Extraction de règles d'association séquentielle à l'aide de modèles semi-paramétriques à risques proportionnels

Nicolas S. Müller*, Matthias Studer*, Gilbert Ritschard*, Alexis Gabadinho*

*Institut d'études démographiques et des parcours de vie, Université de Genève
{nicolas.muller, matthias.studer, gilbert.ritschard, alexis.gabadinho}@unige.ch

Résumé. La recherche de liens entre objets fréquents a été popularisée par les méthodes d'extraction de règles d'association. Dans le cas de séquences d'événements, les méthodes de fouille permettent d'extraire des sous-séquences qui peuvent ensuite être exprimées sous la forme de règles d'association séquentielle entre événements. Cette utilisation de la fouille de séquences pour la recherche de liens entre des événements pose deux problèmes. Premièrement, le critère principal utilisé pour sélectionner les sous-séquences d'événements est la fréquence, or les occurrences de certains événements peuvent être fortement liées entre elles même lorsqu'elles sont peu fréquentes. Deuxièmement, les mesures actuelles utilisées pour caractériser les règles d'association ne tiennent pas compte du caractère temporel des données, comme l'importance du *timing* des événements ou le problème des données censurées. Dans cet article, nous proposons une méthode pour rechercher des liens significatifs entre des événements à l'aide de modèles de durée. Les règles d'association sont construites à partir des motifs séquentiels observés dans un ensemble de séquences. L'influence sur le risque que l'événement « conclusion » se produise après le ou les événements « prémisse » est estimée à l'aide d'un modèle semi-paramétrique à risques proportionnels. Outre la présentation de la méthode, l'article propose une comparaison avec d'autres mesures d'association [1].

1 Introduction

La recherche de motifs fréquents et de règles d'association entre des objets a fait l'objet de nombreux travaux en data mining. Son extension à la recherche de motifs séquentiels fréquents a été également un domaine en plein essor ces dernières années. En revanche, la caractérisation des règles d'association séquentielle restent un problème moins exploré. Il existe néanmoins des critères, comme la confiance ou le rappel, qui ont été reformulés dans le cadre de la fouille de règles d'association à l'intérieur d'une séquence unique (Mannila et al., 1997). Toujours dans le cadre d'une séquence unique, Blanchard et al. (2008) ont développé un indice d'intensité d'implication séquentielle inspiré de l'indice d'implication (Gras et al., 2004).

Nous proposons dans cet article une nouvelle méthode qui extrait des règles d'association entre événements à partir de séquences multiples. Nous utilisons pour cela des modèles

[1] Etude soutenue par le Fonds national suisse de la recherche (FNS) FN-100015-122230

de durée, également appelés modèles de survie. Ces modèles de durée, et en particulier les modèles de régression semi-paramétrique à risques proportionnels, permettent d'évaluer l'influence d'une variable ou d'un événement sur le risque qu'un événement, prédéfini, se produise. L'intérêt de l'utilisation des modèles de durée dans le cadre de l'analyse de séquences d'événements est multiple. Premièrement, ces modèles permettent de tenir compte des données censurées, c'est-à-dire des séquences qui s'arrêtent avant que l'événement d'intérêt ait été observé. Deuxièmement, le *timing* des événements, c'est-à-dire les écarts qui existent entre eux, sont également pris en compte et apparaissent dans la valeur des coefficients estimés. Troisièmement, l'utilisation de modèles de régression permet de mettre en évidence des relations statistiquement significatives entre des événements, même peu fréquents. Cette méthode permet donc de s'affranchir du critère de *support minimum* utilisé pour réduire le nombre de règles explorées. Enfin, les coefficients produits par les modèles de durée que nous utilisons sont facilement interprétables comme des multiplicateurs du risque de subir un événement selon que l'on a ou non subi d'autres événements auparavant.

En plus de la simulation présentée dans cet article, l'extraction de règles d'association séquentielle a été utilisée pour analyser des données sur les événements de vie familiale extraites du Panel suisse de ménages. Cette méthode permet ainsi d'explorer de quelle manière les événements de vie s'influencent et de mettre en évidence une structure séquentielle entre eux. L'intérêt de l'utilisation de cette méthode en sciences sociales réside dans le fait que les données d'observation sont souvent censurées. La prise en compte de la censure permet ainsi d'évaluer correctement le risque de subir un événement en considérant les périodes durant lesquelles les individus sont soumis au risque sans que l'événement se produise.

