Qualité et complexité en évaluation des mesures d'intérêt

Bruno Crémilleux*, Arnaud Giacometti**, Arnaud Soulet**

*Université de Caen Basse-Normandie, GREYC UMR 6072, Campus 2, 14000 Caen
bruno.cremilleux@unicaen.fr
**Université François-Rabelais de Tours, LI EA 6300, Campus de Blois, 41000 Blois
prenom.nom@univ-tours.fr

Résumé. Remplacer des hypothèses sur le modèle de données par des informations mesurées sur les données réelles est l'une des forces de la fouille de données. Cet article étudie cet ajustement entre les données et les méthodes de découverte de motifs pour en évaluer la qualité et la complexité. Nous formalisons ce lien entre données et mesures d'intérêt en identifiant les motifs *liés* qui sont ceux nécessaires pour l'évaluation d'une mesure ou d'une contrainte. Nous formulons alors trois axiomes que devraient satisfaire ces motifs liés pour qu'une méthode d'extraction se comporte bien. En outre, nous définissons la complexité en évaluation qui quantifie finement l'interrelation entre les motifs au sein d'une méthode d'extraction. A la lumière de ces axiomes et de cette complexité en évaluation, nous dressons une typologie de multiples méthodes de découverte de motifs impliquant la fréquence.

1 Introduction

La découverte de motifs locaux introduite par Agrawal et Srikant (1994) consiste à extraire des informations pertinentes décrivant une portion des données. Evaluer et garantir la qualité des motifs extraits demeure une problématique très ouverte malgré le nombre important de propositions (Giacometti et al., 2013). Chacune de ces propositions repose explicitement ou implicitement sur une mesure d'intérêt dont la qualité dépend de la complexité du modèle sous-jacent et de son ajustement aux données. Le modèle repose en général sur des fondements statistiques dont la complexité et la compréhension sont bien connues. A l'inverse, l'ajustement aux données reste une notion difficile à appréhender. Pourtant, c'est probablement cette notion qui distingue la fouille de données des statistiques traditionnelles. L'ajustement aux données est souvent connoté négativement et synonyme de sur-apprentissage par rapport aux données. De notre point de vue, l'ajustement aux données n'est pas un biais d'apprentissage mais un moyen pour lever certaines hypothèses sur le modèle en les remplaçant par des mesures sur les données. Nous proposons d'étudier l'ajustement aux données à travers les interrelations entre motifs lors de l'évaluation d'une mesure d'intérêt ou d'une contrainte d'extraction.

La qualité d'une mesure repose sur sa capacité à isoler un motif singulier qui dévie des autres motifs communs. Pour cette raison, une mesure se doit de mettre en relation le *motif évalué* avec d'autres motifs, dits *motifs liés*. Par exemple, la confiance de la règle d'association $X \rightarrow Y$ met en relation la fréquence de $X \cup Y$ (motif évalué) par rapport à la fréquence

de X (motif lié). La qualité de la règle augmente avec la fréquence de $X \cup Y$ tandis qu'elle diminue si la fréquence de X augmente (lorsque les autres fréquences restent constantes). Ces variations de la confiance sont en partie conformes aux deux axiomes formulés par Piatetsky-Shapiro (1991). De manière intéressante, ces axiomes permettent d'étudier formellement le comportement des mesures d'intérêt dédiées à l'évaluation des règles d'association. Dans cet article, nous proposons de généraliser ce principe en introduisant la notion de motif lié pour s'attaquer à l'évaluation de n'importe quelle méthode de découverte de motifs.

L'objectif de ce travail est de formaliser la qualité et la sémantique des mesures d'intérêt et contraintes en analysant les interrelations entre les motifs nécessaires à l'évaluation de chaque motif extrait. Ce travail s'inscrit dans la lignée des travaux sur l'analyse des propriétés formelles vérifiées par les mesures d'intérêt Piatetsky-Shapiro (1991); Tan et al. (2004); Geng et Hamilton (2006); Lenca et al. (2008); Hämäläinen et al. (2010) en les étendant aux contraintes d'extraction. Nous formaliserons l'interrelation entre motifs en introduisant l'ensemble de motifs liés. Cet ensemble regroupe tous les motifs susceptibles d'impacter l'évaluation d'un motif donné. Nous distinguerons les *motifs liés positivement* qui permettent d'accroître la mesure d'intérêt de ceux qui la font décroître, i.e., les *motifs liés négativement*. Nous formulerons alors trois axiomes que devraient satisfaire une mesure d'intérêt ou contrainte. Chacun de ces axiomes impose des contraintes topologiques que doivent respecter les deux ensembles de motifs liés. Enfin, nous proposerons des critères d'analyse de la complexité et de la sémantique d'une mesure d'intérêt ou contrainte. Nous introduirons finalement la complexité en évaluation qui repose sur la cardinalité de l'ensemble des motifs liés.

2 Travaux relatifs

A notre connaissance, très peu de travaux se sont intéressés à l'interrelation entre les motifs lors de l'évaluation d'une mesure d'intérêt ou d'une contrainte. De manière plus générale, l'évaluation de la qualité des méthodes de découverte de motifs est une tâche ardue et peu étudiée.

