Construction de descripteurs pour classer à partir d'exemples bruités

Nazha Selmaoui*, Dominique Gay*, Jean-François Boulicaut**

*Université de la Nouvelle-Calédonie, ERIM EA3791, PPME EA3325
BP R4 98851 Nouméa, Nouvelle-Calédonie
{nazha.selmaoui, dominique.gay}@univ-nc.nc
**Université de Lyon, CNRS
INSA-Lyon, LIRIS UMR5205, 69621 Villeurbanne, France
jean-francois.boulicaut@insa-lyon.fr

Résumé. En classification supervisée, la présence de bruit sur les valeurs des descripteurs peut avoir des effets désastreux sur la performance des classifieurs et donc sur la pertinence des décisions prises au moyen de ces modèles. Traiter ce problème lorsque le bruit affecte un attribut classe a été très étudié. Il est plus rare de s'intéresser au bruit sur les autres attributs. C'est notre contexte de travail et nous proposons la construction de nouveaux descripteurs robustes lorsque ceux des exemples originaux sont bruités. Les résultats expérimentaux montrent la valeur ajoutée de cette construction par la comparaison des qualités obtenues (e.g., précision) lorsque l'on utilise les méthodes de classification à partir de différentes collections de descripteurs.

1 Introduction

Lorsqu'il s'agit de décrire un ensemble d'objets au moyen de descripteurs, les valeurs de ces derniers peuvent être collectées de façon plus ou moins fiable, par exemple lorsqu'elles sont le résultat d'un processus complexe d'acquisition de mesures. En classification supervisée, nous savons que la présence de bruit dans les exemples d'apprentissage peut avoir un impact négatif sur la performance des modèles construits et donc sur la pertinence des prises de décisions associées. Il existe deux types de problèmes de bruits. Le problème du *bruit de classe* (affectant uniquement l'attribut classe) a été très étudié ces dernières années. Plusieurs approches ont été proposées pour, par exemple, l'élimination, la correction du bruit (Zhu et Wu, 2004), ou encore la pondération des instances (Rebbapragada et Brodley, 2007). Le contexte du *bruit d'attributs* affectant uniquement les attributs non-classe ou descripteurs est moins traité. Nous trouvons des travaux sur la modélisation et l'identification du bruit (Kubica et Moore, 2003; Zhang et Wu, 2007) ainsi que des techniques de filtrage pour "nettoyer" les attributs bruités (Zhu et Wu, 2004; Yang et al., 2004).

Nous nous intéressons à ce problème de la classification en présence de descripteurs (attributs non classe) bruités. Plus précisément, nous voulons apporter une réponse à la question suivante : *comment construire des modèles prédictifs robustes à partir de données dont les attributs Booléens sont a priori bruités ?*

Nous proposons une construction de descripteurs robustes sans pour autant éliminer les exemples bruités ni modifier les valeurs des attributs dans les données d'apprentissage. Cette approche combine deux avancées récentes dans les domaines de la fouille de motifs tolérants aux exceptions d'une part et de la construction de descripteurs d'autre part. Le concept des *itemsets fréquents tolérants au bruit* a été introduit dans Yang et al. (2001) pour permettre la découverte de motifs pertinents dans les données booléennes bruitées. Des synthèses sont disponibles comme, par exemple, Besson et al. (2006). Parmi les extensions robustes étudiées, Besson et al. (2006) présente les motifs FBS[1] qui sont une généralisation des concepts formels (ensembles fermés) et qui tolèrent un nombre borné d'erreurs par colonne (δ) et qui sont construits à partir de la représentation condensée des ensembles δ-libres introduite dans Boulicaut et al. (2000). Ces motifs δ-libres sont au coeur de notre méthode de construction de descripteurs robustes. La *construction de nouveaux descripteurs* basée sur des motifs ensemblistes et des propriétés de fermeture (voir Selmaoui et al. (2006); Cheng et al. (2007); Garriga et al. (2008); Gay et al. (2008) pour des propositions récentes) partent de l'hypothèse que les ensembles d'attributs (ou "itemsets") peuvent être plus pertinents que les attributs seuls pour la caractérisation des classes pourvu que l'on puisse contrôler la redondance dans ces collections. En travaillant sur les propriétés des fermetures, on peut justement donner un cadre à l'élimination de la redondance en groupant les motifs dans des classes d'équivalence. Notre proposition exploite ces avancées. Les étapes de notre processus, comme indiqué en Figure 1, se déroule de la manière suivante : *(i)* nous extrayons des motifs non-redondants et tolérants aux bruits basés sur une représentation condensée approximative des ensembles fréquents ; *(ii)* à partir de ces motifs nous construisons de nouveaux descripteurs pour nos données ; *(ii)* en fin de processus, nous appliquons les algorithmes classiques de classification supervisée, par exemple C4.5 ou encore NB ("Naive Bayes").

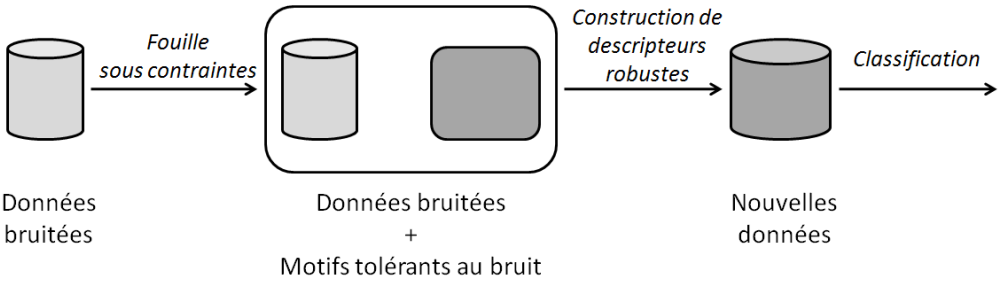


FIG. 1: Construction de descripteurs robustes (Noise-Tolerant Feature Construction NTFC).

