Un critère Bayésien pour évaluer la robustesse des règles de classification

Dominique Gay*, Marc Boullé*

*Orange Labs
2, avenue Pierre Marzin
F-22307, Lannion Cédex
{prenom.nom}@orange-ftgroup.com

Résumé. L'utilisation de règles de classification dans les modèles prédictifs a été très étudiée ces dernières années. La forme simple et interprétable des règles en font des motifs très populaires. Les classifieurs combinant des règles de classification intéressantes (selon une mesure d'intérêt) offrent de bonnes performances de prédictions. Cependant, les performances de ces classifieurs dépendent de la mesure d'intérêt (e.g., confiance, taux d'accroissement,...) et du seuillage (non-trivial) de cette mesure pour déterminer les règles pertinentes. De plus, il est facile de montrer que les règles extraites ne sont pas individuellement robustes. Dans cet article, nous proposons un nouveau critère pour évaluer la robustesse des règles de classification dans les données Booléennes. Notre critère est issu d'une approche Bayésienne : nous proposons une expression analytique de la probabilité d'une règle connaissant les données. Ainsi, les règles les plus probables sont robustes. Le critère Bayésien nous permet alors d'identifier (sans paramètre) les règles robustes parmi un ensemble de règles données.

1 Introduction

Les règles d'association (Agrawal et al., 1993) font certainement partie des motifs les plus étudiés en fouille de données. Dans les données binaires, une règle d'association est une expression de la forme $\pi : X \rightarrow Y$, où X (le corps) et Y (le conséquent) sont des sous-ensembles d'attributs Booléens. Intuitivement, la sémantique de π est : *"lorsqu'on a observé les attributs de X, alors souvent on a observé aussi les attributs de Y"*. Le principal intérêt d'une règle est son pouvoir d'inférence inductive. En effet, si maintenant on observe les attributs de X alors on va probablement aussi observer les attributs de Y. Lorsque Y est un attribut classe, on parle alors de règle de classification ($X \rightarrow c$). Les règles de classification semblent propices aux tâches de prédiction ; puisque si un objet est décrit par les attributs de X alors il est probablement de classe c. Les récentes avancées en extraction de motifs ont donné naissance à de nombreux classifieurs à base de règles (e.g. Liu et al. (1998) pour les pionniers ou Bringmann et al. (2009) pour une vue d'ensemble). Ces méthodes sont connues pour leur interprétabilité et sont performantes en terme de prédiction dans les tâches de classification supervisée. Toutefois, on peut identifier au moins deux verrous :

Le paramétrage. Le seuillage de la mesure d'intérêt utilisée est une étape cruciale et pour autant non-triviale. Le dilemme est bien connu : un seuil de fréquence minimum élevé génère moins de règles mais aussi un faible taux de couverture des données et souvent moins de pouvoir discriminant pour les classes du problème. D'un autre côté, un seuil de fréquence très bas génère un grand nombre de règles parmi lesquelles certaines (de faible fréquence) peuvent être erronées. Le même dilemme persiste pour le seuillage de mesures d'intérêt telles que la confiance (i.e. une estimation de la probabilité $P(c \mid X)$) ou le taux d'accroissement (qui permet d'identifier les motifs émergents, i.e., qui sont fréquents dans une classe de données et infréquents dans le reste de la base (Dong et Li, 1999)). En effet, des seuils élevés de confiance (ou de taux d'accroissement) génèrent un petit nombre de règles de classification presque pures qui sont rares (voire erronées si combinées avec un seuil de fréquence bas) alors que des seuils bas génèrent beaucoup de règles avec un intérêt limité. Ainsi, trouver un compromis entre seuil de fréquence et mesure d'intérêt n'est pas trivial.

L'instabilité des mesures d'intérêt. Bien que des sous-ensembles de règles permettent de bonnes prédictions, il est facile de montrer que des règles à forte confiance ou émergentes ne sont pas individuellement robustes. Dans la figure 1, nous comparons les valeurs de confiance (resp. de taux d'accroissement et de lift (Brin et al., 1997)) en apprentissage et en test de règles extraites de la base `UCI breast-w` (Asuncion et Newman, 2007). Nous voyons clairement que les valeurs de ces mesures sont instables entre phase d'apprentissage et test. Ainsi, une "bonne" règle selon ces mesures peut se révéler "mauvaise" en test. Ces mesures ne permettent donc pas d'identifier les règles robustes.

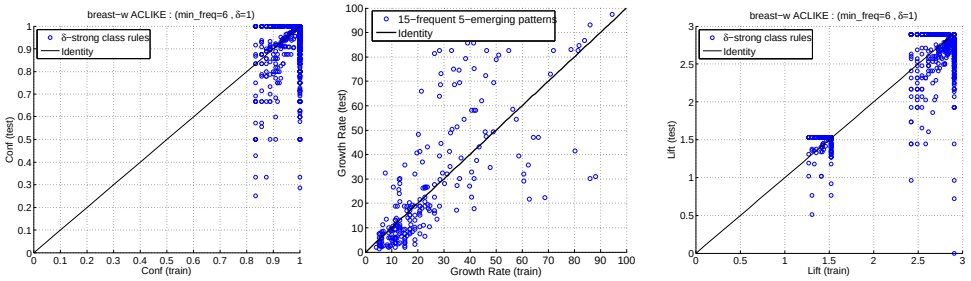

FIG. 1 – *Comparaison des valeurs de confiance, de taux d'accroissement et de lift en apprentissage et test pour des règles de classification : 50% apprentissage / 50% test sur la base* `breast-w`*. Lorsque* $lift \geq 2$*, alors il y a corrélation positive avec la classe.*

Dans cet article, nous proposons un critère Bayésien (issu de l'approche `MODL` de Boullé (2006)) qui nous permet d'identifier les règles robustes de manière naturelle et sans paramétrage aucun.