2.1 Protocoles expérimentaux

La plupart des méthodes d'évaluation ou d'analyse de la qualité concernant la découverte de motifs repose sur des protocoles expérimentaux où l'objectif est de vérifier la conformité du résultat avec un étalon-or. Dans un contexte supervisé, il est possible d'exploiter directement la variable cible comme référence. Ensuite, le cadre précision/rappel, l'analyse ROC (Fawcett, 2006), la validation croisée (Kohavi, 1995), etc sont utilisés pour évaluer l'écart entre les motifs extraits et la référence. L'évaluation de la qualité des méthodes de découverte de motifs dans un contexte non-supervisé s'avère bien plus difficile comme le rappellent de Lin et Chalupsky (2004). En effet, la validation des motifs extraits ne peut pas s'appuyer sur un étalon-or explicite. Plusieurs stratégies sont alors mises en oeuvre pour obtenir un succédané de cet étalon-or.

Premièrement, il est possible de s'appuyer sur la connaissance d'experts d'un domaine (Carvalho et al., 2005). Avec un cas d'utilisation, des experts du domaine sont sollicités pour juger la justesse des motifs découverts et ainsi, évaluer la méthode d'extraction. La stratégie de redécouverte teste si un processus parvient à retrouver les connaissances bien établies dans un

certain domaine. Deuxièmement, il est parfois possible de construire l'étalon-or (Gupta et al., 2008; Zimmermann, 2013). Par exemple, Gupta et al. (2008) propose un protocole d'évaluation quantitative des algorithmes d'extraction de motifs approximés fréquents : (1) l'extraction des motifs fréquents dans un jeu de données classique, (2) l'ajout de bruit dans ce jeu de données, (3) l'extraction des motifs approximés fréquents et enfin, (4) la comparaison des véritables motifs fréquents avec les motifs approximés fréquents. Ici, l'étape 1 explicite l'étalon-or. Enfin, l'hypothèse nulle peut parfois être construite expérimentalement. Gionis et al. (2007) décrivent une technique de randomisation pour évaluer la pertinence des résultats d'exploration de données. Webb (2008) propose de mettre de côté une partie des données pour jouer le rôle d'étalon-or afin de contrôler le risque d'erreur de type 1 : (1) diviser l'ensemble de données initial en données d'apprentissage et données de validation, (2) générer des motifs candidats à partir des données d'apprentissage et (3) réaliser une évaluation statistique de chaque candidat sur les données de validation.

Ces protocoles expérimentaux sont clairement utiles, mais néanmoins ils souffrent de plusieurs limites. Tous ces protocoles expérimentaux reposent sur l'évaluation de la collection de motifs extraits. Ils ne peuvent donc fournir qu'un résultat a posteriori, i.e., après l'implémentation d'un prototype et de son application sur un jeu de données. L'évaluation est forcément dépendante du jeu de données considéré et les résultats sont difficilement généralisables à n'importe quel jeu de données, de n'importe quel domaine.

2.2 Outils formels

Plusieurs outils formels ont été proposés dans la littérature pour analyser qualitativement les méthodes de découverte de motifs. Premièrement, la taille des représentations condensées est souvent utilisée comme une mesure objective d'évaluation de leur intérêt (Calders et al., 2004). Par exemple, une représentation condensée fondée sur les motifs fermés est toujours plus compacte qu'une représentation condensée fondée sur les motifs libres. Les motifs fermés sont donc jugés comme plus intéressants. Cependant, les représentations condensées les plus compactes ne sont pas forcément les plus utilisées. Par exemple, les itemsets non-dérivables (NDI) sont rarement utilisés malgré leur taux de compression impressionnant. La sémantique complexe des NDI expliquerait cette impopularité pour certaines personnes. *Comment identifier formellement cette complexité ?*

Deuxièmement, les mesures d'intérêt doivent satisfaire un certain nombre d'axiomes pour être considérées comme pertinentes. Piatetsky-Shapiro (1991); Tan et al. (2004); Geng et Hamilton (2006); Lenca et al. (2008) ont proposé des axiomes pour les mesures d'intérêt dédiées aux règles d'association et plus récemment, Webb et Vreeken (2013); Hämäläinen et al. (2010) les ont étendus aux motifs ensemblistes. A notre connaissance, de tels axiomes n'ont jamais été appliqués à des contraintes ou des algorithmes de construction de modèles. De plus, ils se sont essentiellement concentrés sur les mesures dédiées à la recherche de corrélations. *Comment généraliser ces axiomes à n'importe quelle mesure d'intérêt ou contrainte ?*

Enfin, à notre connaissance seuls deux travaux ont étudié l'interrelation entre les motifs lors de l'évaluation d'une mesure d'intérêt. Crémilleux et Soulet (2008) ont défini informellement la notion de contrainte globale. Il s'agit de contraintes dont l'évaluation met en correspondance plusieurs motifs. Giacometti et al. (2011) ont ensuite formalisé cette notion de contrainte globale en utilisant une algèbre relationnelle étendue spécifiquement pour la découverte de motifs. Notre cadre propose une définition formelle plus générale et plus précise, mais surtout permet

de mieux analyser l'interrelation entre les motifs lors de l'évaluation. En particulier, nous répondrons aux deux questions énoncées ci-avant.

