



Apport du Data Mining pour prédire la facture de patients hospitalisés

Alex Gnaegi, Mathieu Giotta, René Bonvin

Summary

The aim of this study is to assess the contribution of data mining technology in a medico-administrative context, i.e. to estimate the billing amount of a hospitalisation immediately on the patient's discharge, without having to wait for codification of diagnoses and procedures. The algorithms, decision trees and neural networks for this purpose have been tested. We conclude that from simple administrative data it is possible to estimate the amount of the bill with a coefficient of determination R^2 close to 0.9. The three most crucial variables in the model are length of stay, type of case (i.e. medical specialty) and age. Only outlier cases, namely very expensive stays, remain difficult to estimate. This study opens the door to a wider use of data mining technologies in medicine, e.g. by trying to predict clinical situations at risk of complications.

Objectifs

Lors des périodes de bouclage comptable, il importe de pouvoir valoriser rapidement et facilement les dossiers terminés mais non facturés. Si la facturation est basée sur une classification de type DRG (Diagnosis Related Groups), comme c'est le cas dans les hôpitaux du Réseau Santé Valais (RSV) pour les hospitalisations en chambre commune, il est nécessaire de disposer des diagnostics et des interventions effectuées durant l'hospitalisation en plus de variables administratives avant de pouvoir calculer le montant de la facture. Or la codification des diagnostics n'est parfois terminée, pour de multiples raisons, que plusieurs semaines après la sortie du patient et de ce fait il est extrêmement difficile de provisionner au plus juste les recettes correspondantes. Actuellement, les comptes provisionnent les montants non facturés sur la base du montant facturable moyen par type de cas. On dénombre une trentaine de type de cas qui correspondent en général au service médical qui a pris en charge de manière prépondérante le patient, comme par exemple la chirurgie maxillo-faciale, la gynécologie ou la médecine. Le type de cas est dès lors connu à la sortie du patient.

L'objectif de ce travail est d'évaluer l'apport des techniques de Data Mining dans un contexte médico-administratif, à savoir estimer le montant d'une facture d'un patient hospitalisé dès sa sortie à partir des informations disponibles dans les systèmes opérationnels. Une comparaison avec la méthode actuelle basée sur les types de cas est également prévue.

Méthodes

Logiciel utilisé

Le logiciel utilisé pour l'analyse de données de type Data Mining est Microsoft SQL Server 2005 Analysis Services (SSAS) car disponible sans coût supplémentaire au sein de notre département. Il intègre plusieurs algorithmes de Data Mining comme les arbres de décisions, les clusters, les réseaux bayésiens ou les réseaux neuronaux.

Source des données

Les données ont été extraites du système d'information administratif [1] basé sur le logiciel Opale d'Ordi-Conseils SA. Dans celui-ci se trouvent toutes les données relatives à l'administration du patient: date d'entrée et date de sortie de l'hôpital, durée de séjour, montant facturé, prestations, etc. C'est également dans ce système que se trouvent les codes diagnostiques et de traitements selon les nomenclatures ICD-10 et respectivement ICD-9-CM qui déterminent ensuite le code APDRG (All Patients Diagnosis Related Groups). Chaque code APDRG dispose d'un *cost weight* (CW), dont la valeur est ensuite multipliée par une constante pour obtenir le montant total de la facture du séjour hospitalier. Les données employées dans le cadre de cette étude proviennent de séjours hospitaliers somatiques des années 2006 et 2007 en chambre commune facturés selon les APDRG. Pour 2007, seuls les cas facturés au moment de l'étude (fin 2007) ont été intégrés dans l'analyse. La liste des variables potentiellement utiles figure dans le tableau 1. Celles-ci ont été regroupées en catégories.

Dr Alex Gnaegi
Département d'informatique
médicale et administrative
Institut Central des Hôpitaux
Valaisans
Av. Grand-Champsec 86
CH-1950 Sion
alex.gnaegi@ichv.ch

Procédures

La démarche suivie est classique pour des projets de Data Mining: définition du problème, préparation des données, création du modèle, exploration du modèle et validation du modèle [2].

Pour la préparation des données de test, nous avons extrait aléatoirement 160 cas pour l'année 2006 et 160 autres cas pour l'année 2007 parmi les 27 000 cas des données d'apprentissage. Les 160 cas sont répartis ainsi: 80 cas par centre hospitalier (Haut-Valais comprenant les hôpitaux de Brigue et Viège et Valais central comprenant

les hôpitaux de Sierre, centre valaisan de pneumologie de Montana, Sion et Martigny). Parmi ces 80 cas, 40 sont extraits des inliers [3] et des 40 cas restant, 20 proviennent des low-outliers alors que les derniers 20 cas sont des high-outliers. La répartition par centre hospitalier permet, par exemple, d'évaluer si des différences de codage ou de lourdeur de cas peuvent être remarquées. La répartition par inliers/outliers permet de simuler le fait que les cas non codés à la fin d'une période sont à 50% des cas inliers et 50% des cas outliers. Le pourcentage d'outliers a été fortement augmenté par rapport à la réalité, mais est

Tableau 1. Variables utilisées pour la définition des modèles d'analyse.