Dans cet article, nous introduisons en premier lieu la problématique de la fouille de séquences d'événements et des règles d'association séquentielle entre événements. Ensuite, nous présentons le type de modèle de régression particulier que nous utilisons dans notre méthode, le modèle semi-paramétrique à risques proportionnels de Cox. Un algorithme qui permet la fouille automatique d'un ensemble de sous-séquences pour en extraire les règles significatives est décrit, ainsi qu'un moyen de visualiser les règles extraites. Finalemement, des données artificielles sont créées afin de tester la pertinence de la méthode proposée pour l'extraction de règles. Une comparaison entre les résultats obtenus par notre algorithme et les mesures de support, de confiance et de rappel est également présentée.

2 Fouille de séquences

La fouille de séquences a comme objectif d'extraire à partir d'un ensemble de séquences d'événements des motifs séquentiels fréquents, appelés également sous-séquences. Elle a été formalisée pour la première fois par Agrawal et Srikant (1995). Nous présentons ici une formulation inspirée de celle de Zaki (2001).

On a $I = \{i_1, i_2, ..., i_m\}$, un ensemble d'éléments distincts. Un événement e est composé d'éléments non-ordonnés, noté $(i_1 i_2 ... i_k)$, et une séquence est une liste ordonnée d'événements. Une séquence α se représente de la manière suivante : $e_1 \rightarrow e_2 \rightarrow ... \rightarrow e_q$, chaque e_i étant un événement. On considère α comme une sous-séquence de β si chaque événement de α est un sous-ensemble de β et que l'ordre entre les événements est préservé. Une sous-séquence est dite fréquente si son nombre d'occurrences, aussi appelé *support*, est supérieur à un seuil fixé par l'utilisateur. Le nombre d'occurrences correspond au nombre de séquences

qui contiennent la sous-séquence. Une extension de la formulation originelle de Agrawal et Srikant (1995) a été proposée par Srikant et Agrawal (1996) sous le nom de motifs séquentiels généralisés. Cette extension ajoute au critère de support minimum des contraintes de temps lors de l'extraction des motifs séquentiels, telles que l'écart maximal entre deux événements ou la fenêtre d'observation. Ces méthodes de fouille de motifs séquentiels constituent une approche intéressante pour l'analyse de données séquentielles issues de nombreux domaines, comme par exemple pour l'étude des parcours de vie en sciences sociales (Ritschard et al., 2008).

3 Règles d'association

De la même manière que pour la fouille de motifs fréquents (Agrawal et al., 1993), il est possible d'extraire des règles d'association entre événements à partir des sous-séquences fréquentes. L'extraction de règles à partir de motifs, séquentiels ou non, consiste à séparer le motif en une partie « prémisse » et une partie « conclusion ». Ainsi, à partir d'une sous-séquence composée de trois événements, $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C}$, deux règles d'association peuvent être extraites : $\mathcal{A} \Rightarrow (B \rightarrow \mathcal{C})$ ou $(\mathcal{A} \rightarrow B) \Rightarrow \mathcal{C}$. Dans le cas de règles d'association extraites à partir de séquences, l'ordonnancement des événements doit être pris en compte dans le calcul de la mesure (Mannila et al., 1997; Blanchard et al., 2008). Les trois mesures, dérivées des mesures classiques de règles d'association, que nous utiliserons pour caractériser les règles séquentielles sont les mesures de *support*, c'est-à-dire la probabilité qu'une séquence contienne la règle, la *confiance*, qui est la probabilité conditionnelle d'observer la conclusion sachant qu'on a observé le prémisse (Eq. 1), et le *rappel*, qui représente la probabilité d'observer la règle sur la probabilité d'observer la conclusion (Eq. 2).

$$\text{confiance}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{P(\mathcal{A} \rightarrow \mathcal{B})}{P(\mathcal{A})} \tag{1}$$

$$\text{rappel}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{P(\mathcal{A} \rightarrow \mathcal{B})}{P(\mathcal{B})} \tag{2}$$

Il existe un grand nombre de mesures pour caractériser les règles d'association, et le choix d'une mesure dépend en grande partie du domaine d'application et des critères que la mesure doit satisfaire (Lenca et al., 2004; Lallich et Teytaud, 2004). Dans le cadre des règles séquentielles, les mesures requièrent une adaptation non-triviale pour tenir compte de l'ordre entre les événements qui composent la règle. Par exemple, un indice d'intensité de l'implication séquentielle a été développé (Blanchard et al., 2008) dans le cadre de la fouille d'épisodes fréquents au sein d'une séquence unique. Cet indice nécessiterait une adaptation supplémentaire pour être utilisé dans le cadre de multiples séquences.