3 Formalisation de l'interrelation entre motifs

Nom (référence)	Définition
	Corrélation
Support (Agrawal et Srikant, 1994)	$freq(X)/freq(\emptyset)$
All-confidence Bond (Omiecinski, 2003)	$freq(X)/\max_{i\in X} freq(\{i\})$ $freq(X)/\lvert\{t \in \mathcal{D} \lvert X \cap t \neq \emptyset\}\rvert$
Productivive itemset (Webb et Vreeken, 2013)	$(\forall Y \subset X)(prod(X) \Rightarrow supp(X) > supp(Y) \times supp(X \setminus Y))$
	Non-redondance
Motif non-fréq. minimal Motif fréquent maximal (Mannila et Toivonen, 1997)	$(\forall Y \subset X)(\mathcal{B}d^{-}\text{-}freq(X) \Rightarrow freq(X) < \gamma \wedge freq(Y) \geq \gamma)$ $(\forall Y \supset X)(\mathcal{B}d^{+}\text{-}freq(X) \Rightarrow freq(X) \geq \gamma \wedge freq(Y) < \gamma)$
Motif libre (Boulicaut et al., 2003)	$(\forall Y \subset X)(free(X) \Rightarrow freq(X) < freq(Y))$
Motif fermé (Pasquier et al., 1999)	$(\forall Y \supset X)(closed(X) \Rightarrow freq(X) > freq(Y))$
Motif non-dérivable (Calders et Goethals, 2007)	$(\forall X \in \mathcal{L})(ndi(X) \Leftrightarrow LB(X,\mathcal{D}) \neq UB(X,\mathcal{D}))$ où $LB(X,\mathcal{D})$ et $UB(X,\mathcal{D})$ sont respectivement les bornes inférieure et supérieure dérivables avec les sous-ensembles de X dans $\mathcal{D}$
	Modèle
Motif top-k fréquent (Fu et al., 2000)	$(\forall X \in \mathcal{L})(top_k(X) \Leftrightarrow \lvert\{Y \in \mathcal{L} : freq(Y) > freq(X)\}\rvert < k)$
Bouncer and Picker (Bringmann et Zimmermann, 2009)	Algorithme de sélection avec différentes heuristiques

TAB. 1 – *Définition de plusieurs méthodes de découverte de motifs fondées sur la fréquence*

3.1 Méthode de découverte de motifs

Dans cet article, nous modélisons une méthode de découverte de motifs par une mesure d'intérêt M définie sur $\mathcal{L} \times \Delta$ vers $\Re$ où :

- le *langage* $\mathcal{L}$ correspond à l'ensemble des parties de $\mathcal{I}$ (i.e., $\mathcal{L} = 2^{\mathcal{I}}$) [1] et
- Δ est l'ensemble de tous les *jeux de données* possibles sachant qu'un jeu de données est un multi-ensemble de $\mathcal{L}$.

1. Par simplicité, nous restreignons cet article aux motifs ensemblistes mais les définitions et axiomes sont généralisables à tout langage.

Si le jeu de données visé est clair, $M(X, \mathcal{D})$ est simplement noté $M(X)$. Sans perte de généralité, nous considérons que l'intérêt d'un motif X augmente avec $M(X)$ i.e., si le motif X est plus intéressant que Y alors $M(X) > M(Y)$.

Notre modélisation d'une méthode de découverte de motifs par une mesure d'intérêt M est suffisamment souple pour englober les principaux outils de la littérature :

- **Evaluation par mesure d'intérêt** : l'évaluation par mesure d'intérêt consiste à affecter un score ou un rang à chaque motif reflétant sa qualité (e.g., la all-confidence proposée par Omiecinski (2003)).
- **Extraction sous contraintes** : l'extraction sous contraintes consiste à extraire tous les motifs satisfaisant un prédicat de sélection qui détermine la pertinence d'un motif. Par exemple, il est courant d'utiliser un seuil minimal sur une mesure d'intérêt pour filtrer les motifs (e.g., la contrainte de support minimal (Agrawal et Srikant, 1994)). Lorsque M est une contrainte, M ne prend que deux valeurs à savoir $\{0, 1\}$ correspondant respectivement à faux et vrai.
- **Construction de modèle** : la construction de modèle consiste à sélectionner un ensemble E de motifs ayant un intérêt collectif (e.g., classifieur ou clustering). La condition d'appartenance à cet ensemble E peut être vue comme une contrainte d'extraction. Typiquement, Bringmann et Zimmermann (2009) présente un algorithme de sélection où différentes heuristiques peuvent être choisies.

Notre cadre permettra donc de comparer entre elles des méthodes d'extraction assez variées. Le tableau 1 donne des exemples de méthodes d'extraction afin d'illustrer la suite de notre travail. Il se dégage trois familles de méthodes : les méthodes destinées à trouver des motifs corrélés ; les méthodes visant à réduire les redondances entre les motifs extraits et les méthodes visant à modéliser le jeu de données.

