Chapitre 5 : Analyse Implicative Séquentielle

Julien Blanchard, Fabrice Guillet, Régis Gras

Equipe Connaissances & Décision (COD)
LINA – FRE CNRS 2729 – Ecole Polytechnique de l'Université de Nantes
julien.blanchard@polytech.univ-nantes.fr

Résumé. La découverte de motifs fréquents dans des séquences (généralement des séquences temporelles d'évènements) est l'une des tâches majeures de la fouille de données. Dans cet article, nous nous intéressons à l'évaluation de la qualité des règles séquentielles. Nous proposons une mesure inédite nommée *SII* qui évalue la significativité des règles au regard d'un modèle probabiliste. Les simulations numériques montrent que *SII* que a des caractéristiques uniques en comparaison aux autres mesures de qualité de règles séquentielles.

1 Introduction

La découverte de motifs fréquents dans des séquences symboliques (généralement des séquences temporelles d'évènements) est l'une des tâches majeures de la fouille de données. Les travaux de recherche dans ce domaine se divisent en deux catégories :

- la découverte d'*épisodes* fréquents dans une longue séquence d'évènements (approche initiée par Mannila, Toivonen, et Verkamo Mannila et al. (1995) Mannila et Toivonen (1996)),
- la découverte de *motifs séquentiels* fréquents dans un ensemble de séquences d'évènements (approche initiée par Agrawal et Srikant Agrawal et Srikant (1995) Srikant et Agrawal (1996)).

Episodes et *motifs séquentiels* sont des structures séquentielles, c'est-à-dire définies avec un ordre (partiel ou total). Une telle structure peut être par exemple :

petit-déjeuner then *déjeuner* then *dîner*

La structure est qualifiée par sa fréquence (ou support) et généralement par des contraintes sur les positions des évènements, comme une fenêtre maximale de temps "moins de 12 heures séparent *petit-déjeuner* et *dîner*" Srikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Höppner (2002) Sun et al. (2003).

La différence entre *épisodes* et *motifs séquentiels* réside principalement dans la mesure de leur support : la fréquence des *épisodes* est intra-séquence Mannila et al. (1997) Das et al. (1998) Weiss (2002) Höppner (2002) Sun et al. (2003) Yang et al. (2003), alors que la fréquence des *motifs séquentiels* est inter-séquences Agrawal et Srikant (1995) Srikant et Agrawal (1996) Spiliopoulou (1999) Zaki (2001) Han et al. (2005) (voir Joshi et al. (1999) pour une synthèse sur les différentes manières d'évaluer la fréquence). Ainsi, les algorithmes d'extraction d'*épisode* fréquents recherchent des structures qui se répètent souvent à l'intérieur d'une

même séquence. Au contraire, les algorithmes d'extraction de *motifs séquentiels* fréquents recherchent des structures qui se répètent dans de nombreuses séquences (indépendamment des répétitions dans chaque séquence).

L'extraction des *épisodes*/*motifs séquentiels* est souvent suivie d'une étape de génération de règles séquentielles, permettant d'effectuer des prédictions dans la limite d'une fenêtre de tempsSrikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Spiliopoulou (1999) Zaki (2001) Weiss (2002) Höppner (2002) Sun et al. (2003). De telles règles ont été utilisées par exemple pour faire de la prédiction de cours de bourse Das et al. (1998), ou d'évènements dans un réseau de télécommunication Mannila et al. (1997) Sun et al. (2003). Une règle séquentielle peut être par exemple :

$$\textit{déjeuner} \xrightarrow{6\text{h}} \textit{dîner}$$

Cette règle signifie : "si on observe *déjeuner* alors on observera sûrement aussi *dîner* moins de 6 heures après".