En plus des problèmes d'ordonnancement des événements évoqués plus haut, le cas de la recherche et de la caractérisation de règles d'association séquentielle pose deux nouveaux problèmes inhérents aux données longitudinales se présentant sous la forme de séquences multiples. Le premier problème concerne le phénomène de censure de certaines observations. La question de la censure des observations est particulièrement importante dans des domaines d'application telles que les sciences sociales ou l'épidémiologie, puisqu'il est fréquent de perdre la trace d'un individu. Dans le cas de figure où l'observation d'un individu s'arrête avant qu'il ait connu un événement d'intérêt, il est important de considérer toute la période

d'observation comme une période durant laquelle cet individu a été soumis au risque de subir l'événement. L'utilisation de modèles de durée, aussi appelés modèles de survie, permet de tenir compte de ce phénomène de censure et d'obtenir une estimation plus réaliste de l'influence de variables sur l'occurrence d'un événement.

Le deuxième problème concerne l'écart qui existe entre la survenue de deux événements : dans le cas des mesures classiques d'association, c'est uniquement l'occurrence ou la non-occurrence des événements qui est prise en compte. Les modèles de durées, quant à eux, estiment l'influence de l'occurrence d'un événement sur l'occurrence d'un autre événement en tenant compte des écarts entre les événements. Ainsi, plus les écarts entre deux événements sont courts, plus le coefficient indiquant le risque d'observer l'événement $\mathcal{B}$ après avoir subi l'événement $\mathcal{A}$ sera élevé.

4 Analyse de durée

L'analyse de durée, également appelé analyse de survie, est un domaine de la statistique qui s'intéresse à la modélisation de la durée s'écoulant jusqu'à la survenue d'un événement. Cette durée peut correspondre au temps passé dans un état particulier, ou alors au temps qui s'est écoulé entre deux événements. La particularité de ce type d'analyse est de tenir compte des données censurées. Une observation est dite censurée lorsque l'événement auquel on s'intéresse ne s'est pas produit durant la période d'observation. Ces modèles sont utilisés, entre autres, dans les domaines de la santé, de la biologie, de la microéconométrie, de la démographie ou de la sociologie (Cameron et Trivedi, 2005; Therneau et Grambsch, 2000; Yamaguchi, 1991).

4.1 Modèles de régression semi-paramétriques à risques proportionnels

Nous décrivons ici le modèle de régression semi-paramétrique à risques proportionnels tel qu'il a été défini par Cox (1972). Dans ce type de modèle, on s'intéresse à l'influence de prédicteurs sur le risque qu'un événement se produise. La fonction $\lambda(t)$ est le risque instantané de subir l'événement au temps t sachant qu'il ne s'est pas produit avant le temps t (Eq. 3).

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{3}$$

Les modèles de régression semi-paramétrique à risques proportionnels décomposent le risque conditionnel $\lambda(t|\mathbf{x}, \beta)$, $\mathbf{x}$ étant un vecteur de prédicteurs et β les paramètres à estimer, en deux composants : un risque de base λ_0 et une fonction $\phi(\mathbf{x}, \beta)$. La spécificité de ce modèle est de ne pas spécifier la forme fonctionnelle de λ_0. Dans le cas du modèle de Cox, la fonction ϕ est spécifiée comme étant exponentielle ($\phi(\mathbf{x}, \beta) = exp(\mathbf{x}'\beta)$) (Cox, 1972). Le qualificatif de semi-paramétrique découle de la non-spécification de la forme de λ_0 associé à la spécification complète de la forme de la relation avec les prédicteurs. L'avantage de ne pas spécifier la forme de λ_0 est de s'affranchir d'hypothèses sur la relation entre le temps et le risque. Le modèle se présente ainsi sous la forme de l'équation 4.

$$\lambda(t|\mathbf{x}, \beta) = \lambda_0(t) \cdot exp(\mathbf{x}'\beta) \tag{4}$$

La relation entre le risque et les prédicteurs est multiplicative. Faisons varier le prédicteur x_k de $\mathbf{x}$ en augmentant sa valeur de 1 et notons ce nouveau profil $\mathbf{y}$. Cette variation d'une unité du prédicteur implique un effet multiplicatif sur le risque égal à $\exp(\beta_k)$ (équation 5).

$$\lambda(t|\mathbf{y},\beta) = \lambda_0(t) \cdot exp(\mathbf{x}'\beta + \beta_k) = exp(\beta_k) \cdot \lambda(t|\mathbf{x},\beta) \tag{5}$$

En d'autres termes, l'exponentiel du coefficient β_k peut s'interpréter comme un rapport de risque (*hazard ratio*) entre un profil de référence ($\mathbf{x}$) et un profil pour lequel la valeur du prédicteur x_k a augmenté de 1.