3.2 Notion de motifs liés

La qualité d'une méthode de découverte de motifs repose sur sa capacité à isoler un motif singulier qui dévie des autres motifs communs. Pour cette raison, cette méthode doit mettre en relation le motif *évalué* avec d'autres motifs, dits *liés*. Pour analyser cette interrelation entre motifs, nous proposons de déterminer son ensemble de motifs liés en identifiant l'impact de chacun de ces motifs sur le motif évalué. L'impact d'un motif Y sur le motif évalué X peut se mesurer en observant s'il existe deux jeux de données $\mathcal{D}$ et $\mathcal{D}'$ quasi-équivalents où la seule variation de Y modifie l'évaluation de X. Typiquement, la *all-confidence* d'un itemset X introduite par Omiecinski (2003) correspond à la plus petite confiance des règles d'association $Y \rightarrow Z$ incluses dans X. Il est bien connu que la all-confidence peut être réécrite comme le ratio entre la fréquence de X et la fréquence maximale de ses items : $freq(X)/\max_{i \in X} freq(\{i\})$. Dans ce cas, quand X est évalué, les motifs liés de X sont lui-même et tous ses items. En effet, augmenter la fréquence d'un item peut faire décroître la all-confidence de X. A l'inverse, augmenter la fréquence de X augmente aussi sa all-confidence. Cet exemple conduit à deux observations d'importance :

1. Il y a deux catégories de motifs liés : ceux qui peuvent améliorer la qualité du motif évalué (ici, X) et ceux qui peuvent la détériorer (ici, les items qui constituent X).
2. Les motifs liés impactent la mesure d'intérêt qu'on analyse (ici, la all-confidence) via une autre mesure d'intérêt *élémentaire* (ici, la fréquence).

Suivant ces deux observations, nous définissions formellement la notion de motif lié en nous appuyant sur la définition d'équivalence avec exception :

Définition 1 (Equivalence avec exception) *Les jeux de données $\mathcal{D}$ et $\mathcal{D}'$ sont équivalents par rapport à m excepté en Y, noté $\mathcal{D} \sim_{m/Y} \mathcal{D}'$ si et seulement si $m(Y,\mathcal{D}) \neq m(Y,\mathcal{D}')$ et pour tout motif $X \neq Y$, on a $m(X,\mathcal{D}) = m(X,\mathcal{D}')$.*

Définition 2 (Motifs liés) *Pour une mesure d'intérêt élémentaire m, l'ensemble des motifs liés positivement (ou négativement) pour une mesure M d'un motif évalué X, dénoté par $M_m^+(X)$ (ou $M_m^-(X)$), contient tous les motifs Y tels que $M(X)$ croît (ou décroît) quand $m(Y)$ augmente (et m reste inchangé pour tous les autres motifs) :*

$$M_m^+(X) = \{Y \in \mathcal{L} | \exists \mathcal{D}, \mathcal{D}' : M(X,\mathcal{D}') > M(X,\mathcal{D}) \wedge m(Y,\mathcal{D}') > m(Y,\mathcal{D}) \wedge \mathcal{D} \sim_{m/Y} \mathcal{D}'\}$$

$$M_m^-(X) = \{Y \in \mathcal{L} | \exists \mathcal{D}, \mathcal{D}' : M(X,\mathcal{D}') < M(X,\mathcal{D}) \wedge m(Y,\mathcal{D}') > m(Y,\mathcal{D}) \wedge \mathcal{D} \sim_{m/Y} \mathcal{D}'\}$$

Détaillons le principe de cette définition. Tout d'abord, une mesure d'intérêt élémentaire m sert à impliquer les motifs liés pour modifier la mesure M. Cette mesure m est en général un indicateur assez simple impliqué dans M. Notons que comme pour M, on considère que la qualité de X augmente avec celle de m. Par exemple, avec la all-confidence, cette mesure m correspond à la fréquence. Les motifs liés positivement correspondent donc aux motifs $Y \in \mathcal{L}$ qui font augmenter $M(X)$ (i.e., $M(X,\mathcal{D}') > M(X,\mathcal{D})$) lorsqu'on considère un jeu de données $\mathcal{D}'$ quasi-équivalent à $\mathcal{D}$ selon m. En effet, seul $m(Y)$ a été augmenté (i.e., $m(Y,\mathcal{D}') > m(Y,\mathcal{D})$) car m reste inchangée pour tous les autres motifs (i.e., $\mathcal{D} \sim_{m/Y} \mathcal{D}'$). Nous employons l'expression « *ensemble des motifs liés* » , dénoté par $M_m^{\pm}(X)$, pour désigner l'union des ensembles des motifs liés positivement et négativement : $M_m^{\pm}(X) = M_m^-(X) \cup M_m^+(X)$.

Illustrons la définition 2 avec la all-confidence. Comme la all-confidence ne peut croître qu'avec la fréquence de X, $all\text{-}conf_{freq}^+(X)$ est égal à $\{X\}$. Nous verrons qu'il est courant voire souhaitable que le motif évalué soit aussi un motif lié. Pour l'ensemble des motifs liés négativement, nous obtenons que $all\text{-}conf_{freq}^-(X) = \{\{i\} | i \in X\}$ car seuls l'augmentation de la fréquence d'un item de X peut faire diminuer $all\text{-}conf(X)$. Le tableau 2 donne d'autres exemples d'ensembles de motifs liés. Une force de la notion de motifs liés est de bien identifier les motifs « réellement » impliqués dans l'évaluation. Par exemple, la définition des motifs libres donnée dans le tableau 1 implique tous les sous-ensembles de X (avec le $\forall Y \subset X$). Pourtant, seuls les sous-ensembles directs sont des motifs liés.

