Échantillonnage actif pour la découverte de règles classification via des comparaisons par paires

Tudor Matei Opran*, Samir Loudni*,

*TASC – DAPI, IMT-Atlantique, LS2N – CNRS, Nantes, France
{tudor-matei.opran@imt-atlantique.net, samir.loudni@imt-atlantique.fr}

Résumé. Dans cet article, nous proposons une nouvelle méthode de fouille de données interactive, dédiée à l'extraction de règles de classification. Elle combine un apprentissage interactif des préférences multi-critères, basé sur l'intégrale de Choquet, avec une exploration guidée de l'espace des règles à l'aide d'une technique d'échantillonnage MCMC. Cette approche permet d'identifier efficacement les paires de règles présentant le plus d'incertitudes, qui sont ensuite soumises à l'utilisateur pour comparaison. Nous analysons les propriétés de convergence de la chaîne de Markov associée et démontrons, sous certaines hypothèses, que la probabilité d'échantillonnage d'une règle augmente proportionnellement à son score. Des expériences réalisées sur des jeux de données de l'UCI montrent que notre méthode converge plus rapidement vers des règles pertinentes, en comparaison avec la technique décrite dans (Vernerey et al., 2024).

1 Introduction

En fouille de motifs, l'objectif s'est déplacé d'une extraction exhaustive et efficace de tous les motifs vers des méthodes visant à identifier les motifs les plus significatifs. Face à des ensembles de résultats volumineux, l'analyste doit investir un effort considérable pour identifier les motifs pertinents par rapport à ses intérêts, ce qui nécessite souvent une expertise approfondie. Cela met en évidence la nécessité de formaliser et d'automatiser l'apprentissage des préférences des utilisateurs dans un processus de fouille interactif. À cette fin, de nombreuses approches ont adopté le cadre du Learning to Rank (LTR) (Liu, 2009), en modélisant les préférences utilisateurs au moyen de fonctions de classement de motifs.

L'intégration des retours utilisateurs pour prioriser les motifs d'intérêt offre un cadre prometteur. Cependant, cela soulève des défis, notamment la nécessité de minimiser le nombre de comparaisons tout en garantissant la pertinence des motifs identifiés. L'échantillonnage actif offre une approche efficace pour diminuer le nombre d'évaluations nécessaires en sélectionnant de manière dynamique les motifs les plus pertinents à analyser (Settles, 2009; Mikhailiuk et al., 2021). En particulier, les techniques d'échantillonnage actif basées sur des comparaisons par paires permettent de capturer les préférences relatives entre motifs candidats en s'appuyant sur des modèles probabilistes comme les modèles bayésiens (Xu et al., 2018) et les extensions des modèles de Thurstone (1927) et Bradley et Terry (1952). Ces modèles probabilistes fournissent un cadre rigoureux pour quantifier l'incertitude dans les préférences exprimées et optimiser la sélection des paires à comparer (Chu et Ghahramani, 2005; Chen et al., 2016).

Cet article propose une nouvelle méthode de fouille de données interactive, dédiée à l'extraction de règles de classification. Elle combine un apprentissage interactif des préférences multi-critères, basé sur l'intégrale de Choquet (Grabisch et Roubens, 2000), avec une exploration guidée de l'espace des règles candidates à l'aide d'une technique d'échantillonnage MCMC. Cette approche permet d'identifier efficacement les paires de règles présentant le plus d'incertitudes, qui sont ensuite soumises à l'utilisateur pour comparaison. Nous analysons les propriétés de convergence de la chaîne de Markov associée et démontrons, sous certaines hypothèses, que la probabilité d'échantillonnage d'une règle augmente avec son score. Des expériences réalisées sur des jeux de données de l'UCI montrent que notre méthode converge plus rapidement en comparaison avec la technique décrite dans Vernerey et al. (2024).