Le reste de l'article est organisé de la manière suivante : la section 2 pose le contexte et rappelle les concepts de base de l'approche `MODL`. Puis nous étendons l'approche `MODL` aux règles de classification et définissons notre critère Bayésien. La section 3 rapporte les expérimentations qui valident notre approche. En section 4, nous faisons le lien entre notre approche et plusieurs travaux connexes. Enfin, la section 5 conclut brièvement avant d'ouvrir sur d'autres perspectives de travail.

2 Des règles de classification aux règles MODL

Définitions. Soit $r = \{\mathcal{T}, \mathcal{I}, \mathcal{C}, \mathcal{R}\}$ une base de données binaires, où $\mathcal{T}$ est un ensemble d'objets, $\mathcal{I}$ un ensemble d'attributs Booléens, $\mathcal{C}$ un ensemble de classes et $\mathcal{R} : \mathcal{T} \times \mathcal{I} \mapsto \{0, 1\}$ une relation binaire telle que $\mathcal{R}(t, a) = 1$ indique que l'objet t contient l'attribut a. Chaque objet $t \in \mathcal{T}$ est labellisé par une unique classe $c \in \mathcal{C}$. Une *règle de classification* π dans r est une expression de la forme $\pi : X \rightarrow c$ où $X \subseteq \mathcal{I}$ est un itemset (i.e. un sous-ensemble d'attributs) et $c \in \mathcal{C}$ une classe. La *fréquence* d'un itemset dans r est $freq(X, r) = |\{t \in \mathcal{T} \mid \forall a \in X : \mathcal{R}(t, a)\}|$ et la fréquence de π est $freq(\pi, r) = freq(X \cup \{c\})$. La *confiance* de π dans r est $conf(\pi, r) = freq(\pi, r)/freq(X, r)$. Le *taux d'accroissement* de π est $GR(\pi, r) = freq_r(X, r_c)/freq_r(X, r \setminus r_c)$ où r_c est le sous-ensemble de données restreint aux objets de classe c ($\mathcal{T}_c$) et $freq_r$ dénote la *fréquence relative* (i.e. $freq_r(X, r_c) = freq(X, r_c)/|\mathcal{T}_c|$).

Les premiers auteurs en classification supervisée à base de règles d'association (e.g. CBA (Liu et al., 1998), CAEP (Dong et al., 1999), CMAR (Li et al., 2001)) estiment qu'une règle est potentiellement intéressante si sa fréquence et sa confiance (ou taux d'accroissement) sont supérieurs à des seuils à définir. Comme il n'est pas aisé de définir ces seuils, souvent, des seuils très bas sont décidés arbitrairement – ce qui génère un grand nombre de règles. Un sous-ensemble des règles extraites est alors sélectionné en post-traitement en tenant compte de la couverture, de la redondance, de la corrélation (e.g. en choisissant les k meilleures règles par un test du χ^2). Toutefois certains de ces post-traitements nécessitent des paramétrages non-triviaux supplémentaires.

Dans cet article, nous proposons de suivre l'approche MODL pour évaluer la pertinence des règles de classification. L'approche MODL, déjà appliquée au regroupement de valeurs (Boullé, 2005), à la discrétisation (Boullé, 2006), à la régression (Hue et Boullé, 2007) ou encore aux arbres de décision (Voisine et al., 2009), mise sur un compromis entre *(i)* la précision de l'information prédictive fournie par le modèle et *(ii)* la robustesse afin d'obtenir une bonne généralisation du modèle. Ici, d'un point de vue MODL, un modèle est une règle de classification. Afin de choisir le meilleur modèle de règle, une approche Bayésienne est utilisée : on cherche à maximiser $p(\pi \mid r)$ la probabilité (a posteriori) d'un modèle de règle π sachant les données r. En appliquant le théorème de Bayes et considérant que la probabilité $p(r)$ des données est constante pour un problème donné, cela revient à maximiser l'expression $p(\pi) \times p(r \mid \pi)$ où $p(\pi)$ est la probabilité a priori d'une règle (ou prior) et $p(r \mid \pi)$, la vraisemblance, est la probabilité conditionnelle des données sachant le modèle de règle π. Ainsi, la règle π maximisant cette expression est la règle *la plus probable* issue des données du problème. Notre critère d'évaluation est basé sur le logarithme négatif de $p(\pi \mid r)$, que nous appelons *coût* d'une règle :

$$c(\pi) = -\log(p(\pi) \times p(r \mid \pi))$$

Afin de calculer $p(\pi)$, nous proposons une nouvelle définition de règle de classification basée sur une hiérarchie de paramètres qui identifient de manière unique une règle :

Standard Classification Rule Model. Une règle MODL (SCRM pour *standard classification rule model*) est définie de manière unique par :

- les attributs du corps de la règle
- pour chaque attribut du corps, la valeur (0 ou 1) qui fait partie du corps
- la distribution des classes , dans le corps et hors du corps

Les deux derniers points de la définition étendent la définition classique des règles de classification. En effet, pour un attribut a, les valeurs 0 et 1 sont les deux identifiants possibles. Cette notion se rapproche des règles avec des négations d'attribut dans le corps (Antonie et Zaïane, 2004). Les règles MODL rejoignent aussi la notion de règles de distribution récemment introduite par Jorge et al. (2006). Le conséquent de telles règles est une distribution de probabilité sur l'ensemble des classes (au lieu d'être une classe unique). L'exemple suivant illustre ces deux points.