$$exp(\beta_k) = \frac{\lambda(t|\mathbf{y},\beta)}{\lambda(t|\mathbf{x},\beta)} = \frac{\lambda_0(t) \cdot exp(\mathbf{x}'\beta + \beta_k)}{\lambda_0(t) \cdot exp(\mathbf{x}'\beta)} = \frac{exp(\mathbf{x}'\beta + \beta_k)}{exp(\mathbf{x}'\beta)} \tag{6}$$

Comme on le voit dans l'équation 6, le coefficient β_k est indépendant du temps. Le rapport de risque est donc proportionnel et ne varie pas au cours du temps, d'où le nom de modèle à risques proportionnels. L'estimation des coefficients se fait par la maximimisation d'une vraisemblance partielle (Andersen et al., 1993). La significativé statistique d'un paramètre β_k peut ensuite être évaluée à l'aide de la statistique de Wald qui suit une loi du χ^2 à 1 degré de liberté sous l'hypothèse que le paramètre vaut 0 (équation 7).

$$\frac{\hat{\beta}_k^2}{\hat{\sigma}_k^2} \sim \chi^2 \text{ sous } H_0 : \hat{\beta}_k = 0 \tag{7}$$

4.2 Format de données

Dans un modèle de régression de durée, on s'intéresse à la survenue d'un événement précis (même s'il est possible d'étendre les modèles pour gérer l'occurrence de plusieurs événements, Hougaard 2000). Les données de durée d'un événement sont constituées de paires (t_i, δ_i), où t_i représente la durée de l'observation et δ_i indique si l'observation est censurée. Si $\delta_i = 0$, cela signifie que l'événement ne s'est pas produit entre le début de l'observation et le temps t_i. L'enregistrement d'un événement $\mathcal{A}$ se présente donc sous la forme suivante du tableau 1.

i	$t_i(\mathcal{A})$	$\delta_i(\mathcal{A})$	sexe
1	24	1	M
2	36	0	F

TAB. 1 – *Exemple de données de durée*

L'exemple montre que l'individu 1 a subi l'événement au temps 24, tandis que l'individu 2 a été observé jusqu'au temps 36 sans avoir subi l'événement $\mathcal{A}$. Ce type de représentation permet l'intégration de variables supplémentaires qui ne varient pas dans le temps, tel que le sexe dans l'exemple précédent. Dans le but de modéliser l'influence de la survenue d'un événement sur un autre, par exemple dans le cadre d'une règle $\mathcal{B} \Rightarrow \mathcal{A}$, l'événement $\mathcal{A}$ sera considéré dans le modèle de régression comme l'événement à prédire tandis que l'événement $\mathcal{B}$ devra être considéré comme une variable qui varie avec le temps. Si l'événement $\mathcal{B}$ n'est pas considéré de cette manière, il n'est pas possible de prendre en compte l'écart qui existe entre la survenue des deux événements ; on ne peut que savoir si l'événement s'est produit

à un moment ou à un autre de la période d'observation complète, donc soit avant soit après l'événement $\mathcal{A}$.

i	intervalle	$\mathcal{B}$	$\delta_i(\mathcal{A})$
1	(0 ;18]	0	0
1	(18 ;24]	1	1
2	(0 ;36]	0	0

TAB. 2 – *Exemple de données de durée avec une variable qui varie dans le temps*

Il est possible d'intégrer des variables variant avec le temps en utilisant la reformulation du modèle de Cox sous la forme d'un *counting process*. Cette reformulation a été établie par Andersen et Gill (Andersen et Gill, 1982; Andersen et al., 1993) et implémentée par Therneau et Grambsch (2000) dans les logiciels R et S-plus. Cette formulation permet de spécifier plusieurs épisodes par observation qui correspondent au changement de valeur des variables au cours du temps. Pour reprendre l'exemple du tableau 2, on introduit une variable B qui indique si l'individu a subi ou non l'événement $\mathcal{B}$ durant sa période d'observation.

Dans cet exemple, l'individu 1 a subi l'événement $\mathcal{B}$ au temps 18, puis l'événement $\mathcal{A}$ au temps 24. L'événement $\mathcal{A}$ étant l'événement de censure, la valeur de 1 signifie qu'à la fin de l'intervalle l'événement se produit. L'événement $\mathcal{B}$ étant considéré comme une variable qui varie dans le temps, sa valeur de 1 signifie que *durant* l'intervalle (18,24], l'individu avait déjà subi l'événement $\mathcal{B}$. On remarque que cette représentation des données de durée est spécifique à l'événement qui est considéré comme l'événement à prédire.

