Increasing Secondary Diagnosis Encoding Quality Using Data Mining Techniques

L'Augmentation de la Qualité de Codage de Diagnostic Secondaire en Utilisant des Techniques de Fouille de Données

Ghazar Chahbandarian[1], Nathalie Bricon-Souf[1], Rémi Bastide[1], Jean-Christoph Steinbach[2]

1.University of Toulouse, IRIT/ISIS
F-81100 Castres, France
{ghazar.chahbandarian, nathalie.souf, remi.bastide}@irit.fr

2. Department of Medical Information
Centre Hospitalier Intercommunal de Castres Mazamet
F-81100 Castres, France
jean-christophe.steinbach@chic-cm.fr

Article accepté et présenté à la conférence internationale RCIS, co-localisée avec INFORSID 2016 à Grenoble. La version longue de l'article, en anglais, est disponible dans les actes de RCIS

RESUME. *Afin de mesurer l'activité médicale, les hôpitaux sont tenus de coder manuellement des informations concernant les séjours des patients hospitalisés en utilisant la Classification internationale des maladies (CIM-10). Cette tâche prend du temps et nécessite une formation importante pour le personnel, en particulier pour le codage des diagnostics associés (secondaires) qui ne sont pas toujours bien décrits dans les ressources médicales telles que la lettre de sortie et les dossiers médicaux. Nous proposons d'explorer des outils pour faciliter la tâche fastidieuse de codage de tels diagnostics.*

Notre approche exploite des techniques de fouille de données et plus précisément les arbres de décision dans le but d'explorer les bases de données médicales, et en particulier les diagnostics associés précédemment codés. On utilise les informations structurées comme l'âge, le sexe, le nombre de diagnostics, les actes médicaux présents dans les bases PMSI pour construire un arbre de décision facilement exploitable par un non spécialiste en informatique tel qu'un médecin, afin de souligner les diagnostics associés pour un séjour donné. Nous avons utilisé des données anonymisées extraites de la base de données PMSI de l'hôpital "Centre Hospitalier Intercommunal de Castres Mazamet", il contient environ 90.000 séjours d'hospitalisation entre 2011 et 2014.

Deux niveaux de granularité de diagnostic sont disponibles selon que l'on choisit de représenter le diagnostic de façon très précise (bas niveau de granularité) ou en se contentant de garder une information plus générale (haut niveau de granularité correspondant aux catégories de diagnostics. Les résultats indiquent qu'une amélioration de la performance pourrait être obtenue en utilisant le bas niveau de granularité de diagnostics et en équilibrant la répartition des exemples négatifs et positifs dans l'ensemble de l'apprentissage. En revanche, nous avons trouvé qu'il y a une variation entre les scores d'évaluation des diagnostics étudiés, par exemple, le score le plus élevé est 75% en utilisant la mesure F1 et le score le plus bas est 25% en utilisant la même mesure. En conséquence, des améliorations supplémentaires sont nécessaires pour obtenir une meilleure performance quel que soit le diagnostic codé. Cependant, le score moyen de tous les diagnostics associés étudiés est d'environ 80% en utilisant la mesure "accuracy", ce qui indique la prédiction des exemples négatifs est meilleur donc il pourrait être utile dans la prévention ou la détection des codages erronées dans les séjours hospitalisés.

.

ABSTRACT. In order to measure the medical activity, hospitals are required to manually encode information concerning an inpatient episode using International Classification of Disease (ICD-10). This task is time consuming and requires substantial training for the staff. We propose to help by speeding up and facilitating the tedious task of coding patient information, specially while coding some secondary diagnoses that are not well described in the medical resources such as discharge letter and medical records. Our approach leverages data mining techniques in order to explore medical databases of previously encoded secondary diagnoses and use the stored structured information (age, gender, diagnoses count, medical procedures…) to build a decision tree that assigns the proper secondary diagnosis code into the corresponding inpatient episode or indicates the impatient episodes that contains implausible secondary diagnoses. The results suggest that better performance could be achieved by using low level of diagnoses granularity along with adding some filters to balance the repartition of the negative and positive examples in the training set. The obtained results show that there is big variation in the evaluation scores of the studied diagnoses, the highest score is 75% using F1 measurement and the lowest 25% using F1 measurement which indicates further enhancements are needed to achieve better performance regardless of the encoded diagnosis. However, the average accuracy of all the studied secondary diagnoses is around 80% which indicates better negative predictions therefore it could be useful in the prevention or the detection of wrong coding assignments of secondary diagnoses in the inpatient stay.

MOTS-CLES : Fouille de données, apprentissage, arbre de décision, codage CIM-10.

KEYWORDS: Data mining, Machine learning, Decision tree, coding ICD-10