Dans le tableau 2, les mesures, contraintes, algorithme de construction de modèles choisis reposent exclusivement sur la fréquence (comme mesure m qui implique les motifs liés). Il est à noter que les définitions de motifs libres/fermés auraient pu être présentées avec d'autres mesures élémentaires (e.g., fréquence disjonctive ou fonction d'agrégat). Les bordures négative et positive pourraient être analysées avec d'autres mesures suivant la contrainte monotone ou anti-monotone considérée (Mannila et Toivonen, 1997). De même, la notion de top-k motif est pertinente avec d'autres mesures d'intérêt que la fréquence. En changeant la mesure élémentaire m introduite dans la définition 2, l'analyse des interrelations entre motifs se ferait de manière analogue.

Nom	$M^+_{freq}(X)$	$M^-_{freq}(X)$	A1	A2	A1+2	A3	$\lvert M^{\pm}_{freq}(X)\rvert$
Support	$\{X\}$	$\{\emptyset\}$	$\times$	$\times$			$O(1)$
All-confidence	$\{X\}$	sing.	$\times$	$\times$			$O(k)$
Bond	$\{X\}$	sing.	$\times$	$\times$			$O(k)$
Productive Itemset	$\{X\}$	ss-ens.	$\times$	$\times$			$O(2^k)$
Motif non-fréq. min.	ss-ens. directs	$\{X\}$		$\times$	$\times$		$O(k)$
Motif fréq. max.	$\{X\}$	sur-ens. directs	$\times$	$\times$			$O(n-k)$
Motif libre	ss-ens. directs	$\{X\}$		$\times$	$\times$		$O(k)$
Motif fermé	$\{X\}$	sur-ens. directs	$\times$	$\times$			$O(n-k)$
NDI	ss-ens.	ss-ens.					$O(2^k)$
Top-k fréquent	$\{X\}$	$\{Y \in \mathcal{L} \lvert Y \not\subseteq X \wedge X \not\subseteq Y\}$	$\times$	$\times$			$O(2^n - 2^k - 2^{n-k})$
Bouncer and Picker	$\{X\}$	treillis	$\times$	$\times$		$\times$	$O(2^n)$

où $k = |X|$ et $n = |\mathcal{I}|$; *singletons* correspond à $\{\{i\} | i \in X\}$; *sous-ensembles directs* correspond à $\{X \setminus \{i\} | i \in X\}$; *sur-ensembles directs* correspond à $\{X \cup \{i\} | i \in \mathcal{I} \setminus X\}$; sous-ensembles correspond à $2^X \setminus \{X\}$; treillis correspond à $\mathcal{L} \setminus \{X\}$.

TAB. 2 – *Analyse des méthodes suivant leurs ensembles de motifs liés*

4 Axiomes de qualité

En s'inspirant de ce qui a été fait pour les mesures d'intérêt dédiées aux règles d'association, cette section énonce trois axiomes que devraient satisfaire une mesure ou une contrainte idéale. Le tableau 2 illustre la satisfaction ou non des axiomes par les différentes mesures et contraintes.

4.1 Réflexivité

Revenons sur l'exemple de la all-confidence. Nous avons constaté que l'augmentation de la fréquence de certains motifs avait un impact sur la all-confidence de X. Donc la fréquence est une mesure élémentaire d'importance pour la all-confidence. Par ailleurs, il est souvent considéré que l'intérêt d'un motif augmente avec sa fréquence. Il paraît donc naturel qu'augmenter la fréquence de X augmente également la all-confidence de X. Plus généralement, si l'intérêt d'un motif X augmente avec la mesure élémentaire m, l'intérêt de ce motif X selon la mesure d'intérêt M devrait également augmenter lorsque $m(X)$ croît.

Axiome 1 (Réflexive) *Une mesure d'intérêt M est réflexive par rapport à m ssi tout motif est lié positivement à lui-même pour m : $\forall X \in \mathcal{L}, X \in M^+_m(X)$.*

Comme $all\text{-}conf^+_{freq}(X) = \{X\}$, la all-confidence est bien une mesure réflexive par rapport à la fréquence. A l'inverse, l'extraction des motifs libres n'est pas réflexive par rapport à la fréquence puisque $free^+_{freq}(X) = \{X \setminus \{i\} | i \in X\}$ n'inclut pas X. Tandis qu'un motif est jugé plus intéressant quand sa fréquence augmente, il a moins de chance d'être libre (car sa fréquence sera plus proche de celle de ses sous-ensembles). Par conséquent, la

contrainte de liberté pourrait même être qualifiée d'*irréflexive* par rapport à la fréquence (i.e., $X \in free^{-}_{freq}(X)$). Nous reviendrons dessus dans la sous-section suivante.

L'axiome 1 est une généralisation de plusieurs propositions de la littérature où la mesure m est restreinte au support. Pour les règles d'associations, Piatetsky-Shapiro (1991) a formulé la propriété suivante : « $M(X \rightarrow Y)$ augmente avec le support de $X \cup Y$ (quand les autres paramètres restent constants i.e., les supports de X et de Y) ». Webb et Vreeken (2013); Hämäläinen et al. (2010) ont ensuite proposé cette généralisation : « $M(X)$ augmente avec le support de X quand le support de tout $Y \subset X$ reste inchangé ».