Nos travaux rentrent donc dans le cadre de la classification supervisée basée sur les motifs locaux. Les motifs utilisés furent tout d'abord des règles d'association Liu et al. (1998), puis des motifs émergents[2] (voir Ramamohanarao et Fan (2007) pour une synthèse), et plus récemment des représentations condensées d'ensembles fréquents. L'une des rares études en classification tolérante au bruit et basée sur les motifs locaux exploite des motifs émergents (Sun et al., 2003; Fan et Ramamohanarao, 2004). Le reste de l'article est organisé comme

[1] FBS pour *Free set based Bi-Set*.

[2] Un motif est dit émergent s'il est fréquent relativement à une classe de données et non-fréquent dans le reste des données. Le ratio des fréquences relatives, noté ρ, donne lieu à différentes catégories de motifs dits ρ-émergents.

suit. La Section 2 donne les définitions nécessaires à notre proposition détaillée en Section 3. La Section 4 décrit nos expérimentations et discute nos résultats sur des données d'UCI[3]. La Section 5 conclut.

2 Définitions préliminaires

Soit une base de données binaires $r = \{\mathcal{T}, \mathcal{I}, \mathcal{R}\}$ où $\mathcal{T}$ est un ensemble de transactions (ou objets) décrit par un ensemble $\mathcal{I}$ d'attributs (ou items) Booléens et $\mathcal{R} : \mathcal{T} \times \mathcal{I} \mapsto \{0, 1\}$. Lorsque $\mathcal{R}(t, i) = 1$, on dit que la transaction t contient l'item i (ou l'objet t possède la propriété décrite par l'attribut i). Un itemset $I \in \mathcal{I}$ est simplement un ensemble d'items. La fréquence d'un itemset $I \in \mathcal{I}$ est définie par $freq(I, r) = |Objets(I, r)|$ où $Objets(I, r) = \{t \in \mathcal{T} \mid \forall i \in \mathcal{I} \quad \mathcal{R}(t, i) = 1\}$. On dit que I est γ-fréquent si $freq(I, r) \geq \gamma$.

Définition 2.1 (règle d'association, règle δ-forte, itemset δ-libre) *Une règle d'association π dans r est une expression $I \Rightarrow J$ où $I \subseteq \mathcal{I}$ et $J \subseteq \mathcal{I} \setminus I$. La fréquence d'une règle π est $freq(I \cup J, r)$ et sa confiance $conf(\pi, r) = freq(I \cup J, r)/freq(I, r)$. Soit δ un entier naturel. Une règle* δ-forte *est une règle d'association de la forme $I \Rightarrow^{\delta} J$ qui est violée dans au plus δ transactions (i.e., la transaction implique les items de I mais pas ceux de J). Un itemset $I \subseteq \mathcal{I}$ est dit* δ-libre *ssi il n'existe pas de règle δ-forte entre ses sous-ensembles propres. Lorsque $\delta = 0$, δ est omis, on parle alors de règles fortes et d'itemsets libres.*

Les ensembles δ-libres et les règles δ-fortes ont été introduits dans Boulicaut et al. (2000) pour la définition d'une représentation condensée approximative des ensembles fréquents. Le concept d'ensemble δ-libres généralise clairement celui de motif clé (cas $\delta = 0$) utilisé dans Bastide et al. (2000). Reprenons une présentation en terme de classes d'équivalence qui a justement été introduite dans Bastide et al. (2000).

Définition 2.2 (δ-fermeture, classe d'équivalence) *Soit δ un entier naturel. La δ-fermeture d'un itemset I dans r est l'ensemble d'items $cl_\delta(I, r) = \{i \in \mathcal{I} \mid freq(I, r) - freq(I \cup \{i\}) \leq \delta\}$. Lorsque $\delta = 0$, $cl_0(I, r) = \{i \in \mathcal{I} \mid freq(I, r) = freq(I \cup \{i\})\}$ et cl_0 correspond à l'opérateur de fermeture bien connu. Nous pouvons aussi regrouper les ensembles par* classe d'équivalence sur la δ-fermeture *(δ-CEFs) : ainsi, deux motifs δ-libres I et J sont δ-équivalents (noté $I \sim_{cl_\delta} J$) si $cl_\delta(I, r) = cl_\delta(J, r)$.*

A nouveau, notons que lorsque $\delta = 0$, nous retrouvons la formalisation de Bastide et al. (2000). Il est possible de dériver des règles δ-fortes à partir des δ-CEFs (i.e., à partir des ensembles δ-libres et de leurs δ-fermetures). En effet, une règle δ-forte est constituée d'un ensemble compris (au sens de l'inclusion) entre un ensemble δ-libre et chacun des éléments de sa δ-fermeture. Besson et al. (2006) combine les itemsets δ-libres et leurs δ-fermetures pour construire des motifs appelés δ-bi-ensembles.