Exemple de SCRM. Soit la règle MODL $\pi : (a_1 = 0) \wedge (a_2 = 1) \wedge (a_4 = 1) \rightarrow (p_{c_1} = 0.9, p_{c_2} = 0.1)$. Décrire le corps d'une telle règle consiste à choisir les attributs qui constituent le corps, puis choisir leurs valeurs (0 ou 1). Notons que l'on peut dériver facilement une règle de classification avec négations d'une SCRM en utilisant la classe avec la plus grande probabilité. Par exemple, $\pi : (a_1 = 0) \wedge (a_2 = 1) \wedge (a_4 = 1) \rightarrow c_1$.

Nous utilisons les notations suivantes pour la définition a priori d'une règle puis du critère :

Notations. Soit r une base de données binaires de N objets, m attributs et J classes. Pour une règle MODL $\pi : X \rightarrow (p_{c_1}, p_{c_2}, \ldots, p_{c_J})$ telle que $|X| = k \leq m$, nous notons :

- $X = \{x_1, \ldots, x_k\}$: les attributs du corps ($k \leq m$)
- $i_{x_1}, \ldots, i_{x_k}$: les index des valeurs qui participent au corps
- $N_X = N_{i_{x_1} \ldots i_{x_k}}$: le nombre d'objets dans la cellule $i_{x_1} \ldots i_{x_k}$ (i.e. dans le corps)
- $N_{\neg X} = N_{\neg i_{x_1} \ldots i_{x_k}}$: le nombre d'objets en dehors de la cellule $i_{x_1} \ldots i_{x_k}$ (i.e. hors du corps)
- $N_{Xj} = N_{i_{x_1} \ldots i_{x_k} j}$: le nombre d'objets de classe j dans la cellule $i_{x_1} \ldots i_{x_k}$
- $N_{\neg Xj} = N_{\neg i_{x_1} \ldots i_{x_k} j}$: le nombre d'objets de classe j hors de la cellule $i_{x_1} \ldots i_{x_k}$

A priori hiérarchique MODL. Nous nous appuyons sur la distribution a priori sur les modèles de règles MODL suivante :

- *(i)* le nombre d'attributs du corps est uniformément distribué sur $[0, m]$.
- *(ii)* pour un nombre donné k d'attributs, chaque ensemble de k attributs du corps est équiprobable.
- *(iii)* pour une valeur d'attribut donnée, participer au corps ou pas sont équiprobables.
- *(iv)* les distributions des classes dans le corps (resp. hors du corps) sont équiprobables.
- *(v)* les distributions des classes dans le corps et hors du corps sont indépendantes.

Grâce à la définition de l'espace des modèles et à sa distribution a priori, nous appliquons le théorème de Bayes pour exprimer la probabilité a priori d'un modèle ($p(\pi)$) et la probabilité des données sachant un modèle ($p(r \mid \pi)$). La probabilité a priori $p(\pi)$ d'un modèle règle est :

$$p(\pi) = p(X) \times \prod_{1 \leq l \leq k} p(i_{x_l}) \times \prod_{i \in \{X, \neg X\}} p(\{N_{ij}\} \mid N_X, N_{\neg X})$$

Tout d'abord, considérons $p(X)$, la probabilité d'avoir les attributs de X dans le corps. Considérant les deux premières hypothèses de l'a priori hiérarchique, le nombre de combinaisons $\binom{m}{k}$ serait un bon candidat pour ce prior ; toutefois ce terme combinatoire est symétrique. Au delà de $m/2$, ajouter de nouveaux attributs accroît la probabilité de la sélection. Ainsi l'ajout de variables non-pertinentes peut être favorisé – avec un effet non-significatif sur la vraisemblance du modèle. Préférant les règles les plus simples, nous suggérons d'utiliser le nombre de combinaisons avec remise $\binom{m+k-1}{k}$. Nous avons donc :

$$p(X) = \frac{1}{m+1} \cdot \frac{1}{\binom{m+k-1}{k}}$$

Pour chaque attribut x du corps, la valeur participant au corps de la règle doit être choisie dans $\{0, 1\}$. Ainsi nous avons $p(i_x) = 1/2$ selon l'hypothèse *(iii)* de l'a priori hiérarhique.