2 Préliminaires

Fouille de motifs Soit un jeu de données transactionnelle $\mathcal{D}$, $\mathcal{I}$ l'ensemble des n *items* de $\mathcal{D}$ et $\mathcal{T} = \{1, ..., m\}$ l'ensemble des transactions. Chaque transaction t est un sous-ensemble d'items, *i.e.* $t \subseteq \mathcal{I}$. Nous définissons un motif X comme un sous-ensemble non vide de $\mathcal{I}$ et sa couverture $\boldsymbol{t}_{\mathcal{D}}(X)$ est égale à l'ensemble des transactions qui le supportent, *i.e.* $\boldsymbol{t}_{\mathcal{D}}(X) = \{t \in \mathcal{D} \mid X \subseteq t\}$. Le langage des motifs par rapport à $\mathcal{I}$ correspond à $\mathcal{L}_{\mathcal{I}} = 2^{\mathcal{I}} \backslash \emptyset$. L'une des tâches les plus connues de la fouille est l'*extraction de règles d'association* (Agrawal et Srikant, 1994). Une règle d'association est une implication $R : R_a \Rightarrow R_c$, où R_a est appelé l'antécédent, R_c le conséquent, avec $R_a \cap R_c = \emptyset$ et $R_c \neq \emptyset$. Dans ce travail, nous considérons le cas où R_c est réduit à une étiquette de classe, c'est-à-dire $R_c \in C = \{c_1, c_2, ..., c_k\}$.

L'intégrale de choquet Soit un ensemble de n alternatives $A = \{a_1, \ldots, a_n\}$, évaluées sur une famille de m critères G. La performance d'une alternative a selon le critère $f_j : A \mapsto \mathbb{R}$ est notée $f_j(a)$. Sans perte de généralité, nous supposons que $f_j(a) \in [0, 1]$. L'intégrale de Choquet s'appuie sur la notion de mesure floue ou mesure non additive (aussi appelée capacité) (Grabisch et Roubens, 2000). Pour un ensemble de critères G, une mesure floue est définie, pour chaque sous-ensemble de G, comme étant une fonction $\mu : 2^G \mapsto [0, 1]$ respectant les conditions de normalisation ($\mu(\emptyset) = 0,\ \mu(G) = 1$) et de monotonie ($\forall S, T \subseteq G$, $S \subseteq T$, $\mu(S) \leq \mu(T)$). L'intégrale de Choquet est alors définie comme une fonction $C_\mu : A \mapsto \mathbb{R}$: $C_\mu(a) = \sum_{j=1}^{m} (f_{(j)}(a) - f_{(j-1)}(a))\mu(G(i))$, où $(.)$ est une permutation de $\{1, \ldots, m\}$ telle que : $f_{(0)}(a) = 0 \leq f_{(1)}(a) \leq \ldots \leq f_{(m)}(a)$, et où $G(j) = \{f_{(1)}(a), \ldots, f_{(j)}(a)\}$ représente l'ensemble des j-premiers critères triés par ordre croissant. Le matériel supplémentaire fournit un exemple illustrant le calcul de la valeur de l'intégrale de Choquet pour une alternative donnée. Soit $u(\cdot) : \mathcal{L}_{\mathcal{I}} \times C \mapsto A$ une fonction d'utilité qui retourne le vecteur de performance r d'une règle R, noté $u(R) = (f_1(R), \ldots, f_m(R)) = r$.

Les approches d'apprentissage actif pour l'extraction interactive de motifs s'appuient sur les retours utilisateur, exprimés sous forme de comparaisons entre motifs, pour modéliser leurs préférences au moyen du cadre de Learning-to-Rank (LTR) (Liu, 2009). L'objectif principal est d'apprendre une fonction de classement qui reflète avec précision les choix de l'utilisateur. Ce modèle, une fois appris, est utilisé pour découvrir de nouveaux motifs de meilleure qualité, à mesure que la fonction de classement s'affine tout au long du processus d'apprentissage actif. En supposant l'existence d'un classement cible spécifique à l'utilisateur, noté $\mathcal{R}$, sur l'ensemble $\mathcal{L}_{\mathcal{I}}$, où $X \succ Y$ indique que l'utilisateur trouve subjectivement le motif X plus intéressant que le motif Y, le problème peut être formulé comme suit :