Définition 2.3 (δ-bi-ensemble γ-fréquent) *Un bi-ensemble (T, I) tel que $T \subseteq \mathcal{T}$ et $I \subseteq \mathcal{I}$ est un δ-bi-ensemble γ-fréquent ssi $I = I_1 \cup I_2$, I_1 est un ensemble δ-libre γ-fréquent, $cl_\delta(I_1, r) = I$ et $Objets(I_1, r) = T$.*

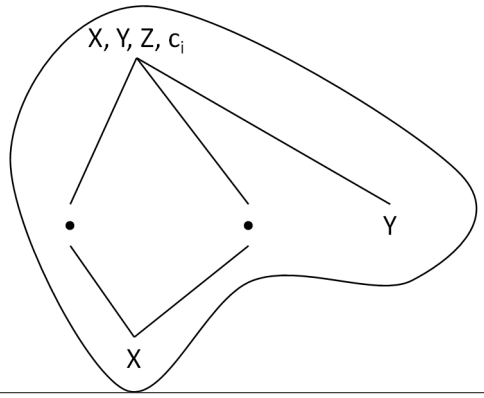
[3] http ://archive.ics.uci.edu/ml/

Ce type de motif est une généralisation de la notion de concept formel (i.e., un 0-bi-ensemble est un concept formel) pour une certaine tolérance aux exceptions Pensa et al. (2006). En fait, le paramètre δ détermine un nombre d'erreurs borné par colonne pour les colonnes correspondant aux attributs de la δ-fermeture. Dans la suite, nous allons décrire comment les δ-CEFs sont utilisées pour la production de nouveaux descripteurs robustes.

3 Des classes d'équivalence aux descripteurs pertinents

Afin de réaliser des tâches de classification, nous nous sommes intéressés aux règles δ-fortes pertinentes contenues dans des δ-CEFs. La Figure 2(a) présente un exemple typique de δ-CEF intéressante : les ensembles δ-libres X et Y ne contiennent pas l'attribut classe et leur δ-fermeture (X, Y, Z, c_i) contient bien l'attribut classe c_i. Ainsi, on peut dériver deux règles δ-fortes potentiellement intéressantes qui concluent sur un attribut classe : $X \rightarrow c_i$ et $Y \rightarrow c_i$.

(a) Cas intéressant d'une classe d'équivalence de δ-fermeture.



(b) Table de contingence pour la règle δ-forte $X \rightarrow^{\delta} c_i$.

$X \rightarrow c_i$	c_i	$\bar{c_i}$	Σ
X	$\gamma - \delta$	δ	γ
$\bar{X}$	$\cdot$	$\cdot$	$\cdot$
Σ	$\lvert r_{c_i} \rvert$	$\lvert r \setminus r_{c_i} \rvert$	$\lvert r \rvert$

FIG. 2: Information dans les classes d'équivalence de δ-fermetures intéressantes.

Selon la formalisation proposée dans Crémilleux et Boulicaut (2002), $\pi : X \rightarrow c_i$ est une règle δ-forte de caractérisation d'une classe si c_i est un attribut classe et le corps X est minimal. X est minimal s'il n'existe pas d'autres règles δ-fortes fréquentes $\pi' : Y \rightarrow c_i$ telle que $Y \subseteq X$ et $conf(\pi', r) \geq 1 - \frac{\delta}{\gamma}$. De plus, pour $\delta \in [0; \lfloor \gamma/2 \rfloor[$, nous savons avec Crémilleux et Boulicaut (2002) que l'ensemble des règles δ-fortes de caractérisation ne renferme pas de conflits d'inclusion ou d'égalité de corps.

La définition des règles de caractérisation basée sur la confiance n'est pas suffisante pour la prédiction. Nous proposons d'utiliser également le taux d'accroissement (TA caractérisant les motifs émergents (Dong et Li, 1999)). Son pouvoir discriminant vient du fait qu'il confronte la fréquence relative d'une règle dans une classe par rapport au reste de la base. Le taux d'accroissement est défini par le rapport des fréquences relatives suivant :

$$TA(\pi, r) = \frac{freq_r(X, r_{c_i})}{freq_r(X, r \setminus r_{c_i})}$$

où r_{c_i} est la base restreinte aux objets de class c_i. Dans, Hébert et Crémilleux (2006), les auteurs placent TA dans le cadre plus général des mesures d'intérêt dites δ-dépendantes, i.e., elles dépendent de la fréquence du corps de la règle (ici γ) et du nombre d'exceptions (ici δ) en concordance avec les deux principes suivants :

(i) Lorsque γ est fixé, $TA(\pi, r)$ augmente avec $freq(\pi, r)$

(ii) Lorsque δ est fixé, $TA(\pi, r)$ augmente avec γ

A partir de ces principes, on trouve dans Hébert et Crémilleux (2006) un résultat concernant les bornes inférieures pour plusieurs mesures d'intérêt (entre autres TA et $conf$) en fonction de γ et δ. En effet, soit la table de contingence pour une règle δ-forte $\pi : X \rightarrow c_i$ en Figure 2(b). Par construction, $freq(X, r_{c_i})$ admet la borne inférieure $\gamma - \delta$ et $freq(X, r \setminus r_{c_i})$ la borne supérieure δ.

Nous pouvons donc déduire une borne inférieure pour les mesures TA et $conf$:

$$TA(\pi, r) \geq \frac{\gamma - \delta}{\delta} \cdot \frac{|r \setminus r_{c_i}|}{|r_{c_i}|} \quad \text{et} \quad conf(\pi, r) \geq 1 - \delta/\gamma$$

Après quelques déductions, nous obtenons :

$$TA(\pi, r) \geq 1 \qquad \Longrightarrow \qquad \delta < \gamma \cdot \frac{|r \setminus r_{c_j}|}{|r|} \tag{1}$$

$$conf(\pi, r) \geq 1/2 \qquad \Longrightarrow \qquad \delta < \gamma/2 \tag{2}$$

où c_j est la classe majoritaire.

Ainsi, pour un seuil de fréquence donné γ, les ensembles γ-fréquents δ-libres dont la δ-fermeture contient un attribut classe tels que δ satisfait les contraintes des équations (1 et 2) sont des motifs émergents et membres gauches de règles non conflictuelles. Dans la suite, nous considérons que les valeurs γ et δ sont contraintes par les équations (1) et (2).