Considérant les hypothèses *(iv)* et *(v)* de l'apriori hiérarchique, dénombrer les distributions des J classes dans et hors du corps se réduit à un calcul combinatoire et nous avons :

$$p(\{N_{Xj}\} \mid N_X, N_{\neg X}) = \frac{1}{\binom{N_X+J-1}{J-1}} \quad \text{et} \quad p(\{N_{\neg Xj}\} \mid N_X, N_{\neg X}) = \frac{1}{\binom{N_{\neg X}+J-1}{J-1}}$$

Pour le terme de la vraisemblance, la probabilité des données sachant le modèle est la probabilité d'observer les données dans et hors du corps de la règle (avec respectivement N_X et $N_{\neg X}$ objets) étant donnée la distribution multinomiale pour N_X et $N_{\neg X}$. Nous avons donc :

$$p(r \mid \pi) = \frac{1}{\frac{N_X!}{\prod_{j=1}^{j=J} N_{Xj}!}} \cdot \frac{1}{\frac{N_{\neg X}!}{\prod_{j=1}^{j=J} N_{\neg Xj}!}}$$

Nous obtenons ainsi la définition complète du coût d'une règle `MODL` (SCRM) π :

$$\begin{aligned} c(\pi) \quad = \quad & \log(m+1) + \log \binom{m+k-1}{k} + k \log(2) \\ & + \log \binom{N_X + J - 1}{J-1} + \log \binom{N_{\neg X} + J - 1}{J-1} \\ & + \left(\log N_X! - \sum_{j=1}^{j=J} \log N_{Xj}! \right) + \left(\log N_{\neg X}! - \sum_{j=1}^{j=J} \log N_{\neg Xj}! \right) \end{aligned}$$

Le coût d'une règle `MODL` est défini par le logarithme négatif de probabilités. Par cette transformation, Shannon (1948) fait le lien entre probabilités et longueur de codage. Ainsi, $c(\pi)$ peut être vu comme la capacité d'une règle `MODL` à coder les classes en fonction des attributs. La première ligne correspond au choix du nombre d'attributs, des attributs et des valeurs participant au corps de la règle. La seconde ligne correspond au choix de la distribution des classes dans et hors du corps et la dernière ligne à la vraisemblance. Par construction du coût d'une règle, nous obtenons le théorème suivant :

Théorème 1 *Considérant l'a priori hiérarchique `MODL`, une SCRM est Bayes-optimal si son coût $c(\pi)$ est minimal sur l'ensemble de tous les SCRM.*

Intuitivement, les règles de faible coût sont les plus probables et donc les meilleures. Notons que $c(\pi)$ est moindre pour k petit (cf. ligne 1), i.e. les règles simples (petit corps) sont les plus probables et donc sont préférées. Par conséquent, les règles fréquentes sont plus probables que les non-fréquentes. Par ailleurs, la notion de pureté de règles apparaît dans la dernière ligne du coût : les règles les plus fortes ont aussi un coût moindre et sont donc les meilleures. Comme le coût d'une règle `MODL` dépend de la taille de la base de données (i.e. N et m), nous définissons notre critère d'évaluation (appelé $level$ [1]) comme la normalisation du coût :

$$level(\pi) = 1 - \frac{c(\pi)}{c(\pi_\emptyset)}$$

1. Le $level$ peut être interprété comme un taux de compression.

où $c(\pi_\emptyset)$ est le coût de la règle par défaut (i.e. au corps vide). Intuitivement, $c(\pi_\emptyset)$ est la longueur de codage des classes lorsqu'aucune information des attributs n'est utilisée. De manière plus formelle le coût de $\pi_\emptyset$ est :

$$c(\pi_\emptyset) = \log(m+1) + \log \binom{N+J-1}{J-1} + \log N! - \sum_{j=1}^{j=J} \log N_j!$$

Ainsi, si $level(\pi) = 0$ alors π a le même coût que $\pi_\emptyset$ et n'est pas plus probable que la règle par défaut. Lorsque $level(\pi) < 0$, alors utiliser π pour "expliquer" les données est plus coûteux qu'utiliser $\pi_\emptyset$. En d'autres termes, π est alors moins probable que la règle par défaut et ne sera donc pas intéressante. Les règles intéressantes sont mises en évidence lorsque $level(\pi) > 0$ (i.e. $c(\pi) < c(\pi_\emptyset)$) car plus probables que la règle par défaut. De plus, pour deux règles π_1 et π_2, si $level(\pi_1) > level(\pi_2)$, alors π_1 sera considérée comme meilleure que π_2 car plus probable. Notons le cas très particulier $level(\pi) = 1$ qui signifie que π, à elle seule, suffit à caractériser exactement la distribution des classes.

3 Validation expérimentale

Dans cette section, nous montrons expérimentalement *(i)* que la confiance et le taux d'accroissement ne sont généralement pas stables de la phase d'apprentissage à la phase de test et ne sont donc pas de bons candidats pour capturer la notion de robustesse des règles de classification, *(ii)* que le $level$, au contraire, est très stable dans les mêmes conditions d'expérience et *(iii)* que le $level$ nous permet d'identifier naturellement les règles robustes et intéressantes.

