Réflexions sur l'extraction de motifs rares

Sandy Maumus, Amedeo Napoli, Laszlo Szathmary, Yannick Toussaint

Réflexions sur l'extraction de motifs rares

Sandy Maumus[1,2], **Amedeo Napoli**[1], **Laszlo Szathmary**[1], **Yannick Toussaint**[1]

[1] *LORIA, 54506 Vandoeuvre-lès-Nancy*
{*maumus, napoli, szathmar, yannick*}*@loria.fr*
[2] *INSERM U525, 54000 Nancy*
Sandy.Maumus@nancy.inserm.fr

RÉSUMÉ. *Les études en fouille de données se sont surtout intéressées jusqu'à présent à l'extraction de motifs fréquents et à la génération de règles d'association à partir des motifs fréquents. L'algorithme le plus célèbre ayant permis d'atteindre ces objectifs est Apriori, qui a été suivi par toute une famille d'algorithmes mis au point par la suite et possédant tous la caractéristique d'extraire l'ensemble des motifs fréquents ou un sous-ensemble de ces motifs (motifs fermés fréquents, motifs fréquents maximaux, générateurs minimaux). Dans cet article, nous posons le problème de la recherche des motifs rares ou non fréquents, qui se trouvent dans le complémentaire de l'ensemble des motifs fréquents. Ce type de motif n'a jamais vraiment fait l'objet d'une étude systématique, malgré l'intérêt et la demande existant dans certains domaines d'application. Ainsi, en biologie ou en médecine, il peut se révéler très important pour un praticien de repérer des symptômes non habituels ou des effets indésirables exceptionnels se déclarant chez un patient pour une pathologie ou un traitement donnés.*

MOTS-CLÉS : *Fouille de données, extraction de motifs, motifs fréquents et rares.*

1. Introduction

La fouille de données a pour objectif d'identifier des motifs et des associations implicites dans de grandes bases de données [HAN 01]. Un motif est un ensemble de propriétés ou attributs tandis qu'une association est de la forme `A` $\longrightarrow$ `B` où `A` et `B` sont des motifs. La recherche de motifs fréquents et de règles d'association sont parmi les tâches les plus importantes en fouille de données. Les études en fouille de données se sont surtout intéressées jusqu'à présent à l'extraction de motifs fréquents — motifs dont la fréquence d'apparition parmi les individus d'une population donnée est supérieure à un seuil donné — et à la génération de règles d'association dérivant des motifs fréquents. L'algorithme le plus célèbre permettant d'extraire des motifs et des règles est Apriori, qui a été suivi par toute une famille d'algorithmes mis au point par la suite et possédant tous la caractéristique d'extraire l'ensemble des motifs fréquents ou un sous-ensemble de ces motifs (motifs fermés fréquents, motifs fréquents maximaux, générateurs minimaux [BAS 02]).

Dans cet article, nous posons le problème de la recherche de *motifs rares* ou *non fréquents*, qui se trouvent dans le complémentaire de l'ensemble des motifs fréquents. Les problèmes de la fouille de motifs rares et de la génération des règles d'association rares qui en dérivent n'ont pas encore été traités en détail dans la littérature (sachant que cet article est une version abrégée, revue et corrigée de [SZA 06] et qu'il existe aussi une étude théorique sur la complexité de la recherche des motifs fréquents et non fréquents dans [BOR 02]). Dans la suite, nous expliquons d'abord l'intérêt que peuvent revêtir les motifs et règles rares, puis nous donnons les définitions et les grandes lignes de la recherche de motifs et de règles rares. L'article se termine par une discussion sur les motifs et règles rares, accompagnée d'une comparaison avec la recherche de motifs fréquents et l'extraction de règles d'association, ainsi que d'une série de questions qui sont en cours d'investigation.