Nom de la variable	Catégorie	Remarques
Type d'admission	Données administratives	7 codes différents
Mode d'entrée	Variables d'entrée	24 codes différents
Provenance	Variables d'entrée	84 codes différents
Décision d'envoi	Variables d'entrée	8 codes différents
Genre d'admission	Variables d'entrée	9 codes différents
Mode de sortie	Variables de sortie	35 codes différents
Destination	Variables de sortie	87 codes différents
Prise en charge après la sortie	Variables de sortie	9 codes différents
Sexe	Données administratives	2 codes différents
Résidence ¹	Variables d'entrée	5 codes différents
Age à l'entrée	Données administratives	Calculé
Type de cas	Données administratives	49 codes différents
Classe	Données administratives	8 codes différents
Type de patient	Données administratives	31 codes différents
Tarif	Données administratives	44 codes différents
Type de taxe	Données administratives	39 codes différents
Groupe de classe	Données administratives	3 codes différents
Division	Variables sur le service	35 codes différents
Service	Variables sur le service	114 codes différents
Unité	Variables sur le service	77 codes différents
Médecin traitant ²	Données sur le médecin	169 codes différents
Spécialité du médecin traitant ²	Données sur le médecin	29 codes différents
Genre de médecin traitant ²	Données sur le médecin	12 codes différents
Durée de séjour	Données administratives	Calculé
Durée de séjour nette	Données administratives	Calculé
Liste des prestations TARMED	Prestations TARMED	4770 codes différents
Points de prestation TARMED	Prestations TARMED	Calculé
Valeur des prestations TARMED	Prestations TARMED	Calculé
Heures de soins intensifs	Données administratives	Calculé

Remarques: 1 résidence correspond à une zone de résidence du patient

2 médecin traitant correspond au médecin chef hospitalier responsable du patient.



justifié car ce sont ces cas qui doivent être estimés au plus proche en raison des montants prévisibles très élevés.

Parmi les différents algorithmes proposés par SSAS, seuls les algorithmes «Microsoft Decision Trees» et «Microsoft Neural Network» sont retenus car la variable à prédire est de type continu. Plusieurs modèles combinant l'un ou l'autre des deux algorithmes susmentionnés avec plusieurs variables d'entrée sont testés. Les comparaisons s'effectuent sur les montants facturables estimés par les différents modèles des 160 cas de test par rapport: a) aux montants facturables réels et b) à la méthode basée sur les types de cas appelée ci-après «méthode RSV». La performance des modèles est évaluée par le R2 (= coefficient de détermination) des droites de régression linéaire. Trente-cinq modèles différents avec l'algorithme des arbres de décisions ont été construits. Les variables des modèles avec un R2 >0,8 ont été reprises pour construire des modèles avec l'algorithme des réseaux neuronaux.

Résultats

Les performances des quatre meilleurs modèles (R2 >0,8) basés sur les arbres de décisions et les quatre modèles correspondants basés sur les réseaux neuronaux figurent dans le tableau 2. En l'occurrence, le meilleur modèle (Test_025_04_MDT) s'appuie sur l'algorithme des arbres de décision et ne nécessite que des variables faciles à collecter. Il obtient un R2 de 0,86 alors que la méthode RSV obtient seulement un R2 de 0,04. Tandis que le montant

total des factures réelles est de 1 502 618,65 francs, le modèle estime le montant total à 1 440 071,73 francs, soit une différence de -4,2%, à comparer au montant total de 1 116 259,01 francs obtenu par la méthode RSV (différence -34,6%). D'autres modèles sous-estiment un peu moins le montant total (en particulier le modèle Test_023_MDT avec une différence de -2,7%), mais par contre ne bénéficient pas de la même robustesse.

La figure 1 présente les différences relatives entre le meilleur modèle de Data Mining et la méthode RSV en rapport avec les résultats attendus pour les 160 cas de tests. Les cas sont regroupés en ordre croissant du pourcentage de différence par tranche de 20% entre 100% et plus de différence sous-estimée vs 100% et plus de différence surestimées. Le graphique permet ainsi de vérifier que les résultats du modèle Data Mining sont bien distribués de manière normale avec une moyenne la plus proche de 0%. Au contraire, la méthode RSV fournit des résultats inhomogènes avec une forte proportion de cas sous-estimés de 100% et plus.

Sans surprise, les trois variables en entrée qui ont le plus grand poids dans l'algorithme basé sur les arbres de décision sont la durée de séjour, le type de cas et l'âge à l'entrée. Par contre, les prestations médicales TARMED n'apportent pas d'informations utiles aux modèles de Data Mining probablement en raison du grand nombre de prestations différentes. Tous les modèles sous-estiment le montant total facturable possiblement en raison des cas extrêmes qui ne peuvent faire l'objet d'une simulation.

Tableau 2. Liste des meilleurs modèles.

Modèle	Algorithme	R2 obtenu	Variables utilisées
Test_023_MDT	Arbres de décision	0,8368	Données administratives
Test_023_MNN	Réseaux neuronaux	0,7827	Données administratives
Test_024_04_MDT	Arbres de décision	0,8572	Données administratives + variables entrées + variables sorties + spécialité du médecin
Test_024_04_MNN	Réseaux neuronaux	0,7515	Données administratives + variables entrées + variables sorties + spécialité du médecin
Test_025_04_MDT	Arbres de décision	0,8611	Données administratives + variables entrées + spécialité du médecin
Test_025_04_MNN	Réseaux neuronaux	0,7676	Données administratives + variables entrées + spécialité du médecin
Test_026_04_MDