

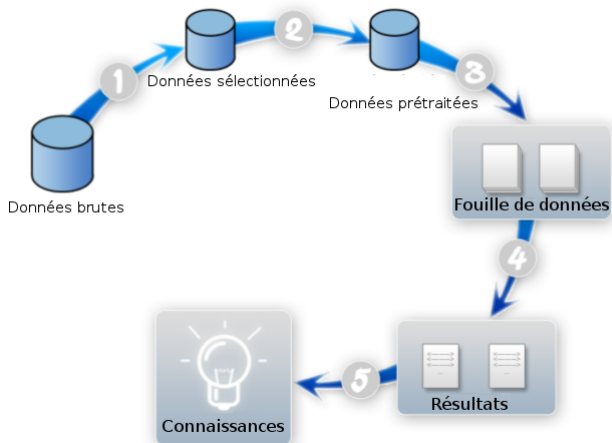
La confiance est dans l'air ! Application à l'identification des parcours hospitaliers

YVES MERCADIER

JESSICA PINAIRE, JÉRÔME AZÉ, SANDRA BRINGAY,
MAGUELONNE TEISSEIRE.



Extraction de connaissances



[Rabatel, 2011]

Plan

- 1 r-confiance
 - 2 Application aux données médicales.
 - 3 Conclusions et perspectives
- Règles d'association
 - Motifs séquentiels
 - Introduction de la mesure d'intérêt
 - Écriture formelle
 - Exemple

la Problématique du Panier de la Ménagère



un objet \iff un panier

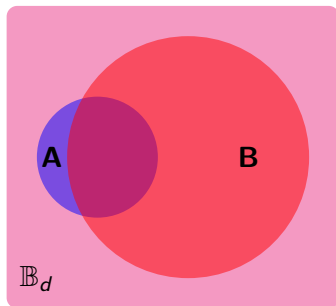
Objet = Pains Poivrons Raisins Pâtes Oeufs

Itemset = (Pains Poivrons Pâtes)

Association = (Pains Poivrons) (Pâtes Oeufs)

La confiance des règles d'association

$$\text{Confiance}_{\mathbb{B}_d}(\mathbf{A} \rightarrow \mathbf{B}) = \frac{\text{support}_{\mathbb{B}_d}(\mathbf{A} \rightarrow \mathbf{B})}{\text{support}_{\mathbb{B}_d}(\mathbf{A})} = \frac{n_{\mathbf{A} \rightarrow \mathbf{B}}}{n_{\mathbf{A}}}$$



L'extraction des règles d'association

$$RA_{\mathbb{B}_d, \sigma_s, \sigma_c}(\mathbf{I}_1 \rightarrow \mathbf{I}_2) = \{\mathbf{O} \in \mathbb{O}_b \wedge (\mathbf{I}_1, \mathbf{I}_2) \in \mathbf{O}^2 \mid \mathbf{I}_1 \cap \mathbf{I}_2 = \emptyset \\ \wedge \text{Support}_{\mathbb{B}_d}(\mathbf{I}_2 \rightarrow \mathbf{I}_2) \geq \sigma_s \wedge \text{Confiance}_{\mathbb{B}_d}(\mathbf{I}_1 \rightarrow \mathbf{I}_2) \geq \sigma_c\}$$

Où σ_s est un seuil de support minimal défini par l'utilisateur.

$$0 \leq \sigma_s \leq 1$$

Où σ_c est un seuil de confiance minimal défini par l'utilisateur.

$$0 \leq \sigma_c \leq 1$$

Motifs séquentiels



un objet \iff une succession de panier

Objet = < (Pain Poivron)(Poivron Raisins)(Pâtes)(Pain Poivron Oeufs) >

L'extraction des motifs séquentiels

$$MS_{\mathbb{B}_d, \sigma_s}(\mathbf{S}) = \{\mathbf{O} \in \mathbb{O}_b \mid \mathbf{S} \preceq \mathbf{O} \wedge \text{Support}_{\mathbb{B}_d}(\mathbf{S}) \geq \sigma_s\}$$

Où σ_s est un seuil de support minimal défini par l'utilisateur.

$$0 \leq \sigma_s \leq 1$$

[Agrawal and Srikant, 1995]

Candidat séquentiel

Un motif élémentaire est une partie du motif débutant par le premier itemset. Le motif complet n'est pas considéré comme un motif élémentaire.

