La confiance est dans l'air ! Application à l'identification des parcours hospitaliers

Yves Mercadier, Jessica Pinaire, Jérôme Azé, Sandra Bringay, Maguelonne Teisseire

La confiance est dans l'air ! Application à l'identification des parcours hospitaliers

Yves Mercadier*, Jessica Pinaire*,**,***
Jérôme Azé*,Sandra Bringay*,**** Maguelonne Teisseire*,♯

* LIRMM, UMR 5506, Université Montpellier, France
prenom.nom@lirmm.fr,
** CHU, Département d'information médicale, BESPIM, Nîmes, France
jessica.pinaire@chu-nimes.fr
*** équipe d'accueil 2415, Institut Universitaire de Recherche Clinique,
Université Montpellier, Montpellier, France
paul.landais@umontpellier.fr
**** AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr
♯ TETIS, IRSTEA, Montpellier, France
maguelonne.teisseire@irstea.fr

Résumé. L'extraction de motifs séquentiels permet d'identifier les séquences fréquentes d'événements ordonnés. Afin de résoudre le problème du grand nombre de motifs obtenus, nous proposons l'extension pour les motifs séquentiels de la confiance, mesure d'intérêt utilisée classiquement pour sélectionner les règles d'association. Dans cet article, après avoir présenté les données, nous définirons formellement la notion de confiance appliquée aux motifs séquentiels. Nous appliquerons cette mesure pour identifier des trajectoires hospitalières, représentées par les motifs séquentiels, dans des données issues du PMSI (Programme de Médicalisation des Systèmes d'Information). Nous nous sommes focalisés sur un cas d'étude hospitalière : l'infarctus du myocarde (IM), et notamment la prédiction de la trajectoire des patients ayant eu un IM entre 2009 et 2013. Les résultats obtenus ont été soumis à un spécialiste pour discussion et validation.

1 Introduction

Parmi les méthodes d'extraction de connaissances non supervisées, nous nous intéressons aux méthodes d'extraction de motifs. Il en existe un très grand nombre permettant d'identifier des régularités dans les jeux de données. L'ingénieur de la connaissance utilise alors son expérience pour proposer un type ou une combinaison de ces motifs selon les besoins des experts métier souhaitant exploiter le jeu de données. Ces méthodes génèrent très souvent un grand nombre de motifs. Un expert métier peut alors être submergé d'informations et ne pas tirer partie des résultats du processus de fouille de données. Pour limiter le nombre de motifs présentés et permettre leur interprétation, l'ingénieur de la connaissance applique alors un filtrage

des motifs en appliquant des mesures d'intérêt puis regroupe les motifs en appliquant des mesures de similarité. Les mesures d'intérêt sont des indicateurs statistiques dont la sémantique est parfois difficile à interpréter par l'expert.

De nombreuses études ont été réalisées en vue de comparer les mesures d'intérêt pour aller vers une amélioration du résultat final de la fouille de données par exemple Lenca et al. (2003), Blanchard (2005b). Dans le cadre de cette étude, nous nous sommes intéressés à une mesure en particulier, la confiance. Cette mesure d'intérêt a été introduite par Agrawal et al. (1993) pour les règles d'associations. Elle consiste à estimer la probabilité dans la base de données d'obtenir une association à partir de ses constituants. L'originalité de notre mesure que nous avons appelée r-confiance est double. Premièrement, elle fonctionne pour tous les types de motifs (règle d'association, motif séquentiel, motif spatio-temporel). Deuxièmement, elle utilise comme opérateur d'agrégation « la proportion de position ». Nous avons également développé une interface qui permet de représenter les motifs extraits mais également de les comparer en prenant en compte différentes mesures d'intérêt, dont la r-confiance.

Nous avons appliqué cette mesure dans un contexte spécifique pour identifier des trajectoires hospitalières, représentées par des motifs séquentiels, dans des données issues du PMSI (Programme de Médicalisation des Systèmes d'Information). Nous nous sommes focalisés sur un cas d'étude hospitalière : l'infarctus du myocarde (IM) et notamment la prédiction de la trajectoire des patients ayant eu un IM entre 2009 et 2013. Nous cherchons à identifier des trajectoires de GHM (Groupe Homogène de Malade), ce dernier est un code renseignant les caractéristiques d'une hospitalisation. Les résultats obtenus ont été soumis à un spécialiste pour discussion et validation.

Dans cet article, nous allons présenter la r-confiance dans la section 2 ainsi que l'interface de l'outil développé dans la section 3. Nous présenterons notre cas d'étude et les données utilisées dans la section 4. Nous analyserons les résultats obtenus et démontrerons l'efficacité de cette méthode, que nous pourrons utiliser à plus grande échelle, par exemple, dans l'identification de trajectoires fréquentes de patients pour un contexte donné.