Notre processus de construction de descripteurs peut maintenant être résumé par l'Algorithme 1. La procédure `ExtractionMotifs` (Ligne 2) extrait tous les ensembles γ-fréquents δ-libres qui sont les membres gauches des règles δ-fortes de caractérisation dans r. Cette étape est effectuée efficacement en utilisant une évolution simple du prototype `AC-like`[4]. Ce prototype est une implémentation de l'algorithme par niveaux décrit dans Boulicaut et al. (2000) et qui profite de l'anti-monotonicité des contraintes de δ-liberté et de γ-fréquence pour extraire efficacement tous les ensembles γ-fréquents δ-libres. Comme nous nous intéressons aux ensembles minimaux dont la δ-fermeture contient un attribut classe, nous pouvons davantage élaguer l'espace de recherche en éliminant les sur-ensembles d'ensembles sélectionnés à un niveau précédent. Ensuite, chaque I' devient un nouveau descripteur (i.e., un attribut) pour r' et (Ligne 5) la valeur de I' pour une transaction t est la proportion des attributs de I' qui décrivent t dans r. Ici, $Items$ est l'opérateur dual pour $Objets$. Nous obtenons alors de nouveaux descripteurs avec des valeurs numériques, $\mathcal{R}' \mapsto [0; 1]$ et $\mathcal{R}'(t, I') \in \{0, \frac{1}{p}, \ldots, \frac{p-1}{p}, 1\}$ avec $p = |I'|$, i.e., le nombre d'attributs de I'. Nous pensons qu'un tel codage est plus pertinent qu'un codage binaire : en effet, avec un tel codage numérique, nous avons toujours la possibilité de discrétiser en séparant le domaine de valeurs de I' d'une manière au moins aussi pertinente que celui imposé par un codage binaire. Enfin (Ligne 6), r' est la nouvelle base de données décrite par des descripteurs numériques robustes et prête pour l'étape d'apprentissage supervisé.

[4]Disponible depuis http ://liris.cnrs.fr/∼turing/

Algorithme 1 : Construire une nouvelle base de données à partir de descripteurs tolérants aux bruits

entrée : une base de données binaires $r = \{\mathcal{T}, \mathcal{I}, \mathcal{R}\}$, deux entiers positifs γ et δ seuils de fréquence et d'erreurs

sortie : une base de données $r' = \{\mathcal{T}, \mathcal{I}', \mathcal{R}'\}$ avec des descripteurs tolérants aux bruits

```
1 begin
2     I' ⟵ ExtractionMotifs(r, γ, δ);
3     for t ∈ T do
4         for I' ∈ I' do
5             R'(t, I') ⟵ |I' ∩ Items(t,r)| / |I'|;
6     r' ⟵ {T, I', R'};
7 end
```

4 Expérimentations

Afin d'évaluer la capacité de notre processus de construction de descripteurs (NTFC) à résister au bruit d'attribut, nous l'expérimentons sur des données UCI (a priori non bruitées) et sur plusieurs versions bruitées de ces dernières. Nous voulons apprendre des modèles prédictifs pertinents à partir d'exemples bruités ; ainsi dans nos expériences les ensembles d'apprentissage sont soumis à un bruit d'attributs et les ensembles de validation restent *propres*. Notre modélisation du bruit consiste en l'injection de bruits *aléatoires* uniquement dans les attributs non-classe et dans l'ensemble d'apprentissage : soit d une base de données, introduire $x\%$ de bruits ($x \in \{10, 20, 30, 40, 50\}$) indique que chaque attribut de d a $x\%$ de chances de voir sa valeur changer et dans chaque transaction de l'ensemble d'apprentissage. Pour les attributs continus, nous utilisons la méthode de discrétisation de Fayyad et Irani (1993) avant d'injecter du bruit (puisque nous nous intéressons au bruit dans les attributs Booléens). A partir de la base de données ainsi bruitée, nous générons une nouvelle base de données en construisant les nouveaux descripteurs par l'Algorithme 1 puis nous appliquons les classifieurs C4.5 (Quinlan, 1993) et Naive Bayes. Notons que les étapes de pré-traitement ainsi que les résultats de précision pour NB et C4.5 sont obtenus par validation croisée grâce à la plateforme Weka (Witten et Frank, 2005).

4.1 Résultats de précision

Dans le Tableau 1, nous reportons les résultats de précisions des classifieurs NB et C4.5 sur différentes versions (bruitées et/ou améliorées) des données. Les seuils de fréquence utilisés pour chaque base sont indiqués en colonne 1 : (f :x-y/z) indique que la fréquence absolue varie de x à y avec un pas de z (lorsque celui-ci est régulier). Les colonnes NTFC rapportent les précisions maximales obtenues parmi toutes les valeurs de γ et δ testées et entre parenthèses, certaines des fréquences absolues pour lesquelles la précision est meilleure que sur les données originales. Dans la colonne *Amélio*, nous reportons aussi l'amélioration de la précision due au processus NTFC, ainsi que la moyenne par niveau de bruit en fin de table.

