La r-confiance pour l'identification de trajectoires de patients

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

* LIRMM, UMR 5506, Université Montpellier, France
yves.mercadier@ac-montpellier.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

1 Introduction

Les méthodes de fouille de données exploratoires génèrent très souvent un grand nombre de motifs qu'il convient de filtrer à l'aide de mesures d'intérêt. 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'association. Nous proposons une mesure originale, appelée **r-confiance**, qui présente un double intérêt : (1) elle fonctionne pour tous les types de motifs (règle d'association, motif séquentiel, motif spatio-temporel) et (2) elle utilise comme opérateur d'agrégation « la proportion de position ».

1.1 La r-confiance

Avant de définir la r-confiance d'un motif de façon générale, nous définissons la r-confiance élémentaire d'un motif séquentiel (Pei et al., 2001).

Étant donné un motif séquentiel $M =< M_1, M_2, ..., M_n >$, un **candidat séquentiel** de M, $C =< M_1, M_2, ..., M_p >$, est défini comme une des sous-séquences préfixes de p items de M telle que $p < n$. Un motif séquentiel de longueur n est ainsi associé à $n - 1$ candidats séquentiels.

Soit M un motif et C un candidat séquentiel de ce motif. La r-confiance élémentaire, notée *r-conf-e*, est définie à partir des supports des séquences impliquées par :

Définition 1.

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

La confiance est dans l'air !

La r-confiance calculée pour le motif M correspond à l'agrégation des $n-1$ r-confiances élémentaires des candidats séquentiels le composant.

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.

Définition 2.

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

2 Application à l'identification de parcours hospitaliers

Nous avons étudié les parcours de soins des patients atteints d'un Infarctus du Myocarde au cours de la période 2009-2013. Ces données sont issues des bases hospitalières nationales du PMSI (Programme de Médicalisation du Système d'Information). L'évaluation de cette mesure par la mise en évidence de parcours connus du corpus médical est encourageante. Par la suite, nous souhaitons appliquer cette mesure dans l'identification de parcours hospitaliers types.

3 Conclusion

Nous avons proposé une nouvelle mesure d'intérêt qui est une extension de la confiance, définie pour les règles d'association, aux motifs séquentiels. Un expert en cardiologie est actuellement sollicité pour évaluer l'impact de la mesure proposée dans la validation des connaissances extraites, réel objectif d'une telle étude.

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.

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.

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. 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.