2 Sélection des motifs d'intérêt selon la r-confiance

2.1 Vers une nouvelle confiance

Depuis les années 90, de nombreuses méthodes ont été proposées pour l'extraction de motifs fréquents dans les bases de données Rabatel (2011). Ces motifs se sont complexifiés avec le temps pour prendre en compte différentes dimensions (temporelles, spatiales...)(Heas, 2005). On peut citer les règles d'association (Agrawal et al., 1996), les motifs séquentiels (Pei et al., 2001), les co-localisations (Sundaram et al., 2012), les trajectoires (Etienne et Devogele, 2012), les graphes (Pennerath et Napoli, 2006). Par ailleurs, afin de résoudre le problème du grand nombre de règles d'association de nombreuses mesures d'intérêt ont été proposées pour leur sélection (Tan et al., 2002). D'après Grissa (2013) on peut dénombrer plus de soixante mesures d'intérêt dédiées aux règles d'association. À notre connaissance il n'existe que très peu de mesure d'intérêt pour les motifs séquentiels, les co-localisations, les graphes ou encore les trajectoires. Nous avons voulu combler ce vide en proposant une mesure d'intérêt spécifique applicable à tout type de motif. Nous avons choisi de faire l'extension de la confiance définie pour les règles d'association car cette mesure permet d'estimer la liaison entre les deux itemsets

constituant l'association. Nous pensons que la qualité des liaisons inter-itemset dans un motif est primordiale pour estimer et discriminer les motifs dans un ensemble de motif.

2.2 Motifs séquentiels et candidats séquentiels

En 1995, la problématique de la recherche de règles d'association a été étendue pour détecter des comportements typiques dans le temps et le concept de motifs séquentiels a été introduit (Agrawal et Srikant, 1995) puis les règles séquentielles ont été proposées par Das et al. (1998). Ces motifs ont été appliqués dans de nombreux domaines comme le panier de la ménagère précédemment introduit (Agrawal et Srikant, 1995), la fouille de données d'usage du Web (Dong Haoyuan et al., 2009), la fouille de texte (Charnois et al., 2009).

Nous présentons quelques définitions préliminaires des motifs séquentiels avant de proposer notre mesure de filtrage.

Définition 1. *Une base de données séquentielles contient un ensemble ordonné d'éléments, généralement par le temps. Soit* $I = \{i_1, i_2, ..., i_k\}$ *l'ensemble de tous les items. Un sous-ensemble de* I *est appelé un itemset. Une séquence* $S =< I_1, I_2, ..., I_n >$ *est une liste ordonnée d'itemset (*$I_i \subseteq I$*). Chaque itemset d'une séquence représente un ensemble d'événements qui apparaissent à la même estampille temporelle. Les différents itemsets d'une séquence sont associés à des estampilles temporelles différentes.*

Définition 2. *Une séquence* $S_1 =< I_1, I_2, ..., I_m >$ *est une sous-séquence de* $S_2 =< I'_1, I'_2, ..., I'_n >$ *(noté* $S_1 \preceq S_2$*) si et seulement si* $\exists i_1, i_2, ..., i_m$ *tels que* $1 < i_1 < i_2 < ... < i_m \leq i_n$ *et* $I_1 \subseteq I'_1, I_2 \subseteq I'_2, ..., I_m \subseteq I'n$.

Exemple 1. Un patient se rend dans un hôpital pour une pathologie : son séjour est codé par un item ex. code -> 01. Il revient plus tard pour un deuxième examen ex.code -> 02. Cela constitue une séquence dans la base que l'on peut noter : $S_0 =< (01), (02) >$ la séquence $S'_0 =< (01) >$ est une sous-séquence de S_0

Définition 3. *Etant donné un ensemble de séquences* $D = \{S_1, S_2, ..., S_n\}$*, le support d'une séquence* S_1 *correspond au nombre de séquences de* D *qui contiennent* S_1*. Si le support d'une séquence* S_1 *satisfait un seuil de support minimum* $minsup$*, alors* S_1 *est une séquence fréquente ou motif séquentiel. L'objectif de la recherche de motifs séquentiels est donc d'extraire l'ensemble complet des séquences fréquentes par rapport à un seuil de support minimum minsup.*

Exemple 2. Un deuxième patient se rend dans un hôpital pour une série d'examen : sont séjour est codé par un item ex. code -> 03. Il revient plus tard pour une deuxième série d'examen ex.code -> 04. Un troisième patient fait de même. $S_1 =< (03), (04) > S_2 =< (03), (04) >$ Nous fixons arbitrairement le seuil de support à deux. La séquence $< (03), (04) >$ sera alors considérée fréquente.