Algorithm 1 Learning to Rank Interesting Decision rules : LETRID

Entrée : Jeu de données transactionnel $\mathcal{D}$, Nombre d'itérations d'apprentissage L, Nombre d'itérations d'échantillonnage k, Certitude de surclassement $\theta(\cdot,\cdot)$, ℓ Nombre de règles à présenter à l'utilisateur
Sortie : Fonction de classement $\mathbf{g}$
1: $\mathcal{U} \leftarrow \emptyset$, $\mathbf{g}_0 \leftarrow$ CHOIXALEACAPA()
2: **Pour** $t = 1$ à L **faire**
3: Requête $Q_t \leftarrow$ CHOISIRREQUÊTE(SIMAS($\mathbf{g}_{t-1}, \theta(\cdot,\cdot),k,\ell$)) ▷ Échantillonner des règles et sélectionner
4: $\mathcal{U} \leftarrow \mathcal{U} \cup$ RETOURUTILISATEUR(Q_t) ▷ Poser la requête Q^t à l'utilisateur et recueillir son retour
5: $\mathbf{g}_t \leftarrow$ APPRENDREMODÈLE($\mathcal{U}$)
6: **Retourner** La fonction finale $\mathbf{g}$

Problème 1 (Learning-to-Rank). Soit $\mathcal{U}$ un ensemble (initialement vide) de comparaisons par paires fournies par l'utilisateur, r le vecteur d'utilité associé à une règle R, et $\mathbf{g} : A \mapsto [0,1]$ une fonction de préférence telle que $\mathbf{g}(r_1) > \mathbf{g}(r_2)$ représente la préférence $R_1 \succ R_2$. L'objectif est d'apprendre $\mathbf{g}$ à partir des comparaisons $\mathcal{U}$, de manière à minimiser la divergence entre le classement produit par $\mathbf{g}$ et le classement cible $\mathcal{R}$.

Une étape clé du *Problème* 1 est d'identifier les paires de règles les plus informatives pour solliciter les retours de l'utilisateur, ce qui est coûteux en raison de la taille de l'espace de recherche. Dans cet article, Nous utilisons une approche par échantillonnage pour sélectionner directement les règles les plus pertinentes selon une fonction de score $\mathbf{g}$, sans exploration exhaustive.

3 Comment échantillonner des règles intéressantes ?

Extraire des règles selon une certaine mesure d'intérêt est nécessaire, aussi bien pour la phase d'apprentissage de la fonction d'ordonnancement $\mathbf{g}(\cdot)$ que pour l'exploitation de cette dernière. Une des contributions de cet article est l'introduction d'un algorithme permettant l'extraction de règles, ou de paires de règles, en s'appuyant sur une mesure d'intérêt avec des hypothèses relâchées. Nous présentons des garanties théoriques et analysons les faiblesses de cette approche, dénommée SIMAS.

A) L'algorithme d'échantillonnage SIMAS repose sur une approche gloutonne probabiliste visant à extraire des règles de décision qui maximisent une mesure d'intérêt. Cela peut être réalisé de manière équivalente soit en échantillonnant à partir d'une distribution de probabilité pré-définie, comme démontré par Qian et al. (2016), soit en sélectionnant de manière probabiliste la règle ayant le score le plus élevé. Dans l'algorithme 2, la certitude du modèle $\theta(\cdot,\cdot)$ est utilisée pour évaluer si une règle est préférée à une autre. Plus précisément, pour un sous-ensemble donné $R_a \in \mathcal{L}_\mathcal{I}$ et un élément $a \in I$, il s'agit de de comparer la règle avec l'antécédent $R_a \cup \{a\}$ à celle avec l'antécédent $R_a \setminus \{a\}$. Le choix d'ajouter a à l'antécédent de la règle actuelle $R_a \setminus \{a\}$ est alors modélisé comme un essai de Bernoulli, avec une probabilité définie par $\theta(\cdot,\cdot)$. L'algorithme initialise la première règle de manière aléatoire. Pour des raisons de commensurabilité, nous employons la normalisation active (Lima et Souza, 2023). Nous initialisons les statistiques de normalisation, telles que le score minimal et maximal, sur un échantillon aléatoire de règles. À chaque nouvelle règle observée, nous mettons à jour ces statistiques. Nous maintenons un historique des n premières règles par score décroissant à l'aide de la fonction UPDATETOPRULE. Pour toute règle ayant pour vecteur de performance r,

on désignera indifféremment par . Pour toute règle R associée à un vecteur de performance r, on notera indifféremment $\mathbf{g}(R)$ ou $\mathbf{g}(r)$ le score de la règle en fonction de ses performances.