4.2 Exclusivité

Pour être facilement compréhensible par l'utilisateur final, le comportement d'une mesure M doit toujours rester le même vis-à-vis de chaque motif lié. En d'autres termes, un motif lié ne devrait pas permettre d'augmenter la mesure M dans certains cas, et de la diminuer dans d'autres.

Axiome 2 (Exclusive) *Une mesure d'intérêt M est exclusive par rapport à m ssi aucun motif est à la fois lié positivement et négativement à un motif donné pour m : $M^{+}_{m}(X) \cap M^{-}_{m}(X) = \emptyset$ pour tout motif $X \in \mathcal{L}$.*

Cet axiome est largement vérifié par les méthodes de la littérature comme le montre le tableau 2 (colonne A2). Par exemple, comme $all\text{-}conf^{+}_{freq}(X) \cap all\text{-}conf^{-}_{freq}(X) = \emptyset$, la all-confidence est exclusive par rapport à la fréquence. A l'inverse, l'extraction des motifs libres fréquents n'est pas exclusive puisque X appartient à la fois à $free \& freq^{+}_{freq}(X)$ à cause de la contrainte de fréquence et à $free \& freq^{-}_{freq}(X)$ à cause de la contrainte de liberté.

Le non-respect de l'axiome 2 complexifie la lecture d'une méthode d'extraction puisqu'il devient nécessaire de se reporter au jeu de données ou à d'autres motifs pour comprendre les motifs extraits. Par exemple, lors de l'extraction des motifs libres fréquents, un motif X peut ne pas être extrait soit si sa fréquence est trop basse, soit si sa fréquence est trop élevée. A l'inverse, pour les motifs fermés fréquents, un motif n'est pas extrait si sa fréquence est trop faible (aussi bien si le motif est non-fréquent ou non-fermé). Nous estimons donc que la violation de l'axiome 2 pourrait expliquer en partie l'échec des motifs NDI même s'ils constituent une représentation condensée extrêmement compacte.

La combinaison des axiomes 1 et 2 (notée A1+2 dans le tableau 2) implique naturellement que $X \notin M^{-}_{m}(X)$. Si une mesure M viole cette propriété, M est dite irréflexive selon m. L'extraction des motifs libres et celle de la bordure négative des motifs fréquents sont irréflexives.

4.3 Exhaustivité

La pertinence d'un motif est d'autant plus forte que son intérêt dépend de la variation de nombreux autres motifs. Pour cette raison, tous les motifs du langage $\mathcal{L}$ devraient avoir un impact sur la mesure M. En d'autres termes, l'ensemble des motifs liés devrait idéalement être égal à la totalité du langage $\mathcal{L}$.

Axiome 3 (Exhaustive) *Une mesure d'intérêt M est exhaustive par rapport à m ssi tous les motifs du langage sont liés à tout motif pour m : $M^{\pm}_{m}(X) = \mathcal{L}$ pour tout motif $X \in \mathcal{L}$.*

Cet axiome exprime que l'interrelation entre les motifs lors de l'évaluation d'une mesure M devrait concerner tous les motifs. Chaque variation du jeu de données mesurable à travers $m(X)$ devrait avoir une incidence sur l'évaluation de M. Bien sûr, la plupart des méthodes d'extraction (mesures ou contraintes de la littérature) ne satisfont pas cet axiome. Cependant, nous pensons que cet axiome donne la direction à suivre et nous revenons longuement dans la section suivante sur la forme de l'ensemble de motifs liés.

5 Complexité et sémantique

En pratique, l'axiome 3 est peu vérifié. Néanmoins, les méthodes d'extraction proposées tendent plus ou moins à le satisfaire. Cette section propose d'étudier deux grandes caractéristiques de l'ensemble des motifs liés à savoir sa taille et sa forme.

5.1 Complexité en évaluation

Suivant l'axiome 3, nous affirmons que la pertinence d'un motif pour une mesure élémentaire m est encore plus forte lorsque sa pertinence dépend de la variation de la pertinence de nombreux autres motifs selon m. Par conséquent, l'intérêt d'une mesure M (au sens de sa globalité) selon m se mesure avec la taille de l'ensemble de motifs liés :

Définition 3 (Complexité en évaluation) *La complexité en évaluation d'une mesure M selon m correspond au comportement asymptotique de la cardinalité de l'ensemble de motifs liés selon m.*

La complexité en évaluation d'une mesure dépend le plus souvent de la cardinalité du motif évalué (notée $k = |X|$) et de la cardinalité de l'ensemble d'items (notée $n = |\mathcal{I}|$). Par exemple, $|\textit{all-conf}^{\pm}_{freq}(X)| = |\textit{all-conf}^{+}_{freq}(X)| \cup |\textit{all-conf}^{-}_{freq}(X)| = 1 + k$. Par conséquent, le comportement du nombre d'évaluations de la all-confidence est linéaire par rapport à la fréquence. De manière similaire, on peut déterminer que la complexité en évaluation de la productivité est exponentielle par rapport à la taille du motif évalué puisque tous les sous-ensembles sont impliqués dans l'évaluation de cette contrainte. Suivant la complexité en évaluation, on dira donc que la productivité est plus intéressante (car plus globale) que la all-confidence car les interrelations sont plus nombreuses.