Définition 4. *Etant donné un motif séquentiel* $M =< M_1, M_2, ..., M_n >$*, un* **candidat séquentiel** *de* M*,* $C_p =< C_1, C_2, ..., C_p >$*, est défini comme une des sous-séquences préfixes de* p *items telle que* $p < m$ *et* $\forall i_1, i_2, ..., i_p$*,* $1 \leq i_1 < i_2 < ... < i_p$ *et* $C_1 = M_{i_1}, ..., C_{p-1} = M_{m-p-1}, C_p = M_{m-p}$*. Un motif séquentiel de longueur* n *peut être associé à* $n-1$ *candidats séquentiels.*

Soit $M =< (ab)(c)(d)(e) >$ un motif. Ce motif génère 3 candidats séquentiels : $< (ab) >$, $< (ab)(c) >$ et $< (ab)(c)(d) >$. $< (ab)(c) >$ est un candidat séquentiel mais pas $< (ab)(d) >$ car ses itemsets ne sont pas consécutifs dans M.

2.3 La r-confiance

Nous proposons de considérer la confiance que l'on peut avoir dans un motif comme étant la mesure de la représentativité de ce motif par rapport aux données. Plus la confiance d'un motif est élevée, plus l'apparition des premiers items le constituant permet l'obtention du motif complet. Concrètement, la confiance exprime alors la qualité de la liaison des itemsets internes au candidat à partir du premier itemset le constituant. Avant de définir la r-confiance d'un motif de façon générale, nous définissons la r-confiance élémentaire d'un motif séquentiel.

Soit M un motif de longueur n et C un candidat séquentiel de ce motif de longueur p selon la définition 4. La r-confiance élémentaire, notée *r-conf-e*, est définie par :

$$\textit{r-conf-e}(M, C) = \frac{support_B(M)}{support_B(C)} \tag{1}$$

La r-confiance calculée pour le motif M correspond à l'agrégation des n-1 r-confiances élémentaires des candidats séquentiels le composant, c'est-à-dire la proportion de candidats séquentiel ayant une r-confiance élémentaire. Afin de conserver la notion de mesure d'intérêt et donc de filtrage des motifs extraits, seules les r-confiances élémentaires dont la valeur est supérieure à un seuil fixé minR seront prises en compte dans cette agrégation. Pour M un motif de longueur n, soit $\mathbf{C}$, l'ensemble des n-1 candidats séquentiels de M.

$$\textit{r-conf}(M) = \begin{cases} 0 & \text{si } Card(\{C \in \mathbf{C}, \textit{r-conf-e}(M, C) > minR\}) = 0 \\ \frac{Card(\{C \in \mathbf{C}, \textit{r-conf-e}(M,C) > minR\}) + 1}{n} & \text{sinon} \end{cases} \tag{2}$$

Cette mesure, ici définie pour les motifs séquentiels, est certainement généralisable à d'autres types de motif. Nous considérons ici que chaque motif est construit structurellement à partir de l'ensemble des candidats de celui-ci. En effet la r-confiance mesure une caractéristique de la structure du motif, elle évalue la proportion de candidat ayant une r-confiance élémentaire supérieure à un seuil. Les motifs de tous types pouvant être considérés comme des conteneurs structurés, il semble possible d'en évaluer la qualité par cette mesure.

2.4 Exemple de calcul

Considérons la base de données du tableau 1 constituée de sept motifs séquentiels.
Nous étudions le motif séquentiel $\mathbf{M} = \langle (a, b)(c)(b, c) \rangle$. Le calcul des supports donne :
$support(\langle (\mathbf{a}, \mathbf{b}) \rangle) = \frac{3}{7} = 0.43$
$support(\langle (\mathbf{a}, \mathbf{b})(\mathbf{c}) \rangle) = \frac{2}{7} = 0.29$

Nous en déduisons d'abord le calcul des r-confiances élémentaires :
r-conf-e$(\mathbf{M}, \langle (a, b) \rangle) = \frac{\frac{2}{7}}{\frac{3}{7}} = 0.66$
r-conf-e$(\mathbf{M}, \langle (a, b)(c) \rangle) = \frac{\frac{2}{7}}{\frac{2}{7}} = 1$

Séquences
$\langle(a,b)(c)(b,c)\rangle$
$\langle(c)(c)(b)\rangle$
$\langle(d)(e)(f)\rangle$
$\langle(d)(f)\rangle$
$\langle(a,b)(d,e)\rangle$
$\langle(d)(a,b)(c)(b)(b,c)\rangle$
$\langle(c)(b)(d)(b,c)\rangle$

TAB. 1: Exemple d'une base de motifs séquentiels.