5 Extraction des règles

La méthode que nous proposons sélectionne des règles d'association dont les statistiques de Wald sur les coefficients estimés par un modèle de régression de Cox sont significatifs. La statistique de Wald correspond au carré du ratio entre le coefficient et son erreur standard. Il serait également envisageable d'utiliser un autre critère pour la sélection des règles, tels que le test du rapport des vraisemblances partielles ou le test du score (Hosmer et Lemeshow, 1999). Une sous-séquence qui contient plus de deux événements produit plusieurs règles d'association. Par exemple, à partir d'une sous-séquence $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C}$, deux règles d'association sont testées : la règle $\mathcal{A} \Rightarrow \mathcal{B}$ et la règle $\mathcal{B} \Rightarrow \mathcal{C}|\mathcal{A}$. La première règle peut s'interpréter directement comme l'influence de l'événement $\mathcal{A}$ sur le risque de subir l'événement $\mathcal{B}$. La deuxième règle évalue l'influence de l'événement $\mathcal{B}$ sur le risque de subir l'événement $\mathcal{C}$ pour la période qui suit la survenue de l'événement $\mathcal{A}$. Plutôt que de créer une nouvelle variable qui varie dans le temps et qui prend la valeur 1 lorsque l'événement $\mathcal{B}$ s'est produit après l'événement $\mathcal{A}$, seuls les épisodes correspondant aux périodes qui suivent la survenue de l'événement $\mathcal{A}$ sont sélectionnés lors de l'estimation du rapport de risque. Notons que cette règle pourrait également s'écrire $\mathcal{A} - \mathcal{B} \Rightarrow \mathcal{C}$ ou encore $\mathcal{A} \Rightarrow \mathcal{B} \Rightarrow \mathcal{C}$ si le risque de subir $\mathcal{B}$ après $\mathcal{A}$ est également estimé.

5.1 Algorithme

On a pu voir dans la section précédente que les données sont formatées en fonction de l'événement qui sera prédit dans le modèle de régression de Cox. La première étape consiste par conséquent à créer une représentation des données pour chaque type d'événement. Soit $E = e_1, e_2, ..., e_m$ l'ensemble des événements distincts. Pour chaque événement e_i, il est nécessaire de créer une représentation des données de durée avec e_i comme variable dépendante, c'est-à-dire qu'une valeur de 1 représente un événement et une valeur de 0 représente une censure, et tout $e_{k \neq i}$ comme variable variant dans le temps. Ces différentes représentations des données serviront à l'ajustement des modèles de régression de Cox, la représentation nécessaire étant choisie en fonction de l'événement « conclusion » , c'est-à-dire à prédire, de la règle. La deuxième étape est d'obtenir la liste des motifs séquentiels grâce une fouille sur l'ensemble des séquences effectuée avec un support minimum fixé à 1. Les règles sont ensuite extraites à partir de ces motifs séquentiels en utilisant l'algorithme suivant :

1. On fixe un seuil p qui correspond au niveau de significativité minimale requis pour conserver une règle.
2. Ce seuil p est ajusté pour pallier le problème des tests d'hypothèse multiples (Schaffer, 1995) en appliquant une correction de Bonferroni ($p_{ajusté} = p/n$ où n = nombre de tests).
3. Pour chaque motif séquentiel, si le nombre d'événement $n > 1$, alors pour k de 1 à n :
   (a) Si $k > 1$, on ne sélectionne que les individus ayant déjà vécu l'événement e_{k-1}, et pour chacun d'entre eux on ne conserve que la période postérieure à l'occurence de e_{k-1}.
   (b) On ajuste un modèle de Cox avec comme variable dépendante e_{k+1} et comme variable indépendante e_k.
   (c) Si le test de Wald sur le coefficient est significatif, on passe à la paire d'événements suivante dans la sous-séquence. S'il n'est pas significatif, la règle est abandonnée et on passe au motif séquentiel suivant.
4. Une représentation graphique des règles retenues est produite.

5.2 Visualisation des résultats

La visualisation des résultats se fait sous la forme de graphes, un par événement. Seules les règles retenues sont représentées, avec à chaque fois l'événement « prémisse » à gauche, l'événement « conclusion » à droite et entre les deux les rapports de risque estimés (HR pour *hazard ratio*) par le modèle de Cox. Les règles retenues par l'algorithme sont représentées sous la forme d'un graphe (Figure 1).