A notre connaissance, la complexité en évaluation est le premier indicateur pour mesurer l'interrelation entre les motifs lors de l'évaluation d'une mesure d'intérêt. Cette complexité permet de comparer plusieurs mesures d'intérêt entre elles. La colonne $|M^{\pm}_{freq}(X)|$ du tableau 2 indique la complexité en évaluation des mesures et contraintes définies dans le tableau 1. Il se dégage clairement 3 grandes classes correspondant à 3 complexités en évaluation : constant, linéaire et exponentiel.

Bien que le tableau 1 ne comporte qu'un échantillon restreint de la découverte de motifs, la complexité en évaluation des méthodes d'extraction de motifs semble avoir augmenté durant ces deux dernières décennies. Au-delà de l'intérêt des motifs extraits, nous pensons que la complexité en évaluation reflète aussi la difficulté algorithmique à les extraire. Ainsi, l'amélioration des techniques d'extraction pourrait expliquer cette augmentation de la qualité des motifs extraits.

Topologie	$M_m^{\pm}(X)$	Sémantique	Complexité	Classe
itemset	X	–	1	constant
singletons	$\{\{i\}\|i \in X\}$	corrélation	k	linéaire
sous-ensembles directes	$\{X \setminus \{i\}\|i \in X\}$	non-redondance	k	
sur-ensembles directes	$\{X \cup \{i\}\|i \in X\}$	non-redondance	$n-k$	
sous-ensembles	2^X	corr./non-red.	2^k	exponentiel
treillis	$2^{\mathcal{I}}$	modèle	2^n	

où $k = |X|$ et $n = |\mathcal{I}|$

TAB. 3 – *Liens entre l'ensemble des motifs liés, la sémantique et la complexité*

5.2 Sémantique de l'ensemble de motifs liés

Contrainte globale Le tableau 3 schématise les principales topologies observées notamment au sein du tableau 2 en les organisant en trois grandes classes de complexité évoquées dans la section précédente. Ces classes font écho à la notion de contrainte globale introduite par Crémilleux et Soulet (2008) puis définie formellement par Giacometti et al. (2011). Il s'agit des prédicats de sélection dont la complexité est au moins linéaire :

Propriété 1 (Contrainte globale) *Une contrainte $q : \mathcal{L} \rightarrow \{1, 0\}$ est globale ssi il existe une mesure élémentaire m telle que la complexité en évaluation de q selon m est au moins linéaire.*

Notre cadre a l'avantage d'affiner l'analyse des contraintes globales en les séparant en deux grandes classes : celles qui requièrent un nombre linéaire d'évaluations et celles qui requièrent un nombre exponentiel.

Corrélations Les motifs corrélés sont clairement associés aux mesures d'intérêt dont les motifs liés sont les sous-ensembles du motif évalué. Cette observation rappelle l'axiome proposé par Webb et Vreeken (2013); Hämäläinen et al. (2010) : « une mesure d'intérêt $m(X)$ se comporte bien [2] si elle décroît quand $supp(Y)$ augmente pour $Y \subset X$ et que tous les autres paramètres restent inchangés ». Ce dernier peut être reformulé dans notre cadre de la manière suivante :

Propriété 2 (Webb et Vreeken (2013); Hämäläinen et al. (2010)) *Une mesure d'intérêt M se comportant bien doit vérifier $\forall X \in \mathcal{L} : 2^X \setminus \{X\} \subseteq M^-_{freq}(X)$*

Non-redondance Toutes les méthodes de découverte de motifs visant à réduire les redondances exploitent les sous- et/ou sur- ensembles du motif évalué. La majorité des représentations condensées s'appuient exclusivement sur les sous-ensembles ou sur-ensembles *directs*.

Modèle Nous avons constaté que tous les algorithmes de construction de modèles ont leurs motifs liés qui couvrent l'intégralité du treillis comme c'est le cas pour Bouncer and Picker (Bringmann et Zimmermann, 2009). Les modèles sont souvent vus comme une amélioration des représentations condensées. La complexité en évaluation confirme que les modèles sont plus intéressants que les représentations condensées. La complexité des motifs top-k fréquents se rapproche de celle des modèles sans toutefois l'atteindre.

2. Dans ce contexte, « un bon comportement » signifie que les corrélations doivent être évaluées plus favorablement que les non-corrélations.

6 Conclusion

Cet article a introduit la notion de motifs liés qui nous semble centrale pour analyser l'interrelation des motifs pour les méthodes de découverte de motifs. Une force de notre approche est son large spectre d'application qui va au-delà des mesures d'intérêt pour traiter aussi bien l'extraction sous contraintes que la construction de modèles. Pour la première fois, des axiomes de qualité concernent la problématique de la non-redondance. L'introduction de la complexité en évaluation permet de dépasser le stade qualitatif pour mieux comparer plusieurs méthodes.