BD	C4.5	NTFC & C4.5	Amélio.	NB	NTFC & NB	Amélio.
breast-w (f :20-40/5)	95.57	**95.85** (*)	+0.28	97.28	**97.42** (*)	+0.14
10%	94.42	**95.14** (*)	+0.72	97.14	**97.42** (*)	+0.28
20%	92.56	**94.28** (*)	+1.72	96.99	**97.28** (*)	+0.29
30%	91.99	**92.28** (40,35,20)	+0.29	96.71	**97.28** (*)	+0.57
40%	89.13	**90.42** (*)	+1.29	96.56	**97.28** (*)	+0.72
50%	88.55	*88.12* (-)	-0.33	96.42	**97.13** (*)	+0.71
colic (f :25-50/5)	85.04	**85.84** (50,45,40)	+0.80	79.90	**83.94** (*)	+4.04
10%	83.14	**85.29** (*)	+2.41	78.81	**84.21** (*)	+5.40
20%	82.85	**83.66** (50,45)	+0.81	79.09	**83.40** (*)	+4.31
30%	80.69	**84.48** (50,45,40,35)	+3.79	78.80	**83.41** (*)	+4.61
40%	76.11	**83.95** (*)	+7.84	78.26	**81.49** (*)	+3.23
50%	80.42	**81.79** (50,45,40)	+1.37	74.73	**81.77** (*)	+7.04
heart-c (f :25-45/5)	80.47	**84.12** (*)	+3.65	81.79	**83.79** (*)	+2.00
10%	79.45	**82.49** (*)	+3.04	82.46	**83.81** (40,35,30,25)	+1.35
20%	79.53	**80.56** (45,40,35,30)	+1.03	82.79	**84.13** (*)	+1.34
30%	77.54	**80.16** (45,40,35)	+2.62	83.14	**84.14** (45,25)	+1.00
40%	70.91	**77.24** (*)	+6.33	83.15	**83.48** (25)	+0.33
50%	71.57	**76.82** (*)	+5.25	80.20	**80.85** (25)	+0.65
heart-h (f :10-22/2)	75.55	**81.64** (*)	+6.11	84.07	**84.71** (22,18,16)	+0.64
10%	78.26	**80.63** (*)	+2.37	84.07	**84.37** (22)	+0.30
20%	80.27	**80.62** (22,20)	+0.35	84.06	**84.37** (22,20,18)	+0.31
30%	79.94	*79.61* (-)	-0.33	83.72	**84.07** (22,20)	+0.35
40%	77.22	**81.99** (*)	+4.77	82.33	**83.37** (22,20)	+1.04
50%	76.25	**82.31** (*)	+6.06	78.94	**82.67** (*)	+3.73
heart-s (f :22-32/2)	81.85	**86.67** (*)	+4.82	81.48	**83.33** (*)	+1.85
10%	80.37	**82.96** (*)	+2.59	81.48	**83.33** (*)	+1.85
20%	75.55	**80.37** (*)	+4.82	82.59	**83.70** (*)	+1.11
30%	72.96	**76.67** (32,30,27)	+3.71	81.11	**83.70** (*)	+2.59
40%	65.92	**68.89** (30,27,22)	+2.97	81.85	74.81 (-)	-7.04
50%	61.11	**62.22** (27,25,22)	+1.11	55.55	**67.03** (*)	+11.48
iris (f :7-30)	93.33	**93.33** (*)	+0.00	92.67	**95.33** (*)	+2.66
10%	92.00	**93.33** (30,25,20,17,15,7)	+1.33	92.67	**95.33** (*)	+2.66
20%	90.67	**94.00** (30,25,20)	+3.33	92.67	**94.00** (30,25,20,15)	+1.33
30%	90.00	**94.00** (30,20)	+4.00	92.67	**94.67** (30,25,20)	+2.00
40%	86.67	**92.67** (30)	+6.00	92.67	*91.33* (-)	-1.34
50%	80.00	*79.33* (-)	-0.67	93.33	*77.33* (-)	-16.00
tic-tac-toe (f :25-40/5)	93.21	**100** (*)	+6.79	68.47	**77.86** (*)	+9.39
10%	92.59	**99.48** (*)	+6.89	68.47	**75.26** (*)	+6.79
20%	88.31	**97.49** (*)	+9.18	70.14	**74.54** (40,35,30)	+4.40
30%	84.65	**95.72** (*)	+11.12	71.60	**73.69** (40,35,30)	+2.09
40%	76.51	**90.30** (*)	+13.79	69.20	**70.79** (*)	+1.59
50%	72.86	**84.87** (*)	+12.01	68.36	*67.01* (-)	-1.35
wine (f :9-18)	91.08	**93.30** (*)	+2.22	98.89	*96.67* (-)	-2.22
10%	91.01	**91.57** (18,15,12)	+1.56	97.19	*96.67* (-)	-0.52
20%	89.38	**91.60** (*)	+2.22	95.52	**96.67** (*)	+1.15
30%	85.98	**92.71** (*)	+6.83	95.52	**96.67** (*)	+1.15
40%	80.88	**91.04** (*)	+10.16	94.90	**96.67** (*)	+1.77
50%	82.61	**87.06** (*)	+4.45	89.31	**95.55** (*)	+6.22
Moyenne			**+3.08**			**+2.31**
10%			**+2.61**			**+2.32**
20%			**+2.93**			**+1.65**
30%			**+4.04**			**+1.79**
40%			**+6.68**			**+0.30**
50%			**+3.65**			**+1.56**

TAB. 1: Comparaison de précisions : données originales v.s. améliorées par NTFC.

Nous remarquons que la précision de C4.5 est affectée par le niveau de bruit croissant. Nous observons jusqu'à 20% de perte de précision. Les résultats expérimentaux montrent que NTFC & C4.5 résiste presque toujours mieux aux bruits que C4.5 seul ; sauf pour des cas

de niveau extrême de bruit (50%) sur les données `breast-Wisconsin` et `iris`. Notons que l'amélioration de la précision n'est souvent pas le fait d'une valeur unique de seuil de fréquence. `NB` est d'origine moins affecté par le bruit (la perte de précision est moins importante). Ainsi, l'amélioration par le processus `NTFC` est moins flagrante mais tout de même significative. Sur les 40 versions de données, `NTFC & NB` (resp. `NTFC & C4.5`) réalise une amélioration de la précision pour 35 (resp. 37) d'entre elles. Notons aussi l'amélioration de la précision par `NTFC` sur les données non-artificiellement bruitées (en moyenne +3.08 (resp. +2.31) par rapport à `C4.5` (resp. `NB`)). La plupart des cas où `NTFC` ne réalise pas d'amélioration sont des cas où le niveau de bruit est de 50%. Nous pensons que ces cas d'échec sont dus à la destruction complète par le bruit des motifs originaux pertinents. En effet, dans ces cas, le processus `NTFC` ne peut extraire assez d'ensembles δ-libres pertinents pour caractériser les objets ; le pire des cas restant bien sûr la génération par le bruit de confusion entre motifs pertinents de différentes classes.