Dans cet article, nous nous intéressons à l'évaluation de la qualité des règles séquentielles. Il s'agit d'une question cruciale pour l'analyse de séquences puisque, du fait de la nature non supervisée des algorithmes d'extraction, les quantités de règles générées peuvent être énormes. Alors que la qualité des règles d'association a été largement étudiée dans la littérature (voir Blanchard et al. (2009) pour une synthèse), il existe peu de mesures dédiées à l'évaluation des règles séquentielles. En plus de la fréquence, on trouve un indice de confiance (ou précision) qui peut être interprêtée comme une estimation de la probabilité conditionnelle de la conclusion étant donnée la prémisse Srikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Spiliopoulou (1999) Zaki (2001) Weiss (2002) Höppner (2002) Sun et al. (2003). Une mesure de rappel est également parfois utilisée ; elle peut être interprêtée comme une estimation de la probabilité conditionnelle de la prémisse étant donnée la conclusion Weiss (2002) Sun et al. (2003). Dans Das et al. (1998) et Höppner (2002), les auteurs ont proposé une adaptation aux règles séquentielles de la J-mesure de Smyth et Goodman, un indice issu de l'information mutuelle[1]. Enfin, une mesure entropique est présentée dans Yang et al. (2003) pour quantifier l'information apportée par un épisode dans une séquence, mais cette approche n'envisage que des épisodes et non des règles de prédiction.

Poursuivant nos travaux débutés dans Blanchard et al. (2002) sur l'adaptation aux règles séquentielles de l'intensité d'implication Gras (1996), nous proposons dans cet article une mesure statistique originale pour la qualité des règles séquentielles[2]. Plus précisément, cette mesure évalue la significativité statistique des règles au regard d'un modèle probabiliste. La section suivante est dédiée à la formalisation des notions de *règle séquentielle*, d'*exemple d'une règle*, et de *contre-exemple d'une règle*, et à la présentation de la nouvelle mesure, nommée *Sequential Implication Intensity (SII)*. Dans la partie 4, nous étudions SII sur plusieurs simulations numériques et la comparons à d'autres mesures.

[1] La J-mesure est la part de l'information mutuelle moyenne relative à la vérité de la prémisse.

[2] Ces travaux ont été également présentés dans Blanchard et al. (2007).

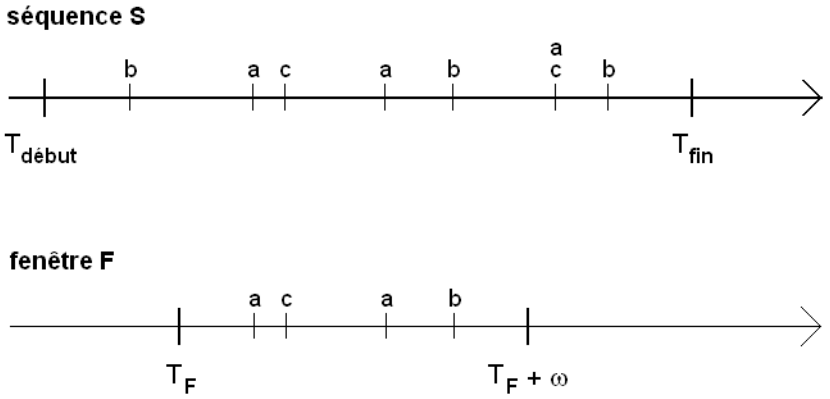


FIG. 1 – *Une séquence S d'évènements de $E = \{a, b, c\}$ et sa fenêtre F de longueur ω débutant en T_F.*

2 Mesurer la significativité des règles séquentielles

2.1 Contexte

Notre mesure SII porte sur des règles séquentielles extraites dans **une unique séquence**. Cette approche a l'avantage d'être facilement généralisable à un ensemble de séquences, par exemple en calculant une SII moyenne ou minimale sur l'ensemble. Les règles sont de la forme $a \xrightarrow{\omega} b$ où a et b sont des motifs séquentiels (ceux-ci peuvent eux-mêmes être régis par des contraintes de temps internes). Toutefois, dans cet article, nous nous limitons à des règles séquentielles où les motifs a et b sont chacun constitués d'un unique évènement.

La séquence étudiée est une séquence continue d'évènements instantanés (l'adaptation aux séquences discrètes est directe). De plus il est possible que deux évènements différents aient lieu au même instant. Ceci revient à se placer dans le cas des séquences étudiées par Mannila, Toivonen, et Verkamo Mannila et al. (1997). Pour extraire de la séquence les cardinaux nécessaires au calcul de SII, il suffit d'utiliser l'un de leurs algorithmes d'extraction de motifs, nommé Winepi Mannila et al. (1995) Mannila et al. (1997) (ou bien l'une de ses déclinaisons). Dans la suite, nous nous positionnons à l'étape des post-traitements, en considérant que Winepi a déjà été appliqué sur la séquence, et travaillons directement sur les cardinaux des motifs qui ont été extraits.