B) Les fonctions de certitude de surclassement $\theta(\cdot,\cdot) : [0,1]^2 \mapsto [0,1]$ sont des fonctions de probabilité définies sur l'ensemble des comparaisons par paires. Elles évaluent la certitude du modèle, $\theta(\mathbf{g}(r_1), \mathbf{g}(r_2))$, que $R_1 \succ R_2$. Ces fonctions possèdent les propriétés suivantes :

(i) Si la certitude que $R_1 \succ R_2$ est élevée, alors la certitude que $R_2 \succ R_1$ est faible : $\forall (r_1, r_2) \in A^2$, $\theta(\mathbf{g}(r_1), \mathbf{g}(r_2)) + \theta(\mathbf{g}(r_2), \mathbf{g}(r_1)) = 1$.

(ii) Si $R_1 \succ R_2$, alors, pour toute règle R_3 associée à un vecteur de performance r_3 $\theta(\mathbf{g}(r_3), \mathbf{g}(r_1)) < \theta(\mathbf{g}(r_3), \mathbf{g}(r_2))$.

Théorème 1 (Voir démonstration dans le Mat. Suppl.)**.** *Pour une fonction de certitude $\theta(\cdot,\cdot)$ donnée, une chaîne de Markov irréductible associée possède une distribution à l'équilibre.*

Théorème 2 (Voir démonstration dans le Mat. Suppl.)**.** *Sous certaines hypothèses, la probabilité d'une règle à l'équilibre et croissante avec son score* $\mathbf{g}(\cdot)$.

Le théorème 2 stipule que les règles ayant un score maximal auront une probabilité plus élevée d'apparaître dans la distribution à l'équilibre. Cependant, la mesure d'intérêt utilisée détermine la façon dont les scores des règles sont calculés et, par conséquent, influence directement la probabilité d'échantillonner une règle. En effet, plus cette mesure attribue un score maximal à un grand nombre de règles, plus la probabilité d'échantillonner une règle ayant un score maximal sera plus élevée. Par ailleurs, le temps de convergence de la chaîne de Markov vers les meilleures règles dépend également de sa connectivité. Plusieurs heuristiques de connectivité sont explorées dans le matériel supplémentaire.

C) Échantillonnage glouton par incertitude (GUS) à pour objectif d'identifier la paire pour laquelle le modèle est le plus incertain, selon une fonction de certitude $\Theta(\cdot,\cdot)$. Les fonction de certitude de différenciation $\Theta(\cdot,\cdot) : [0,1]^2 \mapsto [0,1]$ sont définies sur les scores des règles et sont invariantes à l'ordre des arguments, ce qui signifie que pour toute paire de règles (R_1, R_2) associée aux vecteurs de performance (r_1, r_2), nous avons : $\Theta(\mathbf{g}(r_1), \mathbf{g}(r_2)) = \Theta(\mathbf{g}(r_2), \mathbf{g}(r_1))$. De plus, elles attribuent une valeur nulle lorsque les deux règles sont indiscernables, soit : $\Theta(\mathbf{g}(r_1), \mathbf{g}(r_2)) = 0$ si $\mathbf{g}(r1) = \mathbf{g}(r_2)$. Pour échantillonner la paire de règles la plus incertaine, nous introduisons une fonction $d(R, \mathcal{S}, \Theta(\cdot,\cdot), \mathbf{g}(\cdot))$, qui mesure l'incertitude maximale d'une règle R associée au vecteur de performance r, par rapport à un sous-ensemble $\mathcal{S}$ de règles existantes. Formellement, la fonction $d(R, \mathcal{S}, \Theta(\cdot,\cdot), \mathbf{g}(\cdot))$ est définie comme suit :

$$d(R, \mathcal{S}, \Theta(\cdot,\cdot), \mathbf{g}(\cdot)) = \begin{cases} 1 - \max\limits_{R' \in \mathcal{S}} (\Theta(g(r), g(r'))) & \text{si } R \notin \mathcal{S} \\ 0 & \text{sinon} \end{cases}$$

Étant donné un ensemble de règles $\mathcal{S}$ et une nouvelle règle R à évaluer, la fonction $\delta(\cdot) = d(\cdot, \mathcal{S}, \Theta(\cdot,\cdot), \mathbf{g}(\cdot))$ attribue un score plus élevé aux règles dans $\mathcal{L}_{\mathcal{I}} \times C \setminus \mathcal{S}$ qui forment une paire fortement incertaine avec l'une des règles de $\mathcal{S}$. L'algorithme 3 présente notre heuristique GUS pour sélectionner la paire de règles où l'incertitude du modèle est la plus élevée. Il prend en entrée une règle initiale et définit, à partir d'une fonction multivariée $\Theta(\cdot,\cdot)$, une fonction univariée $\delta(\cdot) = d(\cdot, \mathcal{S}, \Theta(\cdot,\cdot), \mathbf{g}(\cdot))$, en fixant les paramètres de la fonction d, $\Theta(\cdot,\cdot)$ pour l'intégralité de l'algorithme et $\mathcal{S}$ pour une seule itération. À chaque itération de l'algorithme 3, l'algorithme 2 est appelé avec la fonction univariée $\delta(\cdot)$, en fixant le nombre d'itérations de

Algorithm 2 Échantillonnage d'un argument maximisant une fonction univarié (SIMAS)

Entrée : $\mathbf{g}(\cdot), \theta(\cdot,\cdot), k, \ell \in \mathbb{N}$
$R_a \leftarrow$ RANDOMINITANT(), $R_c \leftarrow$ RANDOMINITCONS(), $\mathcal{H} \leftarrow$ INITTASMAX(ℓ)
INITACTIVENORMALIZATION(RANDOMRULES(100))
Pour k itérations **faire**
  Pour $c \in \sigma_k(C)$ **faire** ▷ Échantillonner le conséquent
    UPDATEACTIVENORMALIZATION($R_a \Rightarrow R_c$) ▷ Mettre à jour les statistiques de normalisation
    Si ISVALID($R_a \Rightarrow c$) **alors**
      $p \leftarrow \theta(\mathbf{g}(R_a \Rightarrow c), \mathbf{g}(R_a \Rightarrow R_c))$
      Si Trial($X \sim \mathcal{B}(p)$) **alors**
        $R_c \leftarrow \{c\}$
        UPDATETOPRULE($\mathcal{H}, \{R_a \Rightarrow R_c\}$) **break** ▷ Actualiser l'ensemble $\mathcal{H}$ des meilleures règles
  Pour $a \in \sigma_k(\mathcal{I})$ **faire** ▷ Échantillonner l'antécédent
    UPDATEACTIVENORMALIZATION($R_a \Rightarrow R_c$) ▷ Mettre à jour les statistiques de normalisation
    Si ISVALID($R_a \cup \{a\} \Rightarrow R_c$) **alors**
      $p \leftarrow \theta(\mathbf{g}(R_a \cup \{a\} \Rightarrow R_c), \mathbf{g}(R_a \Rightarrow R_c))$
      Si Trial($X \sim \mathcal{B}(p)$) **alors**
        $R_a \leftarrow R_a \cup \{a\}$
        UPDATETOPRULE($\mathcal{H}, \{R_a \Rightarrow R_c\}$) ▷ Actualiser l'ensemble $\mathcal{H}$ des meilleures règles
      sinon
        $R_a \leftarrow R_a \setminus \{a\}$
Retourner $\arg\max_{r \in \mathcal{H}}(\mathbf{g}(r))$

avec $\mathbf{g}(\cdot) : [0,1]^n \mapsto [0,1]$ le modèle appris, $\theta(\cdot,\cdot) : [0,1]^2 \mapsto [0,1]$ une fonction de certitude de surclassement, et $(\sigma_i(\cdot))_{1 \leq i \leq k}$ une suite de permutations aléatoires, ℓ la taille du tas max.