Dans cet exemple, les informations concernant trois règles sont représentées. Pour chacune des règles, le chiffre entre parenthèses indique la p-valeur de la statistique de Wald. Les résultats s'interprétent de la manière suivante :

- $\mathcal{A} \Rightarrow \mathcal{C}$: lorsque $\mathcal{A}$ s'est produit, le risque de voir $\mathcal{C}$ se produire est multiplié par 2.51.
- $\mathcal{B} \Rightarrow \mathcal{C}$: lorsque $\mathcal{B}$ s'est produit, le risque de voir $\mathcal{C}$ se produire est multiplié par 4.63.
- $\mathcal{A} - \mathcal{B} \Rightarrow \mathcal{C}$ ou $\mathcal{B} \Rightarrow \mathcal{C} \mid \mathcal{A}$: lorsque $\mathcal{B}$ s'est produit après $\mathcal{A}$, le risque que $\mathcal{C}$ se produise est 2.63 plus élevé.

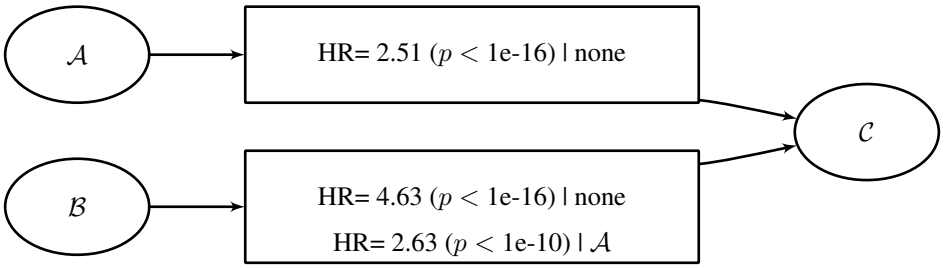

FIG. 1 – *Représentation graphique des règles d'association*

5.3 Complexité

Les performances de l'extraction de l'extraction de règles d'association séquentielle ont été évaluées de manière empirique en faisant varier le nombre de séquences créées artificiellement ainsi que le nombre d'événements 3. Les essais ont été effectués sur une machine avec un processeur Xeon à 3.2Ghz et 8GB de mémoire vive. Pour chaque essai, les durées de la transformation des données, de l'extraction des sous-séquences et de l'extraction des règles d'association ont été additionnées. L'algorithme est implémenté en R et en C++ et sera disponible dans la librairie R TraMineR (Gabadinho et al., 2009).

		nombre de séquences				
		500	**1000**	**2000**	**4000**	**8000**
	5	13.29	23.321	43.617	89.953	199.98
événements	**6**	50.356	93.794	193.871	414.831	963.911
	7	156.855	319.839	721.821	1807.56	4622.17

TAB. 3 – *Performances de l'algorithme en secondes*

On remarque que le temps d'exécution a une relation linéaire avec le nombre de séquences considérées, tandis qu'il semble avoir une relation exponentielle avec le nombre d'événenements.

6 Simulation

Afin d'évaluer la capacité de cet algorithme à mettre en évidence les liens existant entre des événements, des données ont été créées artificiellement à partir d'une distribution de probabilité de Weibull (Eq. 8 représente sa fonction de densité), fréquemment utilisée pour modéliser des durées. Une distribution de Weibull possède deux paramètres, le paramètre a qui est un paramètre d'échelle et le paramètre b qui est un paramètre de forme.

$$f(x) = exp\Big[-\Big(\frac{x}{b}\Big)^a\Big] \cdot \frac{a}{b} \cdot \frac{x}{b}^{(a-1)} \tag{8}$$

	paramètre a	# événements	médiane
$\mathcal{A}$	240	732	182
$\mathcal{B}$	$t_i(\mathcal{A}) + 48$	708	174
$\mathcal{C}$	$t_i(\mathcal{B}) + 48$	706	163
$\mathcal{G}$	192	803	148
$\mathcal{H}$	384	589	302

TAB. 4 – *Données simulées à partir d'une distribution de Weibull*

6.1 Création des données artificielles

Les temps d'occurrence de cinq événements, $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{G}$ et $\mathcal{H}$, ont été simulés en utilisant la procédure suivante. Un premier tirage de 1000 valeurs est effectué à partir d'une distribution de Weibull avec comme paramètres $a = 240$ et $b = 1.5$. Nous avons choisi la valeur de 240 pour simuler des données proches des événements de vie en prenant comme unité de temps le mois ; 240 correspond à un âge de 20 ans. Ce premier tirage correspond à $t_i(\mathcal{A})$, le temps d'occurrence de l'événement $\mathcal{A}$. Ensuite, pour chaque observation i, les temps d'occurrence $t_i(\mathcal{B})$ sont tirés à partir d'une distribution de Weibull avec comme paramètre $a_i = t_i(\mathcal{A}) + 48$. Les temps d'occurrence $t_i(\mathcal{C})$ sont tirés de la même manière en utilisant comme paramètre $a_i = t_i(\mathcal{B}) + 48$. Dans les deux cas, le paramètre b ne varie pas par observation et reste égal à 1.5. Les valeurs $t_i(\mathcal{G})$ et $t_i(\mathcal{H})$ sont quant à elles tirées de manière indépendante aux trois autres événements. Un dernier tirage, indépendant des cinq autres, est effectué afin de fixer les durées d'observation. Tout événement qui apparaît après la durée d'observation est ainsi censuré. Les données résultant de cette simulation sont résumées dans le tableau 4 ; le paramètre a de la loi de Weibull est spécifié, ainsi que le nombre d'événements après censure et le temps médian au moment de l'occurrence de l'événement.