Puis, à partir d'un seuil fixé à $minR = 0.7$, nous pouvons calculer la r-confiance du motif : *r-conf*$(\mathbf{M}) = \frac{1+1}{3} = 0.66$

3 Présentation de l'interface utilisateur

La phase d'extraction des motifs conduit souvent à générer un trop gros volume de motifs à valider. L'expert est alors submergé et démuni devant les nouvelles données ainsi produites. Notre proposition d'une nouvelle mesure de filtrage s'inscrit dans ce contexte et a été implémentée dans une interface interactive avec pour objectif de faciliter l'analyse des résultats de la fouille de données séquentielles. Ce problème a déjà été approché pour le traitement de fouille de règle d'association Blanchard (2005a) ou Hervouet (2011), ou pour les événements temporels Barazzutti et al. (2015).
Notre interface est dotée de trois fonctions principales : la navigation entre les ensembles de motifs, la visualisation de statistiques sur des ensembles de motifs et la manipulation des ensembles de motifs notamment pour leur comparaison que nous allons détailler.

Tout d'abord, nous donnons la possibilité à l'expert de créer des ensembles de motifs à partir de contraintes sur les mesures d'intérêt, ou à partir de plusieurs jeux de résultats d'extraction (avec des supports minimums différents par exemples). Pour cela, nous avons doté l'interface d'une algèbre des ensembles qui nous permet de comparer des ensembles de motifs. L'interface permet les opérations suivantes : union, intersection, soustraction, soustraction symétrique. Il est aussi possible de procéder à la caractérisation des ensembles numériques par les indicateurs statistiques suivants : cardinal, minimum, maximum, moyenne, médiane, mode, écart-type. La navigation, la manipulation et la comparaison des ensembles de motifs par l'expert sont alors facilitées.

La navigation entre les ensembles issus de la manipulation des données est permise par des aller-retours entre les différentes représentations. Cela ne correspond pas au terme d'hyperdata (Kopecky et Pedrinaci, 2011) mais serait plus proche du concept d'hyperset au sens d'une navigation inter-ensembles.

Nous présentons dans la suite un canevas des possibilités d'utilisation de l'outil. Dans un premier temps, en post-traitement de la fouille, nous procédons au calcul de la r-confiance pour une série de seuils élémentaires compris entre zéro et un, avec un pas de 0.1. Nous obtenons 10 ensembles de motifs que nous comparons via l'interface de l'outil.

La confiance est dans l'air !

FIG. 1: Représentation d'un ensemble de motifs séquentiels sous forme de tableau

rang	id	Sequences	Support	Taille	rConfiance
0	1	<(02C051)(28Z04Z)(28Z04Z)(28Z04Z)(28Z04Z)(28Z04Z)>	10	6	0.5
1	2	<(02C051)(28Z04Z)(28Z04Z)(28Z04Z)(28Z04Z)>	11	5	0.4
2	3	<(02C051)(28Z04Z)(28Z04Z)(28Z04Z)>	11	4	0
3	4	<(02C051)(28Z04Z)(28Z04Z)>	14	3	0.666667
4	5	<(02C051)(28Z04Z)>	14	2	0
5	6	<(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(28Z04Z)(28Z04Z)(28Z04Z)(28Z04Z)(28Z04Z)>	11	12	0.833333
6	7	<(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(28Z04Z)(28Z04Z)(28Z04Z)(28Z04Z)>	14	11	0.818182
7	8	<(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(28Z04Z)(28Z04Z)(28Z04Z)>	15	10	0.9
8	9	<(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(28Z04Z)(28Z04Z)>	40	9	0.888889
9	10	<(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(02C05J)(28Z04Z)>	46	8	0.875

3.1 Comment créer des ensembles de motifs ?

Nous considérons ici les motifs associés aux valeurs de leurs mesures d'intérêt pour un seuil de confiance élémentaire de 0.9. Nous obtenons la visualisation présentée sur la figure 1. Afin de générer des ensembles de motifs, nous appliquons diverses contraintes. Pour cela, nous disposons d'une console minimale avec un jeu de commandes. On peut pour un premier exemple rechercher les motifs contenant l'item 05M13T (correspondant à une hospitalisation pour douleur thoracique). Soit M l'ensemble des motifs :

$A = \{M_i | 05M13T \preceq M_i\}$

Nous décrivons cette commande ainsi. L'ensemble A est constitué des motifs du premier fichier chargé tel qu'ils incluent l'item 05M13T.

$B = \{M_i \in A | \textit{r-conf}(M_i) > 0.5\}$

Nous obtenons ici un ensemble B constitué des motifs issus de A et respectant la contrainte de support.