3.1 Protocole expérimental

Nous réalisons nos expériences sur sept bases de données `UCI` et la base de données `meningite` (François et al., 1992). Une brève description des données est reportée en table 1. L'expérience "train-test" consiste à diviser la base de données en deux sous-ensembles en respectant la distribution des classes. La première sert à l'apprentissage (i.e. l'extraction de règles selon des seuils donnés de fréquence, confiance et taux d'accroissement), la deuxième sert à évaluer l'évolution des valeurs des mesures (en test). Nous calculons et comparons aussi les valeurs de $level$ des règles extraites en apprentissage et en test. Nous utilisons le prototype `AClike` pour extraire les règles fréquentes et de confiance : en fait `AClike` extrait les itemsets γ-fréquents δ-libres (Boulicaut et al., 2003) qui sont les corps de règles π telle que $conf(\pi, r) \geq 1 - \delta/\gamma$. Nous utilisons aussi le prototype `consepminer` (Zhang et al., 2000) pour extraire les motifs γ-fréquents ρ-émergents.

Données	#Objets	#Attributs	#Classes et distribution
breast-w $*$	699	9	458/241
credit-a $*$	690	15	307/383
credit-g	1000	21	700/300
diabetes	768	8	500/268
meningitis $*$	329	23	245/84
sonar	208	60	97/111
tic-tac-toe	958	9	626/332
vote $*$	435	17	267/168

TAB. 1 – *Description des bases de données*

3.2 Résultats

Pour des raisons de limitations de pages et de lisibilité, nous reportons uniquement les résultats pour quatre bases de données (marquées par $*$ dans la table 1) et pour le taux d'accroissement. Notons que nous obtenons les mêmes observations et conclusions pour les autres bases et pour le critère de confiance.

Données originales. Dans les graphiques de la figure 2, nous reportons l'évolution train-test des valeurs de taux d'acroissement (GR) pour chaque base de données. Nous comparons aussi les valeurs de $level$ des règles extraites. Nous remarquons que GR est généralement instable : en effet, une règle à fort taux d'accroissement en apprentissage peut avoir un faible GR en test (voir les points éloignés de la droite identité) – ce qui confirme notre hypothèse que le taux d'accroissement (comme la confiance) ne capturent pas la notion de robustesse. Au contraire les valeurs de $level$ sont très stables (points proche de l'identité).

Ces premières expériences montrent qu'il peut être risqué de se reposer sur la confiance ou le taux d'accroissement pour faire des prédictions. La stabilité du $level$ est un signe de robustesse. Dans la suite, nous montrons expérimentalement que les règles à $level$ négatif sont non-significatives alors que celles à $level$ positif sont intéressantes.

Données bruitées. Afin de simuler la présence de bruit de classe dans les données `breast-w`, nous ajoutons de manière uniforme du bruit à l'attribut classe (changement de classe) en utilisant la fonction `AddNoise` de `WEKA` (Witten et Frank, 2005). Nous utilisons deux niveaux de bruit : moyen (20%) et fort (50%). Nous renouvelons l'expérience train-test sur chaque version artificiellement bruitée. Les résultats sont reportés en figure 5. A chaque niveau de bruit, les extracteurs classiques réussissent à extraire des motifs "potentiellement" intéressants – notons tout de même que beaucoup moins de règles sont extraites des contextes fortement bruités. Cependant, l'expérience train-test montre encore une fois l'instabilité des mesures classiques et cette instabilité est d'autant plus grande lorsque le contexte est très bruité (50%). En effet, la plupart des points (règles) sont en-dessous de la droite identité, ce qui indique que le taux d'accroissement (comme la confiance) sont à tort optimistes et peuvent mener à de mauvaises prédictions. Le $level$ est stable dans les contextes bruités aussi. Notons que la plupart des règles (toutes pour un niveau de bruit à 50%) ont un $level$ négatif dans les contextes bruités (i.e. elles ne sont pas plus probables que la règle par défaut et donc statistiquement non-significatives) ce qui est intuitif. Dans la suite, nous montrons expérimentalement que le $level$ positif met en évidence les règles intéressantes.

Règles à level positif. Dans les graphiques de la figure 6, nous reportons les valeurs en apprentissage et en test de μ, d'une mesure basée sur l'entropie de classe :

$$\mu(\pi) = N \times (Ent(\pi_\emptyset) - Ent(\pi))$$

μ est la différence entre les entropies conditionnelles de la règle par défaut et d'une règle π. Plus μ est grand plus π est intéressante. μ peut aussi être vu comme le nombre de bits sauvegardés lorsqu'on compresse les données en utilisant π au lieu de $\pi_\emptyset$. Comme prévu, dans la figure 6, les règles de $level$ positifs (o rouge) sont généralement les plus intéressantes, i.e. avec les plus fortes valeurs de μ et les règles avec un $level$ négatif (+ bleu) non-intéressantes obtiennent un faible score pour μ et sont donc en bas à gauche des graphiques.

4 Travaux connexes & discussions

Notre approche issue de l'approche MODL est à la croisée de la théorie Bayésienne, du principe de Minimum Description Length (MDL (Grünwald, 2007)) et de la complexité de Kolmogorov (Li et Vitányi, 2008).