2.2 Notations

Soit E un ensemble fini de *types d'évènements* $E = \{a, b, c...\}$. Un *évènement* est un couple (e, t) où $e \in E$ est le type de l'évènement et $t \in \mathbb{R}_+$ est le temps d'apparition de l'évènement. Il est à noter que le terme d'évènement est communément employé pour désigner un type d'évènement, sans que ceci nuise à la compréhension.

Une *séquence d'évènements* S observée entre les instants $T_{début}$ et T_{fin} est une suite d'évènements

$$S = \Big((e_1, t_1), (e_2, t_2), (e_3, t_3), ...(e_n, t_n)\Big)$$

telle que :

$$\forall i \in \{1..n\}, (e_i \in E \wedge t_i \in [T_{début}, T_{fin}])$$
$$\forall i \in \{1..n-1\}, t_i \leq t_{i+1}$$
$$\forall (i,j) \in \{1..n\}^2, t_i = t_j \Rightarrow e_i \neq e_j$$

La longueur de la séquence est $L = T_{fin} - T_{début}$.

Une *fenêtre* sur une séquence S est une sous-séquence de S. Par exemple, une fenêtre F de longueur $\omega \leq L$ débutant à l'instant $t_F \in [T_{début}, T_{fin} - \omega]$ contient tous les évènements (e_i, t_i) de S tels que $t_F \leq t_i \leq t_F + \omega$.

Dans la suite, nous considérons une séquence S d'évènements de E.

2.3 Règles séquentielles

Nous établissons un cadre formel pour l'analyse des séquences en définissant les notions de *règle séquentielle*, d'*exemple d'une règle*, et de *contre-exemple d'une règle*.

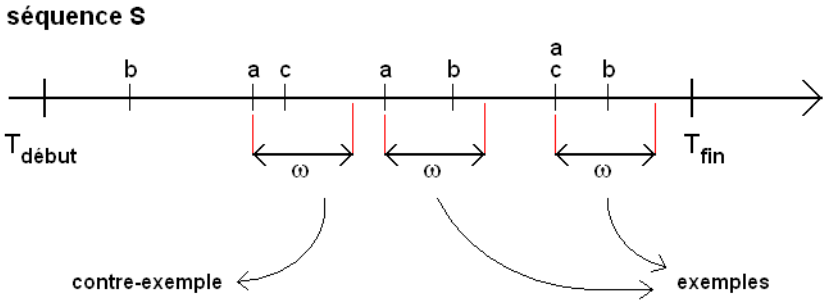


FIG. 2 – *Parmi les 3 fenêtres de longueur ω débutant sur des évènements a, on compte 2 exemples et 1 contre-exemple de la règle $a \xrightarrow{\omega} b$.*

Définition 1 Une **règle séquentielle** est un triplet (a, b, ω) noté $a \xrightarrow{\omega} b$ où a et b sont des évènements de types différents et ω est un réel strictement positif. Elle peut se lire de la manière suivante : "si la séquence comporte un évènement a alors un évènement b apparaît sûrement dans les ω unités de temps qui suivent".

Définition 2 Les **exemples** d'une règle séquentielle $a \xrightarrow{\omega} b$ sont les évènements a qui sont suivis d'au moins un évènement b à moins de ω unités de temps. Le nombre d'exemples de la règle est donc le cardinal noté $n_{ab}(\omega)$:

$$n_{ab}(\omega) = \left| (a,t) \in S \mid \exists (b,t') \in S, 0 \leq t' - t \leq \omega \right|$$

Définition 3 Les **contre-exemples** d'une règle séquentielle $a \xrightarrow{\omega} b$ sont les évènements a qui ne sont suivis d'aucun évènement b à moins de ω unités de temps. Le nombre de contre-exemples de la règle est donc le cardinal noté $n_{a\bar{b}}(\omega)$:

$$n_{a\bar{b}}(\omega) = \left| (a,t) \in S \mid \forall (b,t') \in S, (t' < t \vee t' > t + \omega) \right|$$

Contrairement aux règles d'association, n_{ab} et $n_{a\bar{b}}$ ne sont pas des constantes des données mais dépendent du paramètre ω.