Nous pouvons maintenant appliquer des opérations sur ces deux ensembles. Par exemple si nous désirons obtenir le complémentaire de B dans A. Nous procédons de la façon suivante.

$C = \{A \triangle B\}$

L'ensemble C sera le résultat de la soustraction symétrique de l'ensemble A avec l'ensemble B, soit le complémentaire de B dans A.

3.2 Comment visualiser graphiquement les ensembles de motifs ?

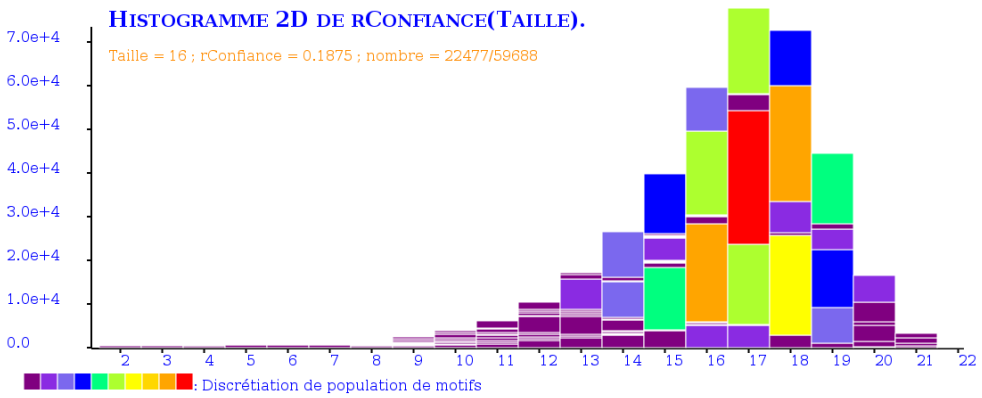
Nous proposons ici deux représentations possibles des ensembles de motifs accessibles via notre interface figure 2 et figure 3.

La figure 2 est l'histogramme de l'ensemble des motifs séquentiels issus de la fouille de données. Chaque barre représente un ensemble de motifs pour une valeur de support. Chaque bloc d'une barre représente les motifs ayant une même valeur de r-confiance. On peut faire

apparaître, par survol de la souris d'un bloc de l'histogramme les informations correspondantes à savoir : la valeur du support, la valeur de la r-confiance, le nombre de motifs constituant le bloc, le nombre de motifs constituant la barre. Afin d'améliorer l'ergonomie visuelle de l'histogramme empilé nous utilisons un code couleur de type arc-en- ciel. Le bloc le plus important en terme de population de motifs sera codé par la couleur rouge ; les blocs les moins importants seront codés en violet.

La figure 3 représente un arbre correspondant à l'agrégation d'un ensemble de motifs séquentiels. Nous procédons comme suit pour réaliser cette agrégation : nous parcourons l'ensemble des motifs séquentiels étudiés, nous extrayons les motifs débutants par un GHM choisi comme racine de l'arbre, nous parcourons les motifs ainsi extraits item par item, nous créons un nœud pour chaque item, nous créons un arc entre deux items successifs, nous itérons cette procédure sur l'ensemble des motifs extraits. Pour obtenir l'ensemble de la figure 3 nous recherchons les motifs contenant le GHM 05K051. Nous obtenons avec ce GHM comme racine un ensemble de neuf motifs. Nous agrégeons ces neuf motifs sur l'arbre de la figure 3. Le dessin de cet arbre a été obtenu avec l'aide de la librairie D3.js (Data-Driven Documents).

FIG. 2: Représentation d'un ensemble de motifs séquentiels sous forme d'un histogramme empilé



3.3 Comment caractériser et comparer statistiquement les ensembles de motifs ?

Les combinaisons de mesures d'intérêt ne sont pas suffisantes pour comparer les grands ensembles de motifs. L'interface développée permet de comparer les ensembles de motifs entre eux à partir d'indicateurs statistiques. Nous procédons ainsi : recherche pour le motif 05K051, codant l'IM. $A = \{M_i | 05K051 \preceq M_i\}, B = \{M_i \in A | \textit{r-conf}(M_i) > 0.2\}$, puis affichage des caractéristiques de A dans la figure 4.

La confiance est dans l'air !