A propos du principe MDL. Siebes et al. (2006) propose une approche d'extraction de motifs basée sur le principe MDL. Les auteurs cherchent à extraire les itemsets qui fournissent une bonne compression des données. Le lien entre probabilités et longueurs de codage permet aux auteurs de réécrire la longueur de codage d'un itemset I en $-log(P(I))$. Ainsi, les "meilleurs" itemsets ont un codage plus court et compressent mieux les données. Dans (van Leeuwen et al., 2006), une extension pour la classification supervisée est proposée. Les deux principales différences avec l'approche MODL sont les suivantes : *(i)*, l'utilisation de l'apriori hiérarchique MODL implique un codage différent ; *(ii)*, Siebes et al. (2006) cherchent un *ensemble* de motifs qui compressent les données alors qu'ici notre critère est défini pour *une* règle.
Notons que récemment, Suzuki (2009) utilise aussi le principe MDL pour intégrer des connaissances du domaine (sous forme de liste de décision) dans un processus d'extraction de règles. La longueur de codage cl d'une liste de décision L à extraire des données D, enrichi avec les connaissances du domaine K est considérée comme une mesure d'intérêt subjective : $cl(L) \equiv -\log P(L) - \log P(D \mid L) - \log P(K \mid L)$.

A propos de la robustesse. Le $level$ s'est montré stable expérimentalement. Ainsi, nous pouvons nous appuyer sur des règles au $level$ positif puisque l'intérêt des règles sera confirmé en test. La notion de robustesse a été étudiée récemment : Le Bras et al. (2010) suggèrent une nouvelle notion de robustesse dépendante de la mesure d'intérêt m utilisée et du seuil pour cette mesure m_{min}. Partant de l'observation qu'une règle peut être caractérisée par trois des valeurs de sa table de contingence (e.g. la fréquence du corps, la fréquence de la cible et le nombre de contre-exemples ; cf table 2), les auteurs définissent la robustesse d'une règle π par la distance Euclidienne normalisée $rob(\pi, m_{min}) = ||\pi - \pi^*||_2/\sqrt{3}$ entre π et une *règle limite* π^* (i.e. une règle qui minimise $g(\pi') = ||\pi' - \pi_{min}||_2$ où π_{min} est tel que $m(\pi_{min}) = m_{min}$). Cette approche capture bien la notion de robustesse, toutefois un paramétrage non-trivial supplémentaire est nécessaire pour le seuil de robustesse.

$X \to c$	c	$\neg c$	$\sum$
X	$freq(Xc, r)$	$freq(X\neg c, r)$	$freq(X, r)$
$\neg X$	$freq(\neg Xc, r)$	$freq(\neg X\neg c, r)$	$freq(\neg X, r)$
$\sum$	$\|c\|$	$\|\neg c\|$	N

TAB. 2 – *Contingency table for a classification rule* $X \to c$

A propos de la redondance. Une règle de classification $\pi_2 : Y \to c_i$ est dite redondante par rapport à $\pi_1 : X \to c_j$ si $c_i = c_j$, $X \subseteq Y$ et π_1 et π_2 ont (à peu près) le même pouvoir discriminant (selon une mesure d'intérêt) – une règle redondante n'est pas utile. Soient deux itemsets X et Y tels que $X \subseteq Y$ et $freq(X, r) = freq(Y, r)$, alors pour une mesure d'intérêt m basée sur la fréquence on a $m(X) = m(Y)$ et Y est redondant. On peut regrouper les itemsets de même support en classes d'équivalence. L'unique plus grand itemset (selon l'inclusion ensembliste) est appelé itemset fermé et les plus petits sont les itemsets libres. Dans la littérature, les classifieurs à base de motifs (e.g. Baralis et Chiusano (2004)) préfèrent les itemsets libres pour des raisons de simplicité. Le critère $level$ va dans le même sens. En effet, si Y est un itemset fermé et X un itemset fermé d'une même classe d'équivalence, alors

$c(\pi_2 : Y \rightarrow c_i) \geq c(\pi_1 : Y \rightarrow c_i)$ puisque le nombre d'attributs favorise π_1 – ce qui peut se formaliser par le théorème suivant :

Théorème 2 *Soient X et Y deux itemsets tels que $X \subset Y$ et $freq(X, r) = freq(Y, r)$; alors X est préférable à Y selon le critère $level$ (i.e. $level(X) > level(Y)$).*

5 Conclusion & perspectives

Dans cet article, nous présentons un nouveau critère Bayésien (le $level$) pour l'évaluation des règles de classification dans les données binaires. Issu de l'approche MODL, le $level$ propose une solution pour deux faiblesses identifiées des approches existantes (basées sur le cadre fréquence-confiance ou taux d'accroissement) : le paramétrage non-trivial des seuils des mesures d'intérêt et la non-stabilité de ces mesures. Le $level$ favorise un compromis entre précision et généralisation et permet d'identifier naturellement les règles intéressantes et robustes sans technique de paramétrage. Les expériences menées confirment la pertinence et la robustesse du critère.
Dans ce travail, le $level$ est utilisé en post-traitement afin de sélectionner les règles robustes à partir d'un ensemble de règles confiantes ou émergentes. La prochaine étape sera une approche constructive pour extraire directement des règles à $level$ positif. De plus, comme l'approche MODL est adaptée aux attributs numériques et catégoriels, une autre étape visera à étendre ce cadre aux règles quantitatives en considérant discrétisation et groupement de valeurs.

Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings ACM SIGMOD'93*, pp. 207–216.

Antonie, M.-L. et O. R. Zaïane (2004). An associative classifier based on positive and negative rules. In *DMKD'04*.

Asuncion, A. et D. Newman (2007). UCI machine learning repository. `http://archive.ics.uci.edu/ml/`.

Baralis, E. et S. Chiusano (2004). Essential classification rule sets. *ACM Transactions on Database Systems 29*(4), 635–674.