4.2 Evolution de la précision en fonction de γ, δ et du niveau de bruit

En Figure 3, nous reportons les résultats de précision de `NTFC` en fonction de δ pour la base `tic-tac-toe` avec un graphique par niveau de bruit et une courbe par seuil de fréquence. La droite constante représente la précision de `C4.5` ou `NB`.

Premièrement, nous remarquons que la précision obtenue avec `NTFC` augmente (pas nécessairement de façon monotone) avec δ jusqu'à un point maximal – souvent meilleur que le même classifieur sur les données originales. Puis, la précision décroît pour `NTFC & C4.5` et un peu plus brutalement pour `NTFC & NB`. En fait, `C4.5` choisit successivement le meilleur motif pour construire ses noeuds de test ainsi que l'arbre de décision. Ainsi, tous les ensembles extraits ne sont pas nécessairement utilisés dans l'arbre. Au contraire, `NB` utilise tous les ensembles extraits pour le calcul du produit d'approximations. C'est pourquoi, la précision de `NTFC & NB` décroît plus vite avec δ. Notons que ce comportement est semblable pour les autres données et accentué lorsque le niveau de bruit augmente. Deuxièmement, lorsque le niveau de bruit augmente, les valeurs faibles de δ engendrent les précisions les plus basses. En effet, lorsque $\delta \rightarrow 0$, nous allons extraire des motifs qui peuvent être rares dans les données bruitées (ou erronés si l'on abaisse la fréquence). De plus, les valeurs intéressantes de δ (i.e., pour lesquelles `NTFC & C4.5` ou `NB` est meilleur) sont différentes selon la quantité de bruit. Étant donné un seuil de fréquence, le paramétrage du nombre d'erreurs autorisées semble crucial. Dans la suite, nous donnons des indications pour déterminer le seuil δ.

4.3 Stratégie pour fixer δ

Le paramétrage automatique du seuil de fréquence reste une question ouverte (voir Cerf et al. (2008); Zhang et al. (2008) pour des travaux préliminaires). Soit un seuil de fréquence γ, comment déterminer les bonnes valeurs pour δ ? L'évolution des mesures δ-dépendantes (telle que la mesure TA) en fonction de δ est intuitive et connue. Lorsque δ diminue, TA augmente pour des motifs (respectant les contraintes de γ et δ) qui peuvent être rares dans les données bruitées. Lorsque δ augmente, les motifs extraits deviennent plus aptes à s'adapter au bruit dans les données d'apprentissage mais pour de plus grandes valeurs de δ, elles seront moins pertinentes car TA sera faible. Dans les graphiques de la Figure 4, nous reportons les résultats

