trois structures de règles d'exception :

*Unique règle d'exception* : l'itemset $X$ de ce type de règle n'est présent que dans cette règle. La confiance pour cette règle est égale à 1. Les items qui composent $X$ ne sont pas fréquents et la règle est générée par la méthode générer-supprimer.

$X \rightarrow c$ - règle d'exception
(support faible, confiance = 1)
$X \rightarrow \neg c$ - $\emptyset$

*Pseudo règle d'exception* : l'itemset $X$ de ce type de règle est composé d'items non fréquents pour la classe $c$ et cet itemset $X$ apparait dans d'autres règles appartenant à des classes différentes de $c$.

$X \rightarrow c$ - règle d'exception
(support faible, confiance faible/élevée)
$X \rightarrow \neg c$ - règle commune
(support élevé, confiance élevée)

*Semi règle d'exception* : l'itemset $X$ de ce type de règle est composé en partie d'items qui ne sont pas fréquents pour la classe $c$. L'itemset $X$ peut composer d'autres règles d'exception pour des classes différentes de $c$.

$X \rightarrow c$ - règle d'exception
(support faible, confiance élevé)
$X \rightarrow \neg c$ - règle d'exception
(support faible, confiance faible)

Cette définition des règles d'exception permet à l'utilisateur d'évaluer les règles de faible poids d'une façon plus pragmatique et de les intégrer dans le processus de décision de la même façon que les règles dites communes. Dans notre méthode, nous combinons la méthode

générer-supprimer et la méthode à double seuils pour générer les règles d'exception. La méthode de double seuils est adaptée à notre approche car elle peut utiliser de manière aisée les fonctions d'appartenance définies dans la phase une. Cette adaptation consiste à définir des seuils de supports locaux, c-à-d chaque classe aura ses propres double supports. Dans notre algorithme, la méthode à double supports minimaux est appliquée dans la phase une et quant à la méthode générer-supprimer, elle est appelée dans la phase deux. Cette combinaison des deux méthodes permet d'extraire des règles d'exception de façon pertinente.

### 3.3 Evaluation des règles

Dans cette partie, nous allons parler d'évaluation des règles générées à partir d'un sous-ensemble d'instances spécifiques à une classe donnée. Cette évaluation permet par la suite de classer les règles par degré d'importance. Notre démarche consiste à évaluer les règles en fonction de leur importance dans la classe à laquelle elles appartiennent. Ceci est effectué par l'attribution de degrés d'appartenance des règles à la classe à laquelle elles appartiennent. La définition des fonctions d'appartenance s'inspire de la méthode de l'histogramme de fréquence normalisée proposée par Chauvin (1995). À partir des supports des règles d'une classe donnée, nous définissons un histogramme de fréquence normalisé par rapport à la fréquence la plus élevée. La raison qui nous motive à définir des fonctions d'appartenance pour évaluer les règles entre elles au lieu de prendre en considération la confiance de règle est la suivante : notre algorithme propose une classification à base de multi-règles, qui nécessite la combinaison de plusieurs règles afin de classer un objet. Or, la combinaison des degrés d'appartenance est plus significative que la combinaison des confiances de règles appartenant à une même classe.

### 3.4 Ordonnancement des règles multi-labels

Une règle multi-labels est le regroupement de règles ayant leur itemset similaire appartenant à des classes différentes d'où l'appellation règle multi-labels. Les différentes classes qui composent ce type de règles sont ordonnancées. $A \prec B$ signifie que la classe $A$ précède la classe $B$. La règle multi-labels $X \rightarrow < A, B >$ sous entend que $A \prec B$. Cela signifie que le degré d'appartenance de l'itemset $X$ à la classe $A$ est plus important que celui à la classe $B$. Si l'itemset $X$ a le même degré d'appartenance pour la classe A et B, c'est le nombre d'instances couvert par les règles $X \rightarrow A$ et $X \rightarrow B$ qui va déterminer l'ordonnancement entre ces deux classes.