FIG. 3: Représentation d'un ensemble de motifs séquentiels sous forme d'un arbre

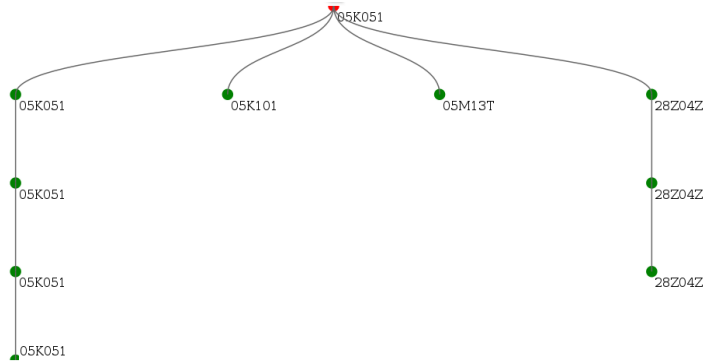


FIG. 4: Caractérisation statistique d'un ensemble de motifs séquentiels

Tableau du Résumé

Estimateur	Support	Taille	rConfiance
Cardinal	9	9	9
Maximun	12	6	0.666667
Minimum	10	3	0.4
Moyenne	10.44	4.56	0.5
Médiane	10	5	0.5
Mode	10	5	0.5
Ecart type	0.68	0.83	0.09

4 Application à l'identification de parcours hospitaliers

4.1 Données

La collecte des données hospitalières dans le cadre du PMSI génère sur le plan national des bases de données de l'ordre de 25 millions d'enregistrements (séjours) par an [1]. Ces données, recueillies à des fins médico-économiques, peuvent *a posteriori* servir à des fins d'analyse et de recherche, pour examiner des questions médicales et épidémiologiques (Quantin et al., 2014; Bocquier et al., 2011). Dans ces bases, grâce au numéro anonyme de patient, il est possible de reconstituer le parcours hospitalier d'un patient.

Nous nous sommes intéressés aux patients atteints d'un IM, au cours de la période 2009-2013. Pour cela, nous requêtons les bases PMSI nationales 2009-2013, en retenant tous les

1. Guide méthodologique de production des résumés de séjour du PMSI en médecine, chirurgie et obstétrique (fascicule spécial 2004/2 bis du Bulletin officiel)
http://www.atih.sante.fr/textes-officiels-du-pmsi-en-mco

patients ayant un séjour, avec un code acte [2] de cardiologie interventionnelle au cours des cinq années d'observation, soit 490 558 patients (75,7% d'hommes et 24,2% de femmes) [3].

Le taux hospitalier de décès sur la période étudiée est peu élevé (6%). Ce qui peut signifier que cette pathologie est bien prise en charge, ou que le patient décède avant son arrivée à l'hôpital. Pour en savoir davantage sur l'évolution de cette pathologie, nous avons récupéré tous les parcours de soins des patients concernés. À terme, nous souhaitons construire un modèle prédictif en sélectionnant les parcours les plus représentatifs en fréquence, en taille et en confiance. Nous souhaitons à ce titre évaluer la faisabilité d'un parcours lorsqu'un patient a entamé ce parcours. Cette modélisation nous permettra de simuler également à court terme ainsi qu'à long terme, le devenir des patients et les issues possibles de cette pathologie.

Pour notre étude, nous définissons la notion de parcours à l'aide du GHM [4], code renseignant les caractéristiques du séjour hospitalier. À chaque patient, est associé une série de GHM, de longueur égale au nombre de ses séjours effectués sur cinq ans : la trajectoire du patient.

Pour l'extraction de motifs séquentiels, nous avons procédé de la manière suivante : dans un premier temps nous avons écarté les patients ayant moins de 2 séjours, soit 34,4% de la population, afin d'observer une évolution de cette pathologie cardiovasculaire ou autres pathologies éventuelles associées. Ensuite, nous avons créé des contextes à l'aide de variables supplémentaires, qui sont l'âge, le sexe et le nombre d'hospitalisations sur cinq ans. Après concertation avec l'expert clinicien, nous avons discrétisé l'âge en trois classes : les moins de 45 ans, entre 45 et 65 ans et les plus de 65 ans. De même, nous avons décomposé le nombre d'hospitalisation en deux classes : ceux qui viennent entre deux et soixante fois, et ceux qui viennent plus de soixante fois. Toutes combinaisons faites, nous obtenons douze contextes-minimaux. Enfin, nous avons appliqué l'algorithme de recherche de motifs fréquents (Rabatel et al., 2010) sur nos données avec un seuil de 3,5%. Nous obtenons 554 955 motifs fréquents que nous devons trier afin d'en extraire les trajectoires d'intérêts à partir desquelles nous construirons un modèle prédictif par contexte.

Dans la suite, après avoir présenté la méthode, nous allons la mettre à l'épreuve, en tentant de répondre aux questions suivantes :

— **Q1** : Y a t'il une hospitalisation pour IM consécutivement à une hospitalisation pour douleur thoracique ?
— **Q2** : Qu'est-ce qui est fréquemment associé à l'insuffisance cardiaque ?