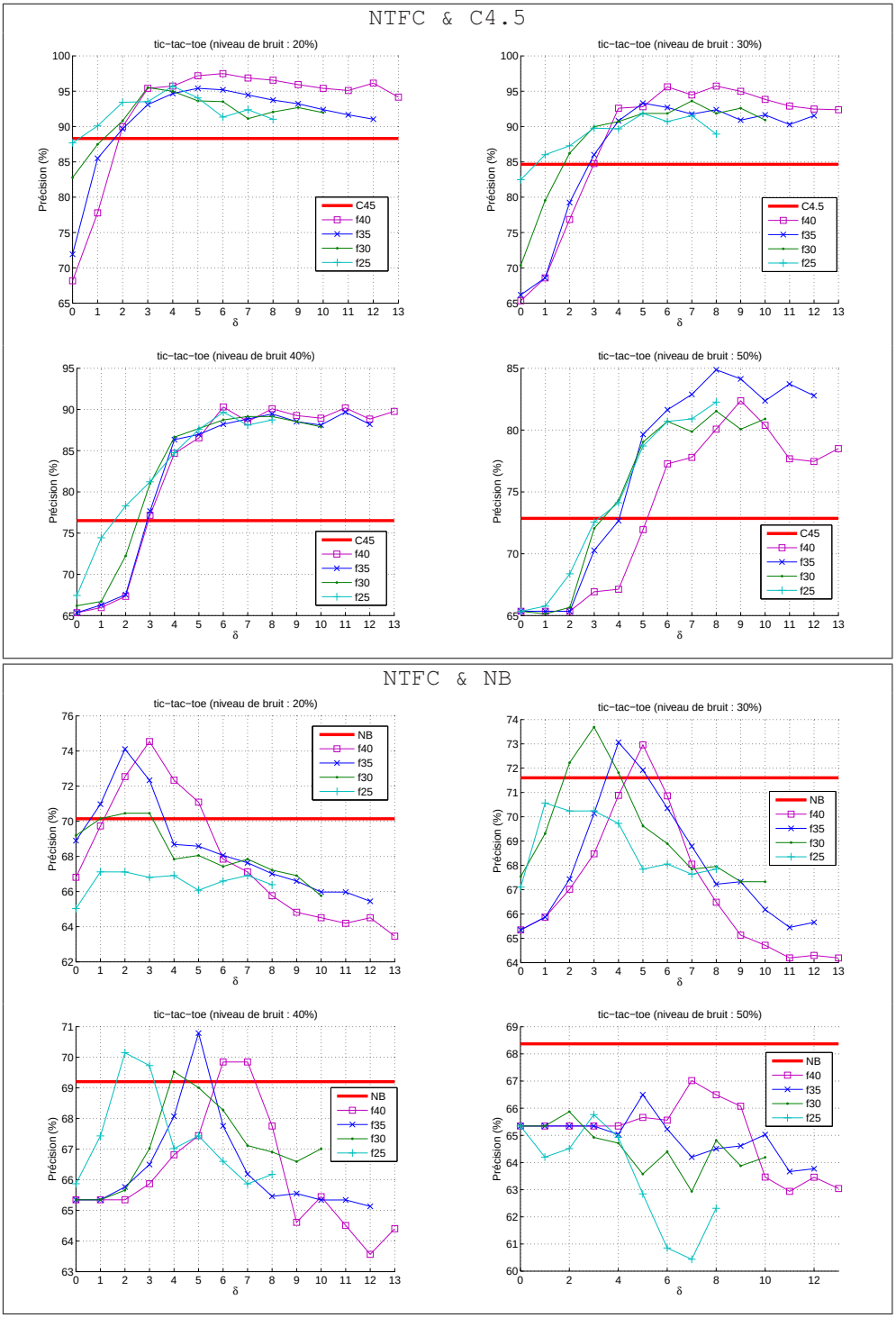


FIG. 3: Evolution de la précision en fonction de γ, δ, et du niveau de bruit pour les données `tic-tac-toe`.

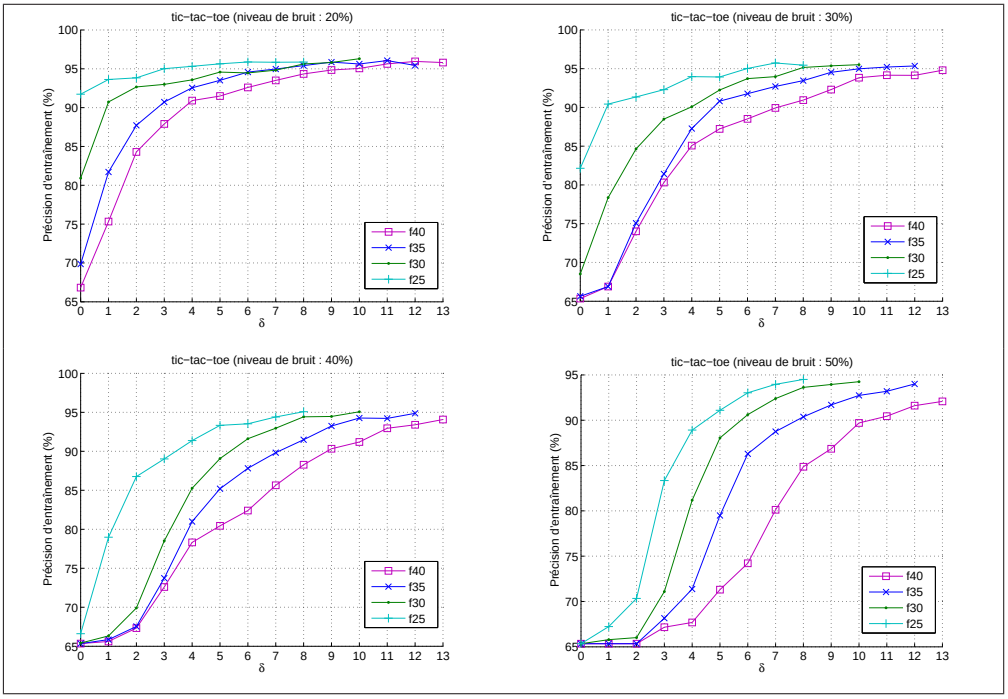


FIG. 4: Evolution de la précision de `NTFC & C4.5` sur l'ensemble d'apprentissage (`tic-ta-toe`) en fonction de δ pour différents seuils de fréquence et niveaux de bruits.

de précision de `NTFC & C4.5` sur la base d'apprentissage (pour `tic-tac-toe`) en fonction de δ. D'une manière générale, la précision augmente avec δ jusqu'à stabilisation (ou fort ralentissement de la croissance). Les valeurs de δ de stabilisation (noté δ_s) sont intéressantes car pour $\delta < \delta_s$, `NTFC & C4.5` est moins performant et $\delta > \delta_s$ n'apporte pas d'amélioration significative. Notons que δ_s dépend du niveau de bruit dans les données. Puisque dans les cas réels, la quantité de bruit dans les données n'est pas connu a priori, nous proposons une méthode raisonnable pour accéder à de *bonnes* valeurs pour δ : (1) poser $\delta = 0$; (2) augmenter δ jusqu'à stabilisation (ou fort ralentissement de la croissance) de la précision de `NTFC & C4.5` sur les données d'apprentissage.

5 Conclusion

Nous proposons un processus original baptisé `NTFC` pour améliorer la qualité des tâches de classification supervisée en présence d'exemples d'apprentissage bruités. Il combine des avancées récentes dans les domaines de l'extraction de motifs tolérants aux exceptions et de la construction de descripteurs. Nous améliorons la description de données initialement booléennes grâce à de nouveaux descripteurs construits sur des motifs ensemblistes pertinents et robustes, plus précisément à partir d'ensembles γ-fréquents δ-libres et de leurs δ-fermetures. Les résultats expérimentaux montrent que, dans un environnement bruité, des classifieurs cal-

culés au moyen d'algorithmes classiques comme NB and C4.5 sont plus performants sur des données améliorées par NTFC que sur les données d'apprentissage originales. De plus, nous proposons une stratégie pour choisir des valeurs de δ pertinentes pour la quantité de bruit dans les données. Les perspectives de travail sont nombreuses. Nous nous sommes restreints au bruit des attributs non-classe mais il est également possible d'étudier l'effet de nos nouveaux descripteurs dans des environnements où l'attribut classe est également soumis à un bruit. D'autre part, nous pourrions considérer l'ensemble des règles δ-fortes de caractérisation extraites comme base pour corriger les exemples peu bruités ou comme filtre pour éliminer les exemples considérés comme très bruités.