### 3.5 Classification

La classification pour les méthodes de règles d'association de classes consiste à utiliser les règles générées afin de déterminer la classe d'une nouvelle instance. L'approche la plus utilisée est celle qui classe les règles selon leur confiance (Liu et al. (1998); Thabtah et al. (2004)). Cette approche attribue à l'instance la classe de la règle ayant la plus forte confiance. Si aucune règle n'est trouvée une classe par défaut est attribuée. Le classement des règles doit respecter certains critères, exemple de critères : soient $r^a$ et $r^b$ deux règles, $r^a$ précède $r^b$ :

1. si $r_a$ a une confiance supérieure à $r_b$, sinon
2. si $r_a$ et $r_b$ ont une même confiance mais $r_a$ a un support supérieur, sinon
3. si $r_a$ et $r_b$ ont leur confiance et leur support identiques mais $r_a$ doit être générée en premier.

La classification des règles se présente de la façon suivante : $< r_1, r_2, \cdots, r_n, c >$ où $r_i$ précède $r_{i+1}$ et $c$ est la classe par défaut.

Dans Yin et Han (2003) une approche est proposée qui utilise l'ensemble des meilleures règles qui couvrent l'instance à classer. Ils se sont basés sur :

- La récupération des règles qui couvrent l'instance à classer.
- La sélection des meilleurs règles.
- Le regroupement des règles selon leur classe et la mesure de leur combinaison.
- La sélection de la classe qui présente la meilleur mesure.

En effet, il y a été développé plusieurs méthodes de combinaison de règles appartenant à une même classe. Une d'elle consiste à calculer la moyenne des confiances des règles. Une autre consiste à utiliser la règle ayant la plus forte mesure comme représentative. Par exemple dans Li et al. (2001) la règle ayant la plus forte mesure est celle qui présente la valeur la plus élevée au test $\chi^2$.

Pour notre méthode, nous définissons la mesure utilisée en calculant la moyenne géométrique des degrés d'appartenance des meilleures règles à une classe donnée. L'utilisation de plusieurs règles permet d'augmenter la précision de prédiction. Cependant, s'appuyer sur une seule règle pour prédire une instance, c'est négliger toutes les autres règles qui couvrent cette même instance. De plus cela ne prend pas en considération des informations qui peuvent donner un poids de certitude ou d'incertitude dans la prédiction. Dans notre cas, nous prenons les meilleures règles de chaque groupe pour la prédiction d'une instance, cela nous évite de prendre les règles qui couvrent l'instance mais dont la confiance est faible et qui ne peuvent que diminuer fortement la mesure.

## 4 Evaluation de l'algorithme MCCAR

Dans cette partie, nous allons évaluer les performances de notre système, mais avant de commencer la présentation de l'expérience, nous précisons que son objectif n'est pas de montrer que notre approche est plus performante que d'autres mais de monter que cette approche a sa place parmi les méthodes de classification. Cependant, la vocation de notre approche ne se limite pas seulement à définir un classificateur mais à apporter de nouvelles connaissances (règles multi-labels, règles d'exception) qui permettent à un système d'aide au diagnostic d'être plus fiable, plus informatif et transparent pour l'utilisateur.

Nous mettons en place une étude comparative de notre méthode avec d'autres méthodes de classification de référence telles que RIPPER Cohen (1995), PART Frank et Witten (1998) et CBA Liu et al. (1998). Comme pour Cohen (1995); Frank et Witten (1998); Liu et al. (1998); Thabtah et al. (2004), 26 bases de UCI Machine Learning Repository Merz et Murphy (1996) sont utilisées comme base de références afin d'évaluer les différentes méthodes. La méthode de la validation croisée à 10 parties est utilisée sur ces 26 bases de cas. L'expérience consiste à prendre chaque base de cas et à la diviser en deux parties : une partie qui sera considérée comme base d'apprentissage et une autre comme base de test. Ces bases sont appliquées aux méthodes de classification citées ci-dessus et à notre méthode, et par la suite, nous comparons les taux de bonne estimation dans la classification des cas de la base de test des différentes méthodes. Nous avons fait appel au logiciel Weka (2006) afin d'exécuter les algorithmes RIPPER et