4.2 Exploration des motifs pour réaliser des découvertes médicales

Dans cette section, nous allons utiliser l'outil décrit précédemment pour répondre à la question **Q1**. Nous procédons de la façon suivante : nous recherchons parmi les motifs fréquents, les ensembles de motifs qui contiennent les hospitalisations pour douleur thoracique (code GHM 05M13T) et IM (05K051), sans se soucier de l'ordre dans un premier temps, et nous calculons leur r-confiance pour chacun d'eux pour un seuil élémentaire de 0.9. Nous obtenons deux motifs avec les résultats résumés dans le tableau 2 ci-dessous :

Nous constatons que pour les deux motifs la r-confiance est nulle. Ainsi, lorsque nous recherchons s'il est fréquent d'être hospitalisé pour IM juste après une hospitalisation pour

2. Article L. 6113-8 du code de la santé publique
3. Base nationale : 44,4% d'hommes et 55,6% de femmes
4. Groupe Homogène de Malades

La confiance est dans l'air !

TAB. 2: Résultat de la recherche des sous-ensembles de motifs contenant 05M13T et 05K051

séquences	support	taille	r-conf
<(05K051)(05M13T)>	215	2	0
<(05M13T)(05K051)>	286	2	0

douleur thoracique (deuxième motif séquentiel du tableau), nous constatons que ce n'est pas le cas. Il est plus probable d'être réhospitalisé pour un (ou plusieurs) autre(s) motif(s) avant de revenir pour IM. Ce qui est cohérent avec les connaissances médicales. En effet, lorsqu'un patient vient pour une douleur thoracique, si cette douleur n'est pas étiquetée IM c'est une alerte. En revanche, le patient peut présenter d'autres complications liées à l'athérosclérose (artériopathie des membres inférieurs, accident vasculaire, *etc.*) , à l'hypercholestérolémie, au diabète, à l'hypertension, à une insuffisance cardiaque gauche, *etc.* avant de revenir cette fois pour un IM.

Pour répondre à la question **Q2**, cette fois-ci, nous recherchons dans l'ensemble des motifs fréquents les motifs qui contiennent le GHM 05M06, codant l'insuffisance cardiaque. Nous extrayons 44 motifs avec des combinaisons de séquences de séances d'hémodialyse (28Z04Z) de longueur variable. Un parcours retient notre attention : <(28Z04Z)(05M043)(28Z04Z)> avec une r-confiance de 0.67. Ce qui signifie que si le patient parcourt 1/3 du motif dans 90% des cas il le terminera. En d'autres termes, lorsque le patient est hospitalisé pour une hémodialyse suivi d'une insuffisance cardiaque, il a 90% de chance d'être à nouveau hospitalisé pour une hémodialyse. Il est important de noter que ce motif a été repéré car sa r-confiance est élevée.

En effet, le rôle de l'hémodialyse est d'enlever au patient l'eau que son organisme n'arrive pas à évacuer naturellement. On donne au patient sujet à ces complications un traitement permettant de palier à cette difficulté et un régime alimentaire à suivre. Cependant, si l'observance du patient au regard de son traitement n'est pas bonne, il peut reprendre du poids, c'est essentiellement de l'eau que le coeur ne peut pas traiter et déclencher une insuffisance cardiaque.

5 Conclusion

Dans cette étude, nous avons proposé une nouvelle mesure d'intérêt qui est une extension aux motifs séquentiels de la confiance définie pour les règles d'association. Nous avons également développé une interface permettant d'explorer les motifs en prenant en compte différentes mesures d'intérêt dont la r-confiance que nous avons proposée. Nous avons utilisé cet outil pour faire émerger des connaissances médicales à partir d'une base issue des données du PMSI traitant de l'infarctus du myocarde. Les connaissances obtenues ont été validées et décryptées par un expert médical.

L'environnement développé et son utilisation par des utilisateurs experts ont permis de soulever plusieurs limites. La première concerne les contraintes de navigation entre les types de représentation. Par exemple, il n'est pas possible de naviguer directement entre les différentes représentations d'un ensemble de motif. Une deuxième limite concerne l'affichage des ensembles de motifs sous la forme d'un arbre. Pour l'instant, l'information sur les valeurs

d'une mesure d'intérêt n'est pas observable sur les arcs. Pour finir, nous prévoyons de consulter un expert en cardiologie pour évaluer l'impact de la mesure proposée dans la validation des connaissances extraites et sur l'accompagnement dans la découverte de nouvelles connaissances, réel objectif d'une telle plateforme.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, pp. 207–216.