2. Motivations

La découverte de motifs rares peut se révéler très intéressante en médecine et en biologie. Considérons d'abord une base de données médicales et le problème de l'identification de la cause de maladies cardio-vasculaires (MCV). Une règle d'association fréquente (extraite d'un motif fréquent) comme "$\{\texttt{niveau élevé de cholestérol}\} \longrightarrow \{\texttt{MCV}\}$" permet de faire émerger l'hypothèse que les individus ayant un fort taux de cholestérol ont un risque élevé de MCV. À l'opposé, s'il existe un nombre conséquent de végétariens dans la base de données, alors une règle d'association rare comme "$\{\texttt{végétarien}\} \longrightarrow \{\texttt{MCV}\}$" permet de faire émerger l'hypothèse qu'un végétarien a un risque faible de contracter une MCV. Dans un tel cas, les motifs $\{\texttt{végétarien}\}$ et $\{\texttt{MCV}\}$ sont tous deux fréquents, mais le motif $\{\texttt{végétarien, MCV}\}$ est lui-même rare.

Le deuxième exemple, qui s'appuie sur les données réelles de la cohorte STANISLAS [MAU 05], montre l'intérêt de l'extraction des motifs rares pour la fouille de cohortes supposées saines. La cohorte STANISLAS est composée d'un millier de familles françaises présumées saines. L'objectif principal de l'étude de la cohorte est de mettre en évidence l'influence des facteurs génétiques et environnementaux sur la variabilité des risques cardio-vasculaires. Parmi les informations intéressantes à extraire de cette base de données figurent les profils associant des données génétiques à des valeurs extrêmes ou limites des paramètres biologiques. Cependant, ces associations sont plutôt rares dans les cohortes supposées saines. Dans ce contexte, l'extraction de motifs rares peut s'avérer très utile pour étudier les variations dans les profils — les profils rares pouvant conduire à des problèmes néanmoins — et ainsi avoir une idée plus complète des associations entre paramètres, ce que ne permet pas la seule recherche de motifs fréquents.

Le troisième exemple est en rapport avec la pharmacovigilance, qui est une branche de la pharmacologie dédiée à la détection et l'étude des effets indésirables des médicaments. L'extraction des motifs rares dans une base de données des effets indésirables de médicaments peut contribuer à un suivi plus efficace des effets indésirables graves et servir ensuite à prévenir les accidents mortels qui aboutissent au retrait de certains médicaments (comme par exemple le retrait de la cérivastatine, médicament hypolipémiant, en août 2001).

3. La recherche de motifs rares

Une méthode générique pour retrouver les motifs rares est présentée ci-après. Dans un premier temps, la méthode identifie un ensemble générateur minimal appelé *ensemble des motifs rares minimaux* ou MRMs. Dans un second temps, les MRMs sont utilisés pour retrouver tous les motifs rares. Avant d'arriver aux détail de la méthode, un rappel des définitions classiques est proposé.

- Une base de données formelle s'appuie sur le produit cartésien $\texttt{O} \times \texttt{A}$ associé à une relation R, où $\texttt{O} = \{\texttt{o}_1, \texttt{o}_2, \ldots, \texttt{o}_\texttt{m}\}$ est un ensemble d'objets, $\texttt{A} = \{\texttt{a}_1, \texttt{a}_2, \ldots, \texttt{a}_\texttt{n}\}$ est un ensemble d'attributs et $\texttt{R} \subseteq \texttt{O} \times \texttt{A}$ est une relation telle que $\texttt{R}(\texttt{o}_\texttt{i}, \texttt{a}_\texttt{j})$ signifie que l'objet $\texttt{o}_\texttt{i}$ possède l'attribut $\texttt{a}_\texttt{j}$.
- Un ensemble d'attributs forme un *motif* dont la taille est le nombre d'attributs qui le composent. Le *support* d'un motif P correspond au nombre d'objets contenant le motif et un motif est *fréquent* si son support est supérieur ou égal à un seuil de fréquence minimum donné (noté minsupp).
- La recherche de motifs fréquents consiste à engendrer tous les motifs dont le support est supérieur ou égal au seuil minsupp, en appliquant les principes suivants [AGR 96] :
  (i) "*la recherche des motifs fréquents commence par traiter les motifs de longueur minimale ; le support des motifs est calculé après un accès à la base données formelle ; les motifs fréquents sont conservés et les motifs non fréquents sont élagués*",
  (ii) "*tous les sous-motifs d'un motif fréquent sont fréquents*",
  (iii) "*tous les super-motifs d'un motif non fréquent sont non fréquents*".
  De plus, un motif P est *fermé* s'il n'existe aucun super-motif Q de P ($\texttt{P} \subseteq \texttt{Q}$) de même support.