Cette génération d'événements artificiels permet d'avoir des données comportant trois événements statistiquement liés par la distribution qui les génère ainsi que deux événements indépendants de tous les autres. Nous comparons ensuite les coefficients obtenus avec les mesures de support, de confiance et de rappel.

6.2 Résultats

Le tableau 5 présente les règles d'association séquentielle retenues par l'algorithme précédemment présentées et classées en fonction de la taille du rapport de risque (HR = *hazard ratio*). Le seuil p a été fixé à 0.01, puis ajusté à $3 \cdot 10^{-5}$. L'algorithme a conclu à la présence de 10 règles d'association significatives. On remarque que les deux premières règles correspondent à l'association entre les deux événements générés à la suite l'un de l'autre (voir les paramètres du tableau 4). Il est également intéressant de constater que la valeur du rapport de risque est moins grande entre deux événements associés mais plus éloignés. L'écart temporel moyen entre les événements $\mathcal{A}$ et $\mathcal{B}$ est de 5.53 mois, tandis qu'entre $\mathcal{A}$ et $\mathcal{C}$ elle est de 14.74 mois. Le coefficient tient compte de la durée entre les événements et respecte donc le caractère temporel des données séquentielles. Le tableau 5 indique également que le support de certaines règles, telles que $\mathcal{A} - \mathcal{C} \Rightarrow \mathcal{B}$, est tellement bas qu'elles auraient certainement été éliminées

durant une fouille de séquences fondée sur un support minimal. On voit pourtant que cette règle est clairement significative selon un modèle de régression de Cox.

Condition	Règle	Rapport de Risque	Confiance	Rappel	Support
	$\mathcal{B} \Rightarrow \mathcal{C}$	4.63	0.465	0.466	0.329
	$\mathcal{A} \Rightarrow \mathcal{B}$	3.6	0.425	0.439	0.311
	$\mathcal{C} \Rightarrow \mathcal{B}$	3.15	0.422	0.421	0.298
$\mathcal{C}$	$\mathcal{A} \Rightarrow \mathcal{B}$	2.65	0.263	0.292	0.087
$\mathcal{A}$	$\mathcal{B} \Rightarrow \mathcal{C}$	2.63	0.341	0.375	0.106
	$\mathcal{A} \Rightarrow \mathcal{C}$	2.51	0.387	0.401	0.283
	$\mathcal{B} \Rightarrow \mathcal{A}$	2.32	0.456	0.441	0.323
$\mathcal{A}$	$\mathcal{C} \Rightarrow \mathcal{B}$	2.2	0.265	0.241	0.075
	$\mathcal{C} \Rightarrow \mathcal{A}$	1.82	0.469	0.452	0.331
$\mathcal{C}$	$\mathcal{B} \Rightarrow \mathcal{A}$	1.81	0.369	0.332	0.110

TAB. 5 – *Résultats de l'extraction des règles*

Le tableau 6 présente les cinq premières règles avec le plus grand support. On remarque clairement que les trois mesures représentées ne permettent pas de distinguer les règles pertinentes. En effet, même des règles entre événements non-associés dans les données, telles que $\mathcal{G} \Rightarrow \mathcal{A}$ ou $\mathcal{C} \Rightarrow \mathcal{G}$, possèdent des valeurs élevées pour le support, la confiance et le rappel.

Règle	Support	Confiance	Rappel
$\mathcal{G} \Rightarrow \mathcal{A}$	0.378	0.47	0.52
$\mathcal{G} \Rightarrow \mathcal{H}$	0.369	0.46	0.63
$\mathcal{C} \Rightarrow \mathcal{A}$	0.331	0.47	0.45
$\mathcal{B} \Rightarrow \mathcal{C}$	0.329	0.46	0.47
$\mathcal{C} \Rightarrow \mathcal{H}$	0.327	0.46	0.56

TAB. 6 – *Règles extraites avec un support minimum*

Il existe des mesures plus appopriées pour le calcul de règles d'association séquentielle, notamment un indice d'implication séquentielle (Blanchard et al., 2008). Il faut cependant noter que l'indice d'implication séquentielle tel qu'il est présenté par Blanchard et al. (2008) nécessiterait une adaptation pour le cas où l'on ne travaille pas sur une longue séquence unique mais, comme dans cet article, sur des séquences multiples.