SIMAS à 10 et en ne considérant que la meilleure règle i.e. $\ell = 1$. Nous avons borne la taille de $\mathcal{S}$ à 10 règles afin de diversifier l'exploration de différentes régions de l'espace de recherche. L'intuition derrière cet algorithme est qu'à chaque itération, nous souhaitons échantillonner une règle qui peut créer une paire incertaine avec l'une des règles présentes dans $\mathcal{S}$. En supprimant la règle la plus ancienne de l'échantillon $\mathcal{S}$, nous permettons à $\mathcal{S}$ d'évoluer et, espérons-le, de converger vers un sous-ensemble de règles pouvant constituer des paires très incertaines.

Échantillonnage des scores élevés (HSS) est une approche visant à sélectionner une paire d'alternatives ayant les scores les plus élevés possibles. L'intuition est que la comparaison d'alternatives avec des scores élevés améliore la précision du modèle à ordonner ces alternatives. Nous utilisons SIMAS pour $10,000$ itérations et sélectionnons les deux règles ayant les meilleurs scores.

4 Expérimentations

Protocole expérimental Nous simulons les retours des utilisateurs à l'aide d'une fonction de classement cachée basée sur la mesure d'intérêt χ^2. Les performances de classement sont évaluées avec la précision moyenne sur quatre ensembles de données réels issus de la bibliothèque UCI. Les détails du protocole expérimental sont disponibles dans le matériel supplémentaire. Les expériences sont implémentées en Java [1].

1. Disponible à `https://github.com/Tudor1415/EGC25-LetRID`

Algorithm 3 Échantillonnage Glouton par Incertitude (GUS)

Entrée : Règle initiale R_0, mesure de certitude de différenciation $\Theta(\cdot,\cdot)$, fonction de préférence $\mathbf{g}(\cdot)$, $\lambda \in \mathbb{N}$, le nombre d'itérations.
$\mathcal{S} \leftarrow \{R_0\}$, paire $\leftarrow (\emptyset, \emptyset)$
Pour i de 1 à λ itérations **faire**
  $\delta(\cdot) = d(\cdot, \mathcal{S}, \Theta(\cdot,\cdot), \mathbf{g}(\cdot))$ ▷ Définition d'une fonction univariée à partir d'une fonction multivariée Θ
  $R_i \leftarrow \text{SIMAS}(\delta(\cdot),\, \theta(\cdot,\cdot)\,, k = 10, \ell = 1)$ ▷ Éch. une règle formant une paire incertaine avec $\mathcal{S}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{R_i\}$
  Si $\text{SIZE}(\mathcal{S}) > 10$ **alors**
    REMOVEOLDEST($\mathcal{S}$)
  $(R_1, R_2) \leftarrow \underset{R_s, R_t \in \mathcal{S}}{\arg\min}\ \Theta(\mathbf{g}(r_s), \mathbf{g}(r_t))$ ▷ Sélectionner la paire la plus incertaine
  paire $\leftarrow \underset{(R_s,R_t)\in\{\text{paire},(R_1,R_2)\}}{\arg\min}\ \Theta(g(r_s), g(r_t))$
Retourner paire

TAB. 1 – *Les fonctions de certitude de surclassement et différenciation utilisées*

Type	$\boldsymbol{\theta(a,b)}$	$\boldsymbol{\Theta(a,b)}$
Affine	$\frac{a-b+1}{2}$	$\lvert a-b\rvert$
Logistic (Bradley et Terry, 1952)	$\frac{\exp a}{\exp a+\exp b}$	$\left\lvert\frac{\exp a-\exp b}{\exp a+\exp b}\right\rvert$
Normal c.d.f. (Thurstone, 1954)	$\Phi(a-b)$	$\lvert\Phi(a-b)-\Phi(b-a)\rvert$