La particularité de notre approche est qu'elle traite la prémisse et la conclusion de manières très différentes : les évènements a servent de référence pour la recherche des évènements b, c'est-à-dire que seules les fenêtres qui débutent par un évènement a sont prises en compte. Au contraire, dans la littérature sur les séquences, les algorithmes de type Winepi décalent (avec un pas constant) une fenêtre sur toute la longueur de la séquence Mannila et al. (1997). Cette démarche revient à considérer comme exemple de la règle séquentielle toute fenêtre qui présente a suivi de b, même si celle-ci ne débute pas par un évènement a. En comparaison, notre approche est moins complexe algorithmiquement.

Notons n_a le nombre d'évènements a dans la séquence. Nous retrouvons l'égalité bien connue $n_a = n_{ab} + n_{a\bar{b}}$. Une règle séquentielle $a \xrightarrow{\omega} b$ est entièrement caractérisée par le quintuplet $(n_{ab}(\omega), n_a, n_b, \omega, L)$. Les exemples d'une règle séquentielle étant maintenant définis, nous pouvons spécifier notre mesure pour la fréquence des règles :

Définition 4 La **fréquence** d'une règle séquentielle $a \xrightarrow{\omega} b$ est la proportion des exemples eu égard à la longueur de la séquence :

$$frquence(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{L}$$

Avec ces notations, la confiance, le rappel, et la J-mesure sont donnés par les formules suivantes :

$$confiance(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{n_a}$$

$$rappel(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{n_b}$$

$$J\text{-}mesure(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{L} \log_2 \frac{n_{ab}(\omega) L}{n_a n_b} + \frac{n_{a\bar{b}}(\omega)}{L} \log_2 \frac{n_{a\bar{b}}(\omega) L}{n_a (L - n_b)}$$

2.4 Modèle aléatoire

A l'instar de l'intensité d'implication pour les règles d'association Gras (1996), l'intensité d'implication séquentielle SII mesure la significativité des règles $a \xrightarrow{\omega} b$. Pour cela, elle quantifie l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}(\omega)$ eu égard à l'hypothèse d'indépendance entre les types évènements a et b. Dans une recherche de modèle aléatoire, nous supposons donc que les types d'évènements a et b sont indépendants. L'objectif est de déterminer la distribution de la variable aléatoire $\mathcal{N}_{a\bar{b}}$ (nombre de contre-exemples de la règle) étant donnés la longueur L de la séquence, les nombres n_a et n_b d'évènements de types a et b, ainsi que la taille ω de la fenêtre de temps utilisée.

Nous supposons que le processus d'arrivée des évènements de type b vérifie les hypothèses suivantes :

- les temps séparant les apparitions successives de b sont des variables aléatoires indépendantes,
- la probabilité qu'un b apparaisse dans un intervalle $[t, t+\omega]$ ne dépend que de ω.

De plus, deux évènements de même type ne peuvent arriver simultanément dans la séquence S (voir section 2.2). Dans ces conditions, le processus d'arrivée des évènements de type b est un processus de Poisson d'intensité $\lambda = \frac{n_b}{L}$. Le nombre de b apparaissant dans une fenêtre de longueur ω suit donc une loi de Poisson de paramètre $\frac{\omega.n_b}{L}$. En particulier, la probabilité pour qu'aucun évènement b ne se produise durant ω unités de temps est :

$$p = \mathrm{P}(Poisson(\frac{\omega.n_b}{L}) = 0) = e^{-\frac{\omega}{L}n_b}$$

Où qu'il apparaisse dans la séquence, un évènement a possède donc la probabilité fixée p d'être un contre-exemple, et $1-p$ d'être un exemple. Répétons n_a fois cette expérience aléatoire pour déterminer la loi du nombre de contre-exemples $\mathcal{N}_{a\bar{b}}$. Si ω est négligeable devant L, alors il est improbable que deux fenêtres de taille ω choisies aléatoirement se chevauchent, et nous pouvons considérer que les n_a répétitions de l'expérience sont indépendantes. Dans ces conditions, la variable aléatoire $\mathcal{N}_{a\bar{b}}$ est binomiale de paramètres n_a et p :

$$\mathcal{N}_{a\bar{b}} = Binomiale(n_a, e^{-\frac{\omega}{L}n_b})$$

Dans les conditions qui conviennent, la distribution binomiale peut être approximée par une seconde distribution de Poisson (même dans le cas de répétitions "faiblement dépendantes" –see Ross (2006)).

Définition 5 L'**intensité d'implication séquentielle (*SII*)** d'une règle $a \xrightarrow{\omega} b$ est définie par :

$$SII(a \xrightarrow{\omega} b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}}(\omega))$$

Numériquement, on a :

$$SII(a \xrightarrow{\omega} b) = 1 - \mathrm{P}(\mathcal{N}_{a\bar{b}} \leq n_{a\bar{b}}(\omega)) = 1 - \sum_{k=0}^{n_{a\bar{b}}(\omega)} \mathrm{C}_{n_a}^{k} (e^{-\frac{\omega}{L}n_b})^k (1 - e^{-\frac{\omega}{L}n_b})^{n_a - k}$$

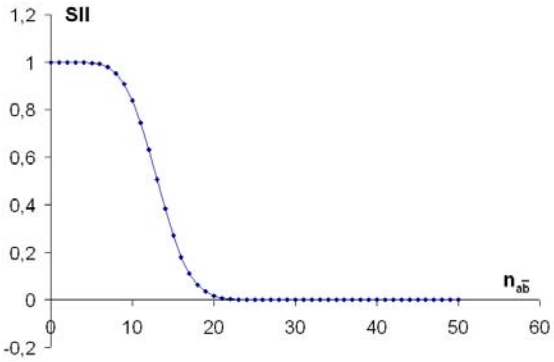


FIG. 3 – *Représentation de SII en fonction du nombre de contre-exemples.*

3 Propriétés et comparaisons

SII quantifie l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}(\omega)$ eu égard à l'hypothèse d'indépendance entre les types évènements a et b. En particulier, si $SII(a \xrightarrow{\omega} b)$ vaut 1 ou 0, alors il est invraisemblable que les types d'évènements a et b soient indépendants (l'écart à l'indépendance est significatif et orienté en faveur des exemples ou des contre-exemples). Ce nouvel indice peut être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse. Toutefois, à l'instar de l'intensité d'implication, il ne s'agit pas ici de tester une hypothèse mais bien de l'utiliser comme référence pour évaluer et ordonner les règles.

Dans la suite, nous testons SII dans plusieurs simulations numériques et le comparons à la confiance, au rappel, et à la J-mesure. Ces simulations soulignent les propriétés intuitives d'une bonne mesure d'intérêt pour des règles séquentielles.

3.1 Augmentation des contre-exemples

Dans cette section, nous étudions les mesures quand le nombre $n_{a\bar{b}}$ de contre-exemples augmente (avec les autres paramètres constants). Pour une règle $a \xrightarrow{\omega} b$, cela revient à espacer davantage les évènements a et b dans la séquence tout en conservant les mêmes quantités de a et de b. Cette opération fait passer les évènements a d'exemples à contre-exemples.

La figure 4 montre que SII fait clairement la distinction entre un nombre de contre-exemples acceptable (associé à des valeurs d'SII proches de 1) et un nombre de contre-exemples non-acceptable (associé à des valeurs proches de 0) au regard des autres paramètres n_a, n_b, ω, et L. Au contraire, la confiance et le rappel varient linéairement, tandis que la J-mesure produit des valeurs très peu discriminantes. A cause de sa nature entropique, la J-mesure peut même augmenter quand le nombre de contre-exemples augmente, ce qui est gênant pour une mesure de qualité de règles.

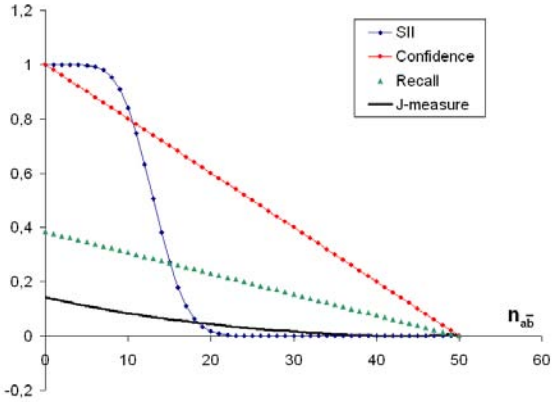


FIG. 4 – *SII, confiance, rappel, et J-mesure en fonction du nombre de contre-exemples.* $n_a = 50$, $n_b = 130$, $\omega = 10$, $L = 1000$

3.2 Allongement de la séquence

Nous désignons par allongement de la séquence l'opération qui consiste à rendre la séquence plus longue en y ajoutant de nouveaux évènements (évènements de nouveaux types) au début ou la fin. Pour une règle $a \xrightarrow{\omega} b$, une telle opération ne modifie pas les effectifs $n_{ab}(\omega)$ et $n_{a\bar{b}}(\omega)$ puisque la répartition des évènements de types a et b reste inchangée. Seule la longueur L de la séquence augmente.

La figure 5 montre que SII augmente avec l'allongement de la séquence. En effet, pour un nombre donné de contre-exemples, une règle est plus surprenante dans une séquence longue plutôt que dans une séquence courte, puisqu'il est moins probable que les a et b soient proches dans une séquence longue. Au contraire, des mesures comme la confiance et le rappel demeurent inchangées car elles ne tiennent pas compte de L (voir figure 6). La J-mesure varie avec L mais faiblement. Elle peut même décroître avec L, ce qui est contre-intuitif.

3.3 Réplication de la séquence

Nous appelons réplication l'opération qui allonge une séquence en la répétant γ fois successivement (nous faisons abstraction des éventuels effets de bord qui pourraient faire apparaître de nouvelles occurrences de motifs à cheval sur la fin d'une séquence et le début de la séquence répétée qui suit). Avec cette opération, les fréquences des évènements a et b et les fréquences des exemples et contre-exemples restent les mêmes.

La figure 7 montre que les valeurs de SII deviennent plus extrêmes (proches de 0 ou 1) avec la réplication. Ce phénomène s'explique par la nature statistique de la mesure. Une règle est en effet d'autant plus significative qu'elle est évaluée sur une séquence longue avec de nombreux évènements : plus la séquence est longue, plus on peut se fier aux déséquilibres entre exemples et contre-exemples observés dans la séquence, et plus on peut confirmer la bonne ou

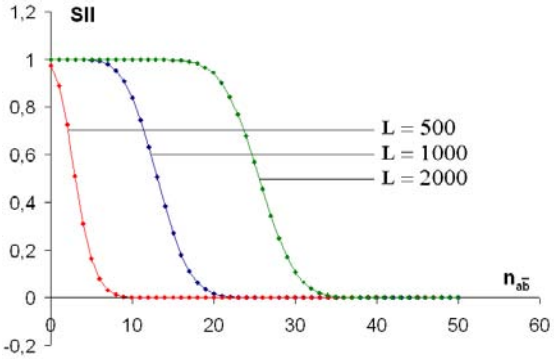


FIG. 5 – *Evolution de SII avec l'allongement de la séquence.* $n_a = 50$, $n_b = 130$, $\omega = 10$

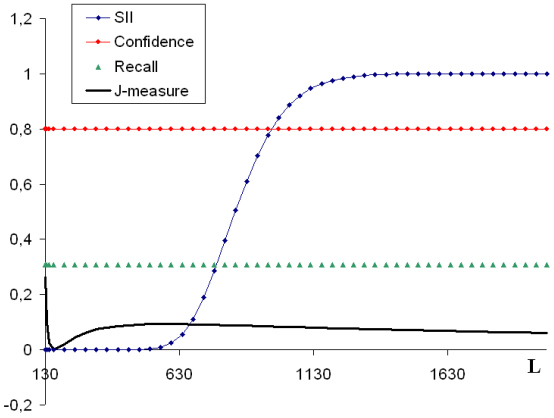


FIG. 6 – *Evolutions de SII, confiance, rappel, et J-mesure avec l'allongement de la séquence.* $n_a = 50$, $n_b = 130$, $n_{a\bar{b}} = 10$, $\omega = 10$

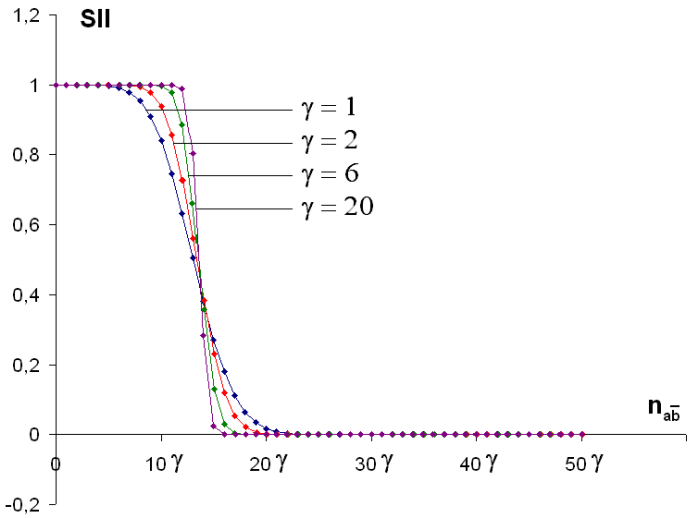
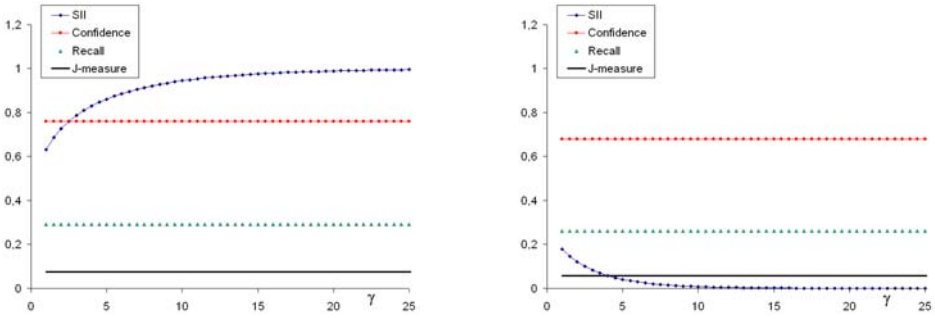


FIG. 7 – *Evolution de SII avec la réplication de la séquence.*
$n_a = 50 \times \gamma$, $n_b = 130 \times \gamma$, $\omega = 10$, $L = 1000 \times \gamma$



(a) $n_{a\bar{b}} = 12 \times \gamma$ (b) $n_{a\bar{b}} = 16 \times \gamma$

FIG. 8 – *Evolutions de SII, confiance, rappel, et J-mesure avec la réplication de la séquence.*
$n_a = 50 \times \gamma$, $n_b = 130 \times \gamma$, $\omega = 10$, $L = 1000 \times \gamma$

mauvaise qualité de la règle. Au contraire, les mesures fréquentielles comme la confiance, le rappel, et la J-mesure ne varient pas avec la réplication (voir la figure 8).

4 Conclusion

Dans cet article, nous avons étudié l'évaluation de la qualité des règles séquentielles. Tout d'abord, nous avons formalisé les notions de *règle séquentielle*, *exemple d'une règle*, et *contre-exemple d'une règle*. Nous avons ensuite présenté l'*Intensité d'Implication Séquentielle (SII)*, une mesure statistique originale qui évalue la significativité des règles séquentielles au regard d'un modèle probabiliste. Les simulations numériques montrent que *SII* que a des caractéristiques uniques en comparaison aux autres mesures de qualité de règles séquentielles. En particulier, SII est la seule mesure qui prenne en compte l'allongement de la séquence et la réplication de la séquence de manière appropriée.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the international conference on data engineering (ICDE)*, pp. 3–14. IEEE Computer Society.