Boulicaut, J.-F., A. Bykowski, et C. Rigotti (2003). Free-sets : A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery 7*(1), 5–22.

Boullé, M. (2005). A bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2006). MODL : A bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : Generalizing association rules to correlations. In *SIGMOD'97*, pp. 265–276. ACM Press.

Bringmann, B., S. Nijssen, et A. Zimmermann (2009). Pattern-based classification : A unifying perspective. In *LeGo'09 workshop co-located with EMCL/PKDD'09*.

Dong, G. et J. Li (1999). Efficient mining of emerging patterns : discovering trends and differences. In *Proceedings KDD'99*, pp. 43–52. ACM Press.

Dong, G., X. Zhang, L. Wong, et J. Li (1999). CAEP : Classification by aggregating emerging patterns. In *Proceedings DS'99*, Volume 1721 of *LNCS*, pp. 30–42. Springer.

François, P., B. Crémilleux, C. Robert, et J. Demongeot (1992). MENINGE : a medical consulting system for child's meningitis study on a series of consecutive cases. *Artificial Intelligence in Medecine 4*(4), 281–292.

Grünwald, P. (2007). *The minimum description length principle*. MIT Press.

Hue, C. et M. Boullé (2007). A new probabilistic approach in rank regression with optimal bayesian partitioning. *Journal of Machine Learning Research 8*, 2727–2754.

Jorge, A. M., P. J. Azevedo, et F. Pereira (2006). Distribution rules with numeric attributes of interest. In *PKDD'06*, pp. 247–258.

Le Bras, Y., P. Meyer, P. Lenca, et S. Lallich (2010). A measure of robustness of association rules. In *ECML/PKDD'10*, Volume 6322 of *LNCS*, pp. 227–242. Springer.

Li, M. et P. M. B. Vitányi (2008). *An Introduction to Kolmogorov Complexity and Its Applications (3^{rd} edition)*. Springer.

Li, W., J. Han, et J. Pei (2001). CMAR : Accurate and efficient classification based on multiple class-association rules. In *Proceedings ICDM'01*, pp. 369–376. IEEE Computer Society.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Proceedings KDD'98*, pp. 80–86. AAAI Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*.

Siebes, A., J. Vreeken, et M. van Leeuwen (2006). Item sets that compress. In *SIAM DM'06*.

Suzuki, E. (2009). Negative encoding length as a subjective interestingness measure for groups of rules. In *PAKDD'09*, pp. 220–231.

van Leeuwen, M., J. Vreeken, et A. Siebes (2006). Compression picks item sets that matter. In *PKDD'06*, pp. 585–592.

Voisine, N., M. Boullé, et C. Hue (2009). Un critère d'évaluation bayésienne pour la construction d'arbre de décision. In *EGC'09*, pp. 67–78.

Witten, I. H. et E. Frank (2005). *Data Mining : Practical machine learning tools and techniques (2nd edition)*. Morgan Kaufmann.

Zhang, X., G. Dong, et K. Ramamohanarao (2000). Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In *KDD'00*, pp. 310–314.

Summary

In this paper, we suggest a new criterion for the evaluation of classification rules' robustness in binary labeled data sets. Our criterion arises from a Bayesian approach : we propose an expression of the probability of a rule given the data. The most probable rules are thus the rules that are robust. Our Bayesian criterion is derived from this defined expression and allows us to mark out the robust rules from a given set of rules without parameter tuning.