Q1) Comparaison de SIMAS avec GIBBSSAMPLING-QIAN. Dans l'algorithme de Qian et al. (2016), l'échantillonnage s'effectue à partir d'une distribution de probabilité connue. Dans notre cas, cette distribution est construite à partir de $\theta(\cdot,\cdot)$ et de l'ordre de considération des items. En contrepoint à GIBBSSAMPLING-QIAN, notre algorithme se restreint aux règles valides, ne satisfaisant donc pas la propriété d'équilibre détaillée de la chaîne de Markov associée. Cette restriction évite l'exploration de règles invalides au support nul mais n'assure pas une exploration complète de l'espace de recherche. Pour cela, il faut relancer l'algorithme itérativement en partant de règles issues d'espaces non connectés. Concernant la Fig. 1, TOMS comprend deux fois moins de transactions pour neuf fois plus de variables que CONNECT, ce qui nous laisse penser qu'il comprend également davantage de règles invalides. Expérimentalement, GIBBSSAMPLING-QIAN se perd parmi les états invalides, ce qui explique l'absence d'états valides enregistrés sur TOMS, biaisant la moyenne des états explorés. Sur CONNECT, où le nombre de règles invalides semble être moins important, GIBBSSAMPLING-QIAN converge, bien que plus lentement, vers des états valides et inévitablement intéressants. Cette expérience montre que notre restriction de l'espace de recherche permet une convergence plus rapide que GIBBSSAMPLING-QIAN. Nous constatons que les trois fonctions de certitude $\theta(\cdot,\cdot)$ présentent un comportement très similaire ; toutefois, SIMAS-AFFINE semble converger plus rapidement.

Q2) Comparaison de LETRID avec CHOQUETRANK. La Fig. 2 montre l'évolution de l'AP@1% sur 100 itérations d'apprentissage interactif (cf. Algorithme 1). LETRID, avec l'heuristique GUS, permet d'atteindre la meilleure qualité d'apprentissage en termes de précision moyenne à 1% et 10%. Ce résultat s'explique par deux principaux facteurs : tout d'abord, un espace de recherche élargi qui induit un temps de calcul plus élevé et une plus grande diversité des règles générées ; ensuite, une optimisation plus efficace de l'incertitude des paires

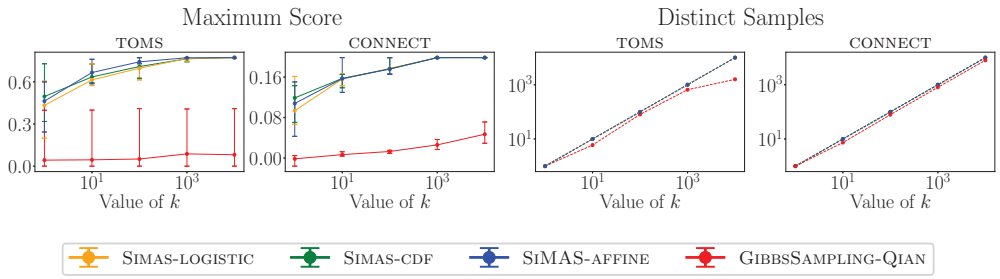

FIG. 1 – *Comparing* SIMAS *with* GIBBSSAMPLING-QIAN *without user feedback.*

sélectionnées. On remarque que GUS parvient à atteindre les plus grandes valeurs de précision moyenne en moins de cinquante itérations.

Q3) Comparaison sur la diversité Pour une paire de motifs, l'indice de Jaccard est un bon indicateur de diversité. Pour autant, cet indice n'est pas utilisable pour un ensemble de motifs. Nous utilisons les fonctions de répartition cumulatives (ou Cumulative Distribution Function (CDF)) sur les indices de Jaccard pour obtenir une forme visuelle de la distribution des paires de motifs en fonction de leurs indices de Jaccard. Pour un ensemble de motifs $X = \{X_1, X_2, \ldots, X_k\}$ et une base de données D :
$\text{CDF}(D, X, \theta) = \frac{|\{(i,j):\text{Jac}(X_i,X_j)\leq\theta,\, 1\leq i<j\leq k\}|\cdot 2}{k(k-1)}$. La Fig. 2 visualise la diversité des règles présentées. Nous observons que, pour les méthodes d'échantillonnage où l'espace de recherche est plus grand, il y a une diversité accrue.