Exemple :

Soit $M = \langle (ab)(c)(d)(e) \rangle$ un motif.

Ce motif génère 3 candidats séquentiels :

- $\langle (ab) \rangle$;
- $\langle (ab)(c) \rangle$;
- $\langle (ab)(c)(d) \rangle$.

La r-confiance élémentaire

$$r\text{-conf-e}(M, C) = \frac{\text{support}_B(M)}{\text{support}_B(C)} \quad (1)$$

Un motif de n itemsets génère n-1 valeurs de r-confiance élémentaire.

Agrégation de la r-confiance élémentaire

La r-confiance sera la proportion des motifs élémentaires issus du motif et ayant une r-confiance élémentaire supérieure à σ_{rce} par rapport au nombre d'itemsets constituant le motif.

Où σ_{rce} est un seuil de confiance élémentaire minimal défini par l'utilisateur.

$$0 \leq \sigma_{rce} \leq 1$$

La r-confiance

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

Base de données de séquences.

Nous étudions le motif séquentiel $\mathbf{M} = \langle (a, b)(c)(b, c) \rangle$ avec $\sigma_s > \frac{1}{7}$

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

TABLE –

Les candidats séquentiels du motif : $\mathbf{M} = \langle (a, b)(c)(b, c) \rangle$

$$\mathbf{C}_1 = \langle (a, b) \rangle$$

$$\mathbf{C}_2 = \langle (a, b)(c) \rangle$$

$$\text{support}(\langle (a, b) \rangle) = \frac{3}{7} = 0.43$$

$$\text{support}(\langle (a, b)(c) \rangle) = \frac{2}{7} = 0.29$$


$$r\text{-conf-e}(\mathbf{M}, \langle (a, b) \rangle) = \frac{\frac{2}{7}}{\frac{3}{7}} = 0.66$$

$$r\text{-conf-e}(\mathbf{M}, \langle (a, b)(c) \rangle) = \frac{\frac{2}{7}}{\frac{2}{7}} = 1$$

Calcul de la r-confiance

Puis, à partir d'un seuil fixé à $\sigma_{rce} = 0.7$,

$$r\text{-conf}(\mathbf{M}) = \frac{1+1}{3} = 0.66$$

$$\mathbf{M} = \langle (a, b)(c)(b, c) \rangle$$


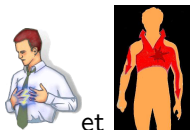
Plan

- 1 r-confiance
- 2 Application aux données médicales.
- 3 Conclusions et perspectives

Questions





- Y a-t-il une hospitalisation pour infarctus du myocarde consécutivement à une hospitalisation pour douleur thoracique ?
- Qu'est-ce qui est fréquemment associé à l'insuffisance cardiaque ?

Y a-t-il une hospitalisation pour IM consécutivement à une hospitalisation pour douleur thoracique ?



- Recherche des ensembles de motifs contenant : et
- pour $\sigma_{rce} = 0,9$ $r\text{-conf} = 0 \Rightarrow$ ré-hospitalisation pour d'autres motifs.

Qu'est-ce qui est fréquemment associé à l'insuffisance cardiaque ?

- Recherche des ensembles de motifs contenant : 
-    $\Rightarrow r\text{-conf} = 0.67$ pour $\sigma_{rce} = 0,9$.

Résumé des contributions :

- Proposition une nouvelle mesure d'intérêt : la r-confiance ;
- Validation de cette mesure dans le cadre d'un questionnaire médical.

La poursuite de notre travail :

- Validation de la mesure r-confiance sur les motifs séquentiels et autres ;
- Étendre d'autres mesures d'intérêt comme la mesure de Jaccard ou le lift ;
- Application à l'identification de trajectoires de patients.

References I

Agrawal, R. and Srikant, R. (1995).

Mining sequential patterns.

In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, Washington, DC, USA.

<http://dl.acm.org/citation.cfm?id=645480.655281>.

Rabatel, J. (2011).

Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles.

Thèse de doctorat, Université Montpellier II, LIRMM - Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier.

<http://www.lirmm.fr/~poncelet/publications/papers/theseRabatel.pdf>.