Remerciements : Ce travail a été partiellement financé par le contrat européen IST-FET IQ FP6-516169 et l'ANR MDCO-2007 Bingo2.

Références

Bastide, Y., R. Taouil, N. Pasquier, G. Stumme, et L. Lakhal (2000). Mining frequent patterns with counting inference. *SIGKDD Explorations 2*(2), 66–75.

Besson, J., R. G. Pensa, C. Robardet, et J.-F. Boulicaut (2006). Constraint-based mining of fault-tolerant patterns from boolean data. In *Proceedings KDID'05*, Volume 3933 of *LNCS*, pp. 55–71. Springer.

Boulicaut, J.-F., A. Bykowski, et C. Rigotti (2000). Approximation of frequency queries by means of free-sets. In *Proceedings PKDD'00*, Volume 1910 of *LNCS*, pp. 75–85. Springer.

Cerf, L., D. Gay, N. Selmaoui, et J.-F. Boulicaut (2008). A parameter free associative classifier. In *Proceedings DaWaK'08*, Volume 5182 of *LNCS*, pp. 238–247. Springer.

Cheng, H., X. Yan, J. Han, et C.-W. Hsu (2007). Discriminative frequent pattern analysis for effective classification. In *Proceedings ICDE'07*, pp. 716–725. IEEE Computer Society Press.

Crémilleux, B. et J.-F. Boulicaut (2002). Simplest rules characterizing classes generated by delta-free sets. In *Proceedings BCS SGAI ES'02*, pp. 33–46. Springer.

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

Fan, H. et K. Ramamohanarao (2004). Noise tolerant classification by chi emerging patterns. In *Proceedings PAKDD'04*, Volume 3056 of *LNCS*, pp. 201–206. Springer.

Fayyad, U. M. et K. B. Irani (1993). Multi-interval discretization of continous-valued attributes for classification learning. In *Proceedings IJCAI'93*, pp. 1022–1027. Morgan Kaufmann.

Garriga, G. C., P. Kralj, et N. Lavrac (2008). Closed sets for labeled data. *Journal of Machine Learning Research 9*, 559–580.

Gay, D., N. Selmaoui, et J.-F. Boulicaut (2008). Feature construction based on closedness properties is not that simple. In *Proceedings PAKDD'08*, Volume 5012 of *LNCS*, pp. 112–123. Springer.

Hébert, C. et B. Crémilleux (2006). Optimized rule mining through a unified framework for interestingness measures. In *Proceedings DaWaK'06*, Volume 4081 of *LNCS*, pp. 238–247. Springer.

Kubica, J. et A. W. Moore (2003). Probabilistic noise identification and data cleaning. In *Proceedings ICDM'03*, pp. 131–138. IEEE Computer Society Press.

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

Pensa, R. G., C. Robardet, et J.-F. Boulicaut (2006). Supporting bi-cluster interpretation in 0/1 data by means of local patterns. *Intelligent data analysis 10*(5), 457–472.

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

Ramamohanarao, K. et H. Fan (2007). Patterns based classifiers. *World Wide Web 10*(1), 71–83.

Rebbapragada, U. et C. E. Brodley (2007). Class noise mitigation through instance weighting. In *Proceedings ECML'07*, Volume 4701 of *LNCS*, pp. 708–715. Springer.

Selmaoui, N., C. Leschi, D. Gay, et J.-F. Boulicaut (2006). Feature construction and delta-free sets in 0/1 samples. In *Proceedings DS'06*, Volume 4265 of *LNCS*, pp. 363–367. Springer.

Sun, Q., X. Zhang, et K. Ramamohanarao (2003). Noise tolerance of EP-based classifiers. In *Proceedings AusAI'03*, Volume 2903 of *LNCS*, pp. 796–806. Springer.

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

Yang, C., U. M. Fayyad, et P. S. Bradley (2001). Efficient discovery of error-tolerant frequent itemsets in high dimensions. In *Proceedings ACM SIGKDD'01*, pp. 194–203. ACM Press.

Yang, Y., X. Wu, et X. Zhu (2004). Dealing with predictive-but-unpredictable attributes in noisy data sources. In *Proceedings PKDD'04*, Volume 3202 of *LNCS*, pp. 471–483. Springer.

Zhang, S., X. Wu, C. Zhang, et J. Lu (2008). Computing the minimum-support for mining frequent patterns. *Knowledge and Information Systems 15*(2), 233–257.

Zhang, Y. et X. Wu (2007). Noise modeling with associative corruption rules. In *Proceedings IEEE ICDM'07*, pp. 733–738. IEEE Computer Society Press.

Zhu, X. et X. Wu (2004). Class noise vs. attribute noise : A quantitative study. *Artificial Intelligence Revue 22*(3), 177–210.

Summary

When training classifiers, the presence of noise can severely harm their performance. One may differentiate class noise from attribute noise. The earlier has been extensively studied while a few methods have been developed to handle the latter. We focus on attribute noise and we consider robust feature construction based on a frequent fault-tolerant pattern mining task. Indeed, our method is based on an application independent strategy for feature construction based on the so-called δ-free pattern type which has been proposed earlier as an approximate condensed representations of frequent sets. Our experimental evaluation on noisy training data sets shows accuracy improvement when using the computed features instead of the original ones.