Cela étant, un motif est dit *rare* ou *non fréquent* si son support est inférieur ou égal à un *support maximum*, noté maxsupp. Dans ce qui suit, la valeur de maxsupp se calcule à partir de celle de minsupp, à savoir maxsupp =

minsupp − 1 (ici minsupp et maxsupp sont donnés en valeur absolue). La recherche de motifs rares consiste à engendrer tous les motifs dont le support est inférieur ou égal au seuil *maxsupp*.

Il peut exister un intervalle de valeurs entre *minsupp* et *maxsupp*. Mais dans cet article, nous avons travaillé avec un cas particulier sans intervalle, c'est à dire que pour nous un motif est rare s'il n'est pas fréquent. Cela implique l'existence d'une seule frontière entre motifs rares et fréquents. Une telle frontière est étudiée et discutée dans [BOU 03, CAL 05].

L'ensemble des motifs rares et l'ensemble des motifs fréquents ont tous deux un sous-ensemble minimal générateur. Dans le cas des motifs fréquents, ce sous-ensemble est l'ensemble des *motifs fréquents maximaux* (MFMs). Un motif est un motif fréquent maximal s'il est fréquent et si tous ses super-motifs ne sont pas fréquents.

De façon complémentaire, un *motif rare minimal* (MRM) est un motif rare dont tous les sous-motifs ne sont pas rares. L'ensemble des motifs rares minimaux forme un ensemble générateur minimal à partir duquel tous les motifs rares peuvent être retrouvés, comme tous les motifs fréquents peuvent être retrouvés à partir des motifs fréquents maximaux. Pour les motifs fréquents maximaux, tous les sous-motifs possibles des MFMs sont considérés et leur support est calculé après un passage sur la base de données. De façon duale, tous les super-motifs des motifs rares minimaux sont considérés, puis le calcul du support de ces motifs se fait grâce grâce à un passage sur la base de données.

Parmi les motifs rares se distinguent les motifs rares de support 0 (zéro), appelés *motifs zéros*, et les motifs rares de support non nul, ou *motifs non zéros*. Le nombre de motifs rares zéros peut être très élevé. De façon analogue à un motif rare minimal, un motif est *générateur zéro minimal* (GZM) si c'est un motif zéro et si tous ses sous-motifs sont des motifs non zéros (tous ses super-motifs sont bien sûr des motifs zéros).

Les motifs rares minimaux peuvent être retrouvés simplement à l'aide de l'algorithme Apriori de la façon suivante : quand un motif non fréquent donc rare P est détecté — son support est inférieur ou égal à maxsupp (ou strictement inférieur à minsupp) — aucun des super-motifs de P n'est considéré par la suite, car ces super-motifs sont de manière sûre non fréquents. Puisque l'algorithme Apriori explore le treillis des motifs niveau par niveau — du "bas vers le haut" ou des tailles minimales aux tailles maximales —, il calcule nécessairement le support des motifs rares minimaux. Les motifs rares minimaux sont élagués et l'algorithme Apriori construit ensuite les motifs candidats de longueur k dont tous les sous-motifs de longueur $(k-1)$ sont fréquents. Si, pour un candidat P de longueur k, un des sous-motifs de P, soit Q, de longueur $(k-1)$, n'est pas fréquent, alors P est rare ; et en plus cela signifie que Q est un sous-motif rare minimal. Ainsi, l'espace de recherche dans le treillis des motifs est réduit de façon significative.

Une légère modification d'Apriori suffit pour conserver les motifs rares minimaux : dès que le support d'un motif candidat P est inférieur au support minimum, alors P est enregistré dans l'ensemble des motifs rares minimaux. Ensuite, tous les motifs rares sont retrouvés à partir des motifs rares minimaux. Pour cela, il faut engendrer tous les super-motifs possibles des motifs rares minimaux. Les générateurs zéros minimaux permettent de filtrer les motifs zéros pendant la génération des super-motifs rares.

4. Synthèse et questions

Dans cet article, une méthode pour extraire les motifs rares dans une base de données a été présentée. La méthode s'appuie sur l'algorithme de recherche de motifs fréquents Apriori et se compose de deux parties : (i) recherche d'un sous-ensemble générateur minimal des motifs rares (MRMs), (ii) recherche à partir des MRMs des motifs rares dont le support n'est pas nul. Ce travail de recherche est l'un des premiers à s'intéresser de façon systématique et spécifique aux motifs rares. L'algorithme Apriori a été le premier algorithme de recherche de motifs fréquents et il a été suivi de nombreux autres algorithmes plus efficaces et performants. De manière similaire, il ne fait aucun doute que la méthode de recherche de motifs rares présentée ici sera dans un avenir proche améliorée et les auteurs de l'article s'y emploient. Ainsi, des sous-ensembles utiles pour la recherche de motifs fréquents ont été découverts, comme les motifs fermés fréquents, les motifs fréquents maximaux, les générateurs (clés) minimaux, etc. De façon duale, de tels sous-ensembles doivent pouvoir être définis pour les motifs rares, puisque

par exemple le complémentaire des motifs fréquents maximaux est l'ensemble des motifs rares minimaux. Une autre question intéressante est la suivante : comme les motifs fermés fréquents déterminent sans ambiguïté tous les motifs fréquents et leur support, existe-t-il un sous-ensemble analogue qui déterminerait les motifs rares ? En outre, va suivre l'exploitation de la recherche des motifs rares pour la génération de règles d'association rares.

Pour terminer, il faut évoquer une série de questions plus théoriques qui se posent également :

– Des représentations condensées des motifs fréquents sont introduites dans [BOU 03, CAL 05], ainsi qu'une bordure négative et une bordure positive entre motifs fréquents et rares, et des ensembles disjonctifs libres et δ-libres (δ mesure le nombre de contre-exemples à une règle d'association). Quel est le dual des ces ensembles pour les motifs rares ?
– La base de Duquenne-Guigues permet d'engendrer les règles d'association exactes ou de confiance 1 [GUI 86] (ces règles sont aussi celles qui peuvent être extraites du treillis de concepts associé à la base de données formelle). Cette base peut être calculée à l'aide des motifs dits *pseudo-fermés* qui se définissent comme suit : un motif P est pseudo-fermé s'il n'est pas fermé et si tous les sous-motifs pseudo-fermés $Q \subset P$ qu'il contient strictement ont une fermeture contenue dans P (voir par exemple [GAN 99, STU 01]). Il est intéressant d'étudier le rapport exact existant entre les motifs pseudo-fermés et les motifs rares : est-ce que l'un se dérive de l'autre et est-ce que l'un permet de calculer l'autre.

5. Bibliographie

[AGR 96] AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A. I., Fast Discovery of Association Rules, FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSAMY R., Eds., *Advances in Knowledge Discovery and Data Mining*, Menlo Park, California, 1996, AAAI Press / MIT Press, p. 307–328.

[BAS 02] BASTIDE Y., TAOUIL R., PASQUIER N., STUMME G., LAKHAL L., Pascal : un algorithme d'extraction des motifs fréquents, *Technique et science informatiques*, vol. 21, n^o 1, 2002, p. 65–95.

[BOR 02] BOROS E., GURVICH V., KHACHIYAN L., MAKINO K., On the Complexity of Generating Maximal Frequent and Minimal Infrequent Sets, *Symposium on Theoretical Aspects of Computer Science*, 2002, p. 133-141.

[BOU 03] BOULICAUT J.-F., BYKOWSKI A., RIGOTTI C., Free-Sets : A Condensed Representation of Boolean Data for the Approximation of Frequency Queries, *Data Mining and Knowledge Discovery*, vol. 7, n^o 1, 2003, p. 5–22.

[CAL 05] CALDERS T., RIGOTTI C., BOULICAUT J.-F., A survey on condensed representations for frequent sets, BOULICAUT J.-F., RAEDT L. D., MANNILA H., Eds., *Constraint-based mining and Inductive Databases*, Lecture Notes in Computer Science 3848, Springer-Verlag, Berlin, 2005, p. 64–80.

[GAN 99] GANTER B., WILLE R., *Formal Concept Analysis*, Springer, Berlin, 1999.

[GUI 86] GUIGUES J., DUQUENNE V., Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Mathématiques et Sciences Humaines*, vol. 95, 1986, p. 5–18.

[HAN 01] HAN J., KAMBER M., *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.

[MAU 05] MAUMUS S., NAPOLI A., SZATHMARY L., VISVIKIS-SIEST S., Exploitation des données de la cohorte STANISLAS par des techniques de fouille de données numériques et symboliques utilisées seules ou en combinaison, *Atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances - EGC 2005, Paris, France*, 2005, p. 73–76.

[STU 01] STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N., LAKHAL L., Intelligent structuring and reducing of association rules with formal concept analysis, BAADER F., BREWKA G., EITER T., Eds., *KI 2001*, Lecture Notes in Artificial Intelligence 2174, Springer-Verlag, Berlin, 2001, p. 335–350.

[SZA 06] SZATHMARY L., MAUMUS S., PETRONIN P., TOUSSAINT Y., NAPOLI A., Vers l'extraction de motifs rares, RITSCHARD G., DJERABA C., Eds., *Extraction et gestion des connaissances (EGC'2006), Lille*, RNTI-E-6, Cépaduès-Éditions Toulouse, 2006, p. 499–510.

6. Annexe : un exemple de recherche de motifs fréquents et rares

Dans cet article, la base de données formelle suivante, reprise de [BAS 02], est utilisée (notée $\mathcal{D}$, table 1). Le seuil de fréquence (absolu), `minsupp`, est fixé à 3, et donc le seuil de non fréquence, `maxsupp`, à 2. Les motifs rares et fréquents issus de cette base de données formelle peuvent être visualisés sur la figure 1.

	A	B	C	D	E
1	x		x	x	
2		x	x		x
3	x	x	x		x
4		x			x
5	x	x	x		x

TAB. 1. *La base de données formelle $\mathcal{D}$ prise en exemple.*

Étant donné un seuil de fréquence, les motifs peuvent être classifiés en deux catégories : les motifs rares, dont la fréquence est en-dessous du seuil, et les motifs fréquents, dont la fréquence est au-dessus du seuil. Une frontière existe entre ces deux catégories, qui peut être visualisée sur le treillis des parties de l'ensemble des attributs considérés (voir figure 1). En bas du treillis se trouve le plus petit motif, ou motif de longueur nulle, qui correspond à l'ensemble vide. À chaque niveau se situent les motifs de même taille. Au sommet du treillis se trouve le motif le plus long qui contient tous les attributs. Le support de chaque motif par rapport à la base de données formelle $\mathcal{D}$ est indiqué dans le coin en haut à droite à côté de chaque motif.

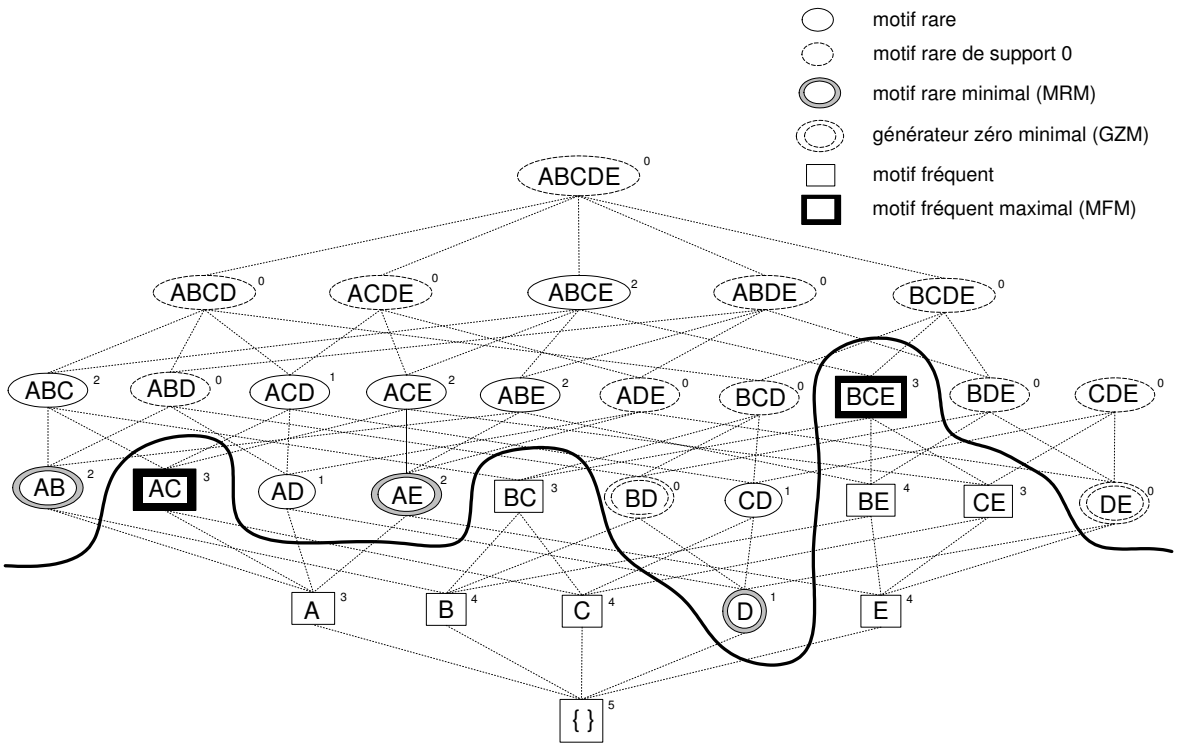

FIG. 1. *Le treillis des parties de la base de données formelle $\mathcal{D}$ avec les motifs fréquents et les motifs rares.*

L'ensemble des motifs rares forme un sup-demi-treillis car il est est fermé pour l'opération sup — tout sup de deux rares est rare, mais il ne forme pas un inf-demi-treillis, car l'inf de deux rares n'est pas forcément rare. De façon duale, les motifs fréquents forment un inf-demi-treillis mais pas un sup-demi-treillis.

Sur la figure 1, les deux générateurs zéros minimaux sont {BD} et {DE}. Les GZMs forment une représentation condensée et sans perte d'information des motifs zéros : à partir des GZMs, tous les motifs zéros peuvent être retrouvés — avec leur support, qui est toujours 0 — ; pour cela, il suffit d'engendrer tous les super-motifs possibles des GZMs en utilisant les attributs de la base de données. Mais, cette génération n'est pas effectuée à cause du trop grand nombre de motifs zéros mais aussi parce que seuls les GZMs sont utiles. La seconde partie de la méthode de recherche permet de retrouver tous les motifs rares non-zéros à partir des MRMsà l'aide d'une approche par niveau. Si un candidat intègre un sous-motif GZM, alors ce candidat est de manière sûre un motif zéro et peut donc être élagué. Les GZMs permettent ainsi de réduire l'espace de recherche lors de la recherche des motifs rares.