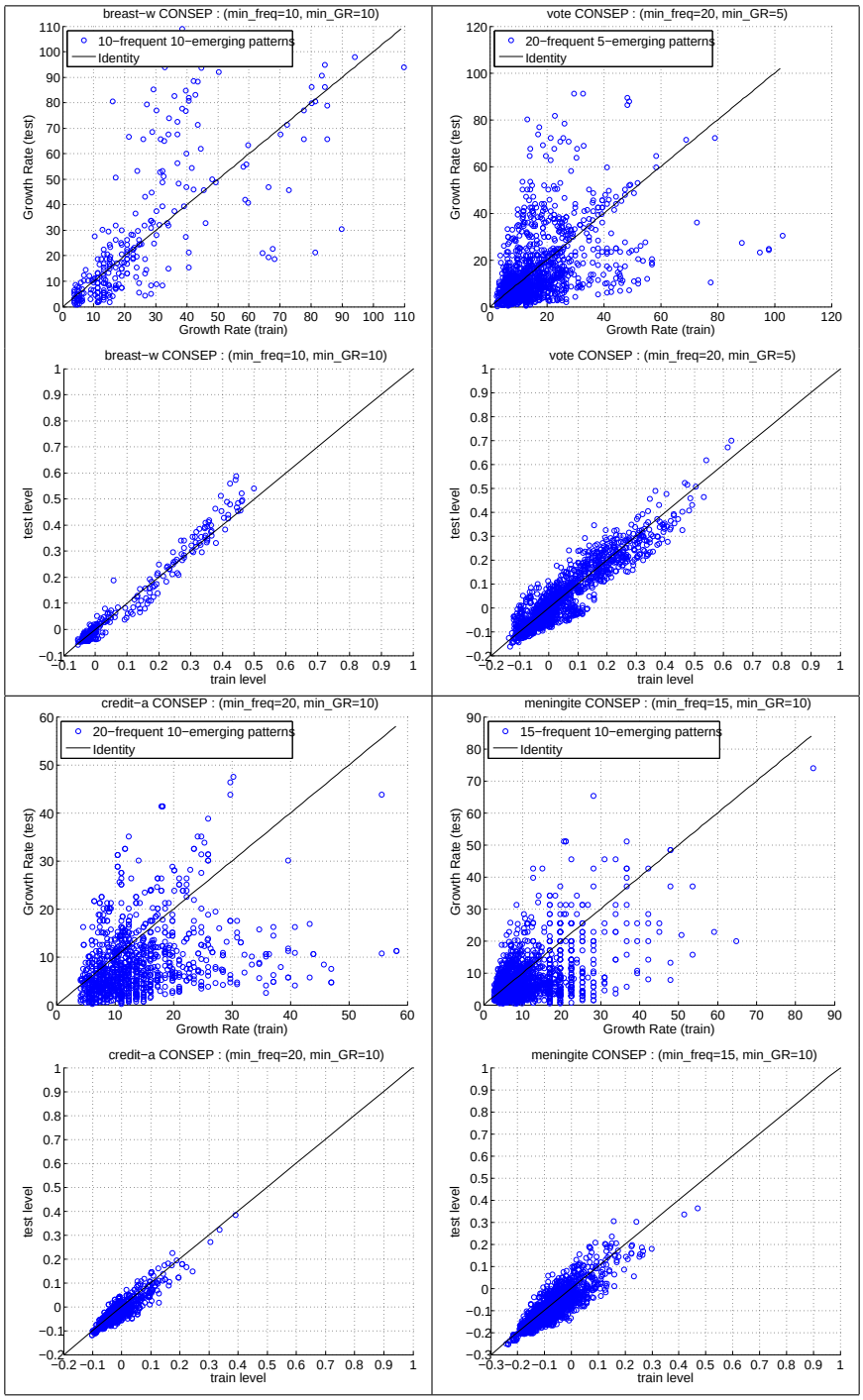

FIG. 2 – *Comparaison des valeurs de GR et level : apprentissage vs test. Les valeurs de GR sont instables entre l'apprentissage et la phase test alors que les valeurs de level sont plus stables (points proches de la droite identité), ce qui confirme la robustesse du critère.*

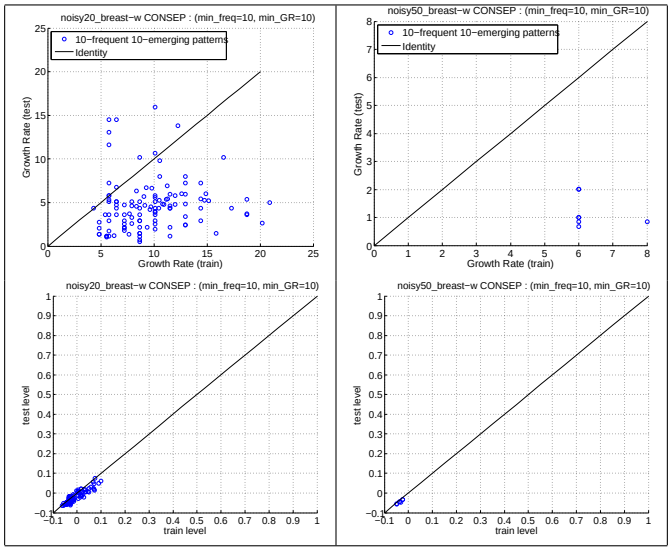

FIG. 3 – *Comparaison des valeurs de GR, $conf$ et $level$ dans un contexte bruité : apprentissage vs test. Les règles considérées comme intéressantes par le GR se révèlent très instables, ce qui se traduit par un level négatif.*

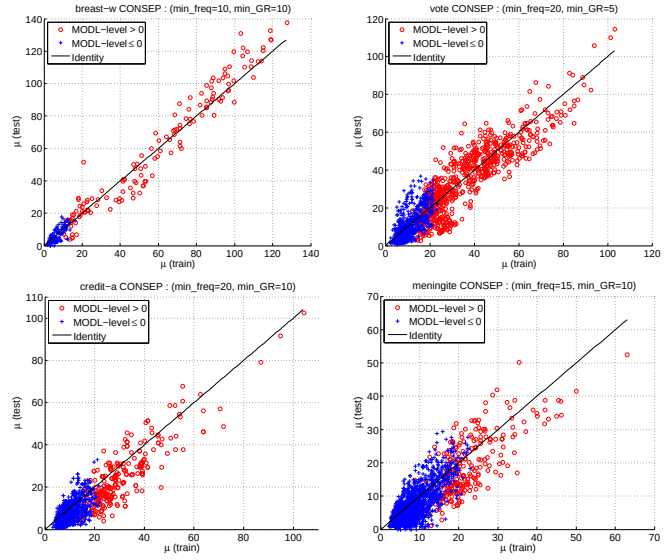

FIG. 4 – *Comparaison des valeurs de μ pour les règles émergentes : apprentissage vs test. Les meilleures règles (i.e., les plus probables, avec level positif, 'o' rouge) sont généralement localisées en haut à droite du graphique, alors les règles non robustes (avec level négatif, '+' bleu) sont proches de l'origine.*