Blanchard, J., F. Guillet, et H. Briand (2002). L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans les séquences de pannes d'ascenseurs. *Extraction des Connaissances et Apprentissage 1*(4), 77–88. Actes des journées Extraction et Gestion des Connaissances (EGC) 2002.

Blanchard, J., F. Guillet, et R. Gras (2007). On the discovery of significant temporal rules. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics SMC'2007*, pp. 443–450. IEEE Computer Society.

Blanchard, J., F. Guillet, et P. Kuntz (2009). Semantics-based classification of rule interestingness measures. In Y. Zhao, C. Zhang, et L. Cao (Eds.), *Post-Mining of Association Rules : Techniques for Effective Knowledge Extraction*, pp. 56–79. IGI Global.

Das, G., K.-I. Lin, H. Mannila, G. Renganathan, et P. Smyth (1998). Rule discovery from time series. In R. Agrawal, P. E. Stolorz, et G. Piatetsky-Shapiro (Eds.), *Proceedings of the fourth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 16–22. AAAI Press.

Gras, R. (1996). *L'implication statistique : nouvelle méthode exploratoire de données*. La Pensée Sauvage Editions. in French.

Han, J., J. Pei, et X. Yan (2005). Sequential pattern mining by pattern-growth : Principles and extensions. In W. W. Chu et T. Y. Lin (Eds.), *Recent Advances in Data Mining and Granular Computing (Mathematical Aspects of Knowledge Discovery)*, pp. 183–220. Springer-Verlag.

Höppner, F. (2002). Learning dependencies in multivariate time series. In *Proceedings of the ECAI'02 workshop on knowledge discovery in spatio-temporal data*, pp. 25–31.

Joshi, M., G. Karypis, et V. Kumar (1999). A universal formulation of sequential patterns. Technical report, University of Minnesota. TR 99-021.

Mannila, H. et H. Toivonen (1996). Discovering generalized episodes using minimal occurrences. In *Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 146–151. AAAI Press.

Mannila, H., H. Toivonen, et A. I. Verkamo (1995). Discovering frequent episodes in sequences. In *Proceedings of the first ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 210–215. AAAI Press.

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.

Ross, S. M. (2006). *Introduction to Probability Models*. 9th edition.

Spiliopoulou, M. (1999). Managing interesting rules in sequence mining. In *PKDD'99 : Proceedings of the third European conference on principles of data mining and knowledge discovery*, pp. 554–560. Springer-Verlag.

Srikant, R. et R. Agrawal (1996). Mining sequential patterns : generalizations and performance improvements. In *EDBT'96 : Proceedings of the fifth International Conference on Extending Database Technology*, pp. 3–17. Springer-Verlag.

Sun, X., M. E. Orlowska, et X. Zhou (2003). Finding event-oriented patterns in long temporal sequences. In K.-Y. Whang, J. Jeon, K. Shim, et J. Srivastava (Eds.), *Proceedings of the seventh Pacific-Asia conference on knowledge discovery and data mining (PAKDD2003)*, Volume 2637 of *Lecture Notes in Computer Science*, pp. 15–26. Springer-Verlag.

Weiss, G. M. (2002). Predicting telecommunication equipment failures from sequences of network alarms. In *Handbook of knowledge discovery and data mining*, pp. 891–896. Oxford University Press, Inc.

Yang, J., W. Wang, et P. S. Yu (2003). Stamp : On discovery of statistically important pattern repeats in long sequential data. In D. Barbará et C. Kamath (Eds.), *Proceedings of the third SIAM international conference on data mining*. SIAM.

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

Summary

In this article, we study the assessment of the interestingness of sequential rules (generally temporal rules). This is a crucial problem in sequence analysis since the frequent pattern mining algorithms are unsupervised and can produce huge amounts of rules. While association rule interestingness has been widely studied in the literature, there are few measures dedicated to sequential rules. Continuing with our work on the adaptation of implication intensity to sequential rules, we propose an original statistical measure for assessing sequential rule interestingness. More precisely, this measure named Sequential Implication Intensity (SII) evaluates the statistical significance of the rules in comparison with a probabilistic model. Numerical simulations show that SII has unique features for a sequential rule interestingness measure.