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo (1996). Advances in knowledge discovery and data mining. Chapter Fast Discovery of Association Rules, pp. 307–328. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, Washington, DC, USA, pp. 3–14.

Barazzutti, P.-L., A. Cordier, et B. Fuchs (2015). Transmute: an Interactive Tool for Assisting Knowledge Discovery in Interaction Traces. Research report, Universite Claude Bernard Lyon 1 ; Universite Jean Moulin Lyon 3.

Blanchard, J. (2005a). *An interactive visualization system for mining, assessing, and exploring association rules.* Theses, Université de Nantes.

Blanchard, J. (2005b). Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.

Bocquier, A., N. Thomas, J. Zitouni, E. Lewandowski, S. Cortaredona, M. Jardin, O. Favier, S. Finkel, F. Champion, A. Bernardy, A. Trugeon, et P. Verger (2011). Evaluation of hospital stays linkage quality to study health spatial variation. a feasibility study in three french regions. *Revue d'épidémiologie et de santé publique 59*(4), 243–249.

Charnois, T., M. Plantevit, C. Rigotti, et B. Crémilleux (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Traitement Automatique des Langues(TAL) 50*.

Das, G., K. Lin, H. Mannila, G. Renganathan, et P. Smyth (1998). Rule discovery from time series. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, New York, USA, August 27-31, 1998*, pp. 16–22.

Dong Haoyuan, L., A. Laurent, et P. Poncelet (2009). Extraction de comportements inattendus dans le cadre du web usage mining. *Revue Nouvelles Technologies de l'Information (RNTI) 2ème Numéro Spécial : Fouille de Données Complexes (2009)*, 113–132.

Etienne, L. et T. Devogele (2012). Mesures de similarité de trajectoires basées sur l'utilisation de patrons spatio-temporels. *Ingénierie des Systèmes d'Information 17*(1), 11–34.

Grissa, D. (2013). *Behavioral study of interestingness measures of knowledge extraction*. Theses, Université Blaise Pascal - Clermont-Ferrand II ; Université de Tunis-El Manar (Tunisie).

Heas, P. (2005). *Apprentissage bayésien de structures spatio-temporelles (application à la fouille visuelle de séries temporelles d'images de satellites)*. Theses, Ecole nationale supé-

rieure de l'aéronautique et de l'espace, Toulouse.

Hervouet, D. (2011). Visualisation des règles d'association en environnement virtuel 3D interactif. Master's thesis.

Kopecky, J. et C. Pedrinaci (2011). RESTful write-oriented API for hyperdata in custom RDF knowledge bases. pp. 199 – 204. IEEE.

Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2003). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (RNTI)*, 220–246.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2001). Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. *2014 IEEE 30th International Conference on Data Engineering 0*, 0215.

Pennerath, F. et A. Napoli (2006). La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique. In C. D. Gilbert Ritschard (Ed.), *6èmes Journées Francophones "Extraction et gestion des connaissances" - EGC 2006*, Volume 2/RNTI-E-6, Lille, France, pp. 517–528. Cépaduès-éditions.

Quantin, C., Éric Benzenine, M. Hägi, B. Auverlot, M. Abrahamowicz, J. Cottenet, Évelyne Fournier, C. Binquet, D. Compain, Élisabeth Monnet, A.-M. Bouvier, et A. Danzon (2014). Évaluation du pmsi comme moyen d'identification des cas incidents de cancer colorectal. *Santé Publique 26*(1), 55–63.

Rabatel, J. (2011). *Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles)*. Theses, Université MontpellierII Sciences et Techniques du Languedoc.

Rabatel, J., S. Bringay, et P. Poncelet (2010). Aide à la décision pour la maintenance ferroviaire préventive. In *Extraction et Gestion des Connaissances*, EGC'10, Revue des Nouvelles Technologies de l'Information, pp. 363–368. Cépaduès-Éditions.

Sundaram, V. M., A. thnagavelu, et P. Paneer (2012). Discovering co-location patterns from spatial domain using a delaunay approach. *Procedia Engineering 38*, 2832 – 2845. ICMOC12.

Tan, P.-N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA, pp. 32–41. ACM.

Summary

Sequential patterns mining consist in identifying frequent sequences of ordered events. To solve the problem of the large number of patterns obtained, we extend the interest measure called confidence, conventionally used to select association rules to sequential patterns. In this paper, after presenting the data, we formally define the notion of confidence applied to the sequential patterns. We will apply this measure to identify hospital trajectories, represented by the sequential patterns in data from the PMSI (Medicalization Programme of Information Systems). We focused on a case study : myocardial infarction (MI), in order to predict the

trajectory of patients with MI between 2009 and 2013. The results were submitted to an expert for discussion and validation.