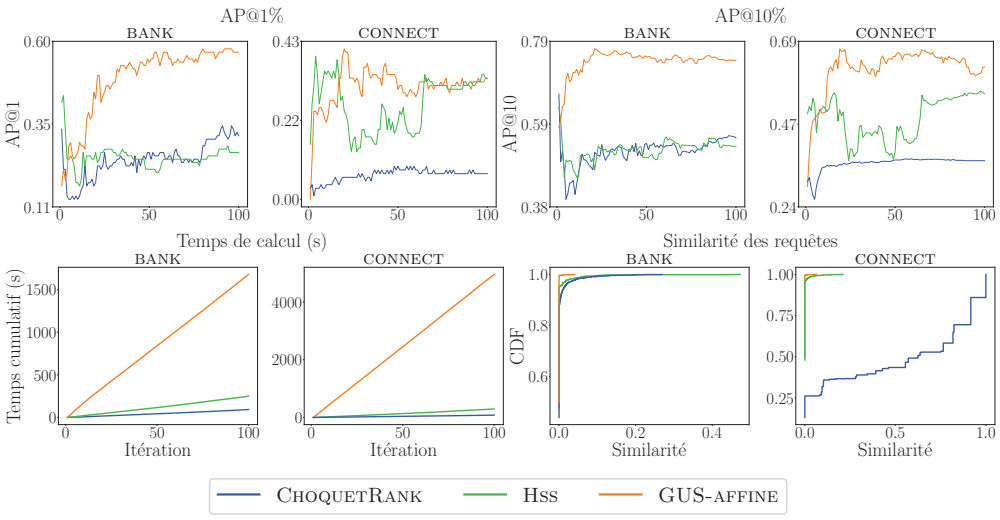

FIG. 2 – *Résultats d'apprentissage actif sur les jeux de données Connect 4 et Bank, avec l'oracle* χ^2*. Précision moyenne à 1% (à gauche) et à 10% (à droite). En dessous, le temps de calcul en secondes et la distribution de la similarité de Jaccard sur les couvertures des règles présentées à l'utilisateur.*

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB*, San Francisco, CA, USA, pp. 487–499.

Bradley, R. et M. Terry (1952). Rank analysis of incomplete block designs : I. the method of paired comparisons. *Biometrika 39*, 324.

Chen, X., K. Jiao, et Q. Lin (2016). Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *J. Mach. Learn. Res. 17*, 217 :1–217 :40.

Chu, W. et Z. Ghahramani (2005). Preference learning with gaussian processes. *Proceedings of the 22nd international conference on Machine learning*, 137–144.

Grabisch, M. et M. Roubens (2000). Application of the Choquet integral in multicriteria decision making. In *Fuzzy Measures and Integrals - Theory and Applications*, pp. 348–374.

Lima, F. T. et V. M. Souza (2023). A large comparison of normalization methods on time series. *Big Data Research 34*, 100407.

Liu, T. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr. 3*(3), 225–331.

Mikhailiuk, A., C. Wilmot, M. Perez-Ortiz, D. Yue, et R. Mantiuk (2021). Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *IEEE International Conference on Pattern Recognition (ICPR)*.

Qian, G., C. Rao, X. Sun, et Y. Wu (2016). Boosting association rule mining in large datasets via gibbs sampling. *Proceedings of the National Academy of Sciences 113*, 201604553.

Settles, B. (2009). Active learning literature survey.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review 34*, 273–286.

Thurstone, L. L. (1954). The measurement of values. *Psychological review 61*(1), 47.

Vernerey, C., N. Aribi, S. Loudni, Y. Lebbah, et N. Belmecheri (2024). Learning to rank based on choquet integral : Application to association rules. In *PAKDD 2024*, pp. 313–326.

Xu, Q., J. Xiong, X. Chen, Q. Huang, et Y. Yao (2018). Hodgerank with information maximization for crowdsourced pairwise ranking aggregation. In *Proceedings of AAAI 2018*, pp. 4326–4334. AAAI Press.

Summary

In this paper, we present a novel interactive data mining method for extracting classification rules. It combines interactive multi-criteria preference learning based on the Choquet integral with guided exploration of the rule space through an MCMC sampling technique. This approach efficiently identifies pairs of rules with the highest uncertainty, which are then presented to the user for comparison. We investigate the convergence properties of the associated Markov chain and show that, under certain conditions, the probability of sampling a rule increases in proportion to its score. Experiments on UCI datasets demonstrate that our method converges more rapidly to relevant rules when compared to the technique presented in (Vernerey et al., 2024).