Plusieurs axes de progression subsistent au sein de notre cadre. La définition actuelle des motifs liés repose sur une notion d'équivalence entre jeux de données où seule une mesure élémentaire m est impliquée dans l'évaluation de M ; comment tenir compte qu'une autre mesure m' peut potentiellement impacter M en parallèle ? Une réflexion sur la définition de mesure élémentaire et des interrelations entre mesures élémentaires est nécessaire pour répondre à cette question. Par ailleurs, d'autres axiomes importants sur les mesures d'intérêts mériteraient d'être étendus en s'appuyant sur la notion de motifs liés. Bien que l'incidence de la mesure élémentaire m dans l'évaluation de M soit cruciale sur la sémantique des motifs liés, nous n'avons pas encore étudié les implications des propriétés de m sur celles de M.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pp. 487–499. Morgan Kaufmann.

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

Bringmann, B. et A. Zimmermann (2009). One in a million : picking the right patterns. *Knowl. Inf. Syst. 18*(1), 61–81.

Calders, T. et B. Goethals (2007). Non-derivable itemset mining. *Data Min. Knowl. Discov. 14*(1), 171–206.

Calders, T., C. Rigotti, et J.-F. Boulicaut (2004). A survey on condensed representations for frequent sets. In J.-F. Boulicaut, L. D. Raedt, et H. Mannila (Eds.), *European Workshop on Inductive Databases and Constraint Based Mining*, Volume 3848 of *LNCS*, pp. 64–80. Springer.

Carvalho, D. R., A. A. Freitas, et N. F. F. Ebecken (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, et J. Gama (Eds.), *PKDD*, Volume 3721 of *Lecture Notes in Computer Science*, pp. 453–461. Springer.

Crémilleux, B. et A. Soulet (2008). Discovering knowledge from local patterns with global constraints. In O. Gervasi, B. Murgante, A. Laganà, D. Taniar, Y. Mun, et M. L. Gavrilova (Eds.), *ICCSA (2)*, Volume 5073 of *Lecture Notes in Computer Science*, pp. 1242–1257. Springer.

de Lin, S. et H. Chalupsky (2004). Issues of verification for unsupervised discovery systems. In *Workshop on Link Analysis and Group Detection (LinkKDD2004) at KDD'04*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Fu, A., R. W., W. Kwong, et J. Tang (2000). Mining *n*-most interesting itemsets. In *proceedings of the 12th International Symposium ISMIS*, Volume 1932 of *LNCS*, USA, pp. 59–67. Springer.

Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining : A survey. *ACM Comput. Surv. 38*(3).

Giacometti, A., D. H. Li, P. Marcel, et A. Soulet (2013). 20 years of pattern mining : a bibliometric survey. *SIGKDD Explorations 15*(1), 41–50.

Giacometti, A., P. Marcel, et A. Soulet (2011). A relational view of pattern discovery. In J. X. Yu, M.-H. Kim, et R. Unland (Eds.), *DASFAA (1)*, Volume 6587 of *LNCS*, pp. 153–167. Springer.

Gionis, A., H. Mannila, T. Mielikäinen, et P. Tsaparas (2007). Assessing data mining results via swap randomization. *TKDD 1*(3).

Gupta, R., G. Fang, B. Field, M. Steinbach, et V. Kumar (2008). Quantitative evaluation of approximate frequent pattern mining algorithms. In Y. Li, B. Liu, et S. Sarawagi (Eds.), *KDD*, pp. 301–309. ACM.

Hämäläinen, W. et al. (2010). *Efficient search for statistically significant dependency rules in binary data*. Ph. D. thesis, University of Helsinki.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pp. 1137–1145.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On selecting interestingness measures for association rules : User oriented description and multiple criteria decision aid. *European Journal of Operational Research 184*(2), 610–626.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov. 1*(3), 241–258.

Omiecinski, E. (2003). Alternative interest measures for mining associations in databases. *IEEE Trans. Knowl. Data Eng. 15*(1), 57–69.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Inf. Syst. 24*(1), 25–46.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pp. 229–248. AAAI/MIT Press.

Tan, P.-N., V. Kumar, et J. Srivastava (2004). Selecting the right objective measure for association analysis. *Inf. Syst. 29*(4), 293–313.

Webb, G. I. (2008). Discovering significant patterns. *Machine Learning 71*(1), 131.

Webb, G. I. et J. Vreeken (2013). Efficient discovery of the most interesting associations. *ACM Trans. Knowl. Discov. Data 8*(3), 15 :1–15 :31.

Zimmermann, A. (2013). Objectively evaluating interestingness measures for frequent itemset mining. In J. Li, L. Cao, C. Wang, K. C. Tan, B. Liu, J. Pei, et V. S. Tseng (Eds.), *PAKDD Workshops*, Volume 7867 of *Lecture Notes in Computer Science*, pp. 354–366. Springer.

Summary

One of the strengths of data mining is to replace assumptions about the data model with information directly measured from real data. This paper analyzes this relationship between the mining process and the data for pattern discovery methods. We formalize this notion by identifying patterns, called linked patterns, which are necessary for the evaluation of a measure or a constraint. We then formulate three axioms that a well-behaving pattern mining method should satisfy. We also define the evaluation complexity that quantifies the data fitness of a method. These axioms and evaluation complexity are illustrated with many examples.