7 Conclusion

La méthode présentée dans cet article pour extraire des règle d'association séquentielle à partir d'un ensemble de séquences possède plusieurs avantages. Premièrement, elle permet une gestion des données censurées, un problème récurrent dans de nombreux domaines tels que les sciences sociales ou la recherche médicale. Deuxièmement, cette méthode fournit un coefficient de rapport de risque qui tient compte de la durée « moyenne » entre deux événements

et permet d'estimer l'influence de l'un sur l'autre ainsi que d'évaluer la rapidité avec laquelle un événement en entraîne un autre. Le coefficient de rapport de risque a également l'avantage d'être interprétable très facilement. Finalement, la significativité statistique nous donne un critère de sélection des règles, sans se baser sur leur fréquence, même si la fréquence joue un rôle dans le calcul de la significativité. Notons que ceci pourrait conduire à retenir trop de règles dès lors que l'on applique la méthode sur de grands jeux de données et pourrait lui faire perdre son caractère discriminant.

Cet article ouvre la perspective à d'autres travaux. En premier lieu, une adaptation de l'indice d'implication séquentielle (Blanchard et al., 2008) aux séquences multiples permettrait d'obtenir une comparaison des performances de notre méthode avec une mesure plus pertinente que la confiance ou le rappel. En deuxième lieu, nous avons montré que la méthode fournit des résultats pertinents sur des données générées artificiellement. Il est par conséquent envisageable d'utiliser cette méthode sur des données réelles, comme des données sur les événements de vie issues de questionnaires rétrospectifs.

Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., pp. 207–216.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. S. P. Chen (Eds.), *Eleventh International Conference on Data Engineering*, Taipei, Taiwan, pp. 3–14. IEEE Computer Society Press.

Andersen, P., O. Borgan, R. D. Gill, et N. Keiding (1993). *Statistical models based on counting processes*. Springer Series in Statistics. New York : Springer-Verlag.

Andersen, P. et R. Gill (1982). Cox's regression model for counting processes, a large sample study. *Annals of statistics 10*, pp. 1100–1120.

Blanchard, J., F. Guillet, et R. Gras (2008). Assessing the interestingness of temporal rules with sequential implication intensity. In R. Gras, E. Suzuki, F. Guillet, et F. Spagnolo (Eds.), *Statistical Implicative Analysis : Theory and Applications*, pp. 55–71. Springer.

Cameron, A. C. et P. K. Trivedi (2005). *Microeconometrics : Methods and applications*. Cambridge : Cambridge University Press.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), pp. 187–220.

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2009). Mining sequence data in R with the TraMineR package : A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). Quelques critères pour une mesure de la qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information RNTI E-1*, pp. 3–30.

Hosmer, D. W. J. et S. Lemeshow (1999). *Applied Survival Analysis, Regression Modeling of Time to Event Data*. New York: Wiley.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

Lallich, S. et O. Teytaud (2004). Evaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information RNTI-E-1*, pp. 193–218.

Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2004). Evaluation et analyse multi-critères des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information RNTI-E-1*, pp. 219–246.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

Ritschard, G., A. Gabadinho, N. S. Müller, et M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Schaffer, J. (1995). Multiple hypothesis testing. *Annual Review of Psychology 46*, 561–582.

Srikant, R. et R. Agrawal (1996). Mining sequential patterns: Generalizations and performance improvements. In P. M. G. Apers, M. Bouzeghoub, et G. Gardarin (Eds.), *Proc. 5th Int. Conf. Extending Database Technology, EDBT*, Volume 1057, pp. 3–17. Springer-Verlag.

Therneau, T. et P. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Yamaguchi, K. (1991). *Event History Analysis*, Volume 28 of *Applied Social Research Methods Series*. Newbury Park: SAGE.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1/2), 31–60.

Summary

Association rules mining is a thriving research field in data mining. These methods can also be applied to sequential data. Two problems arise when one wants to apply association rules mining to sequential data. First, the main criterium used to extract sequential patterns is their frequency. However, two events might be strongly associated even if they do not happen frequently. Second, association rules measures do not take into account the temporal aspect of sequential data, like the importance of the duration between two events or the problem of censured obsevations. In this article, we propose a method to extract significant associations between events using duration models. Association rules are extracted from each sequential pattern observed in a set of sequences. Then, the influence on the risk that the "conclusion" event occurs after the "premise" event(s) is estimated using a proportional hazard semi-parametric duration model. This paper presents the method and a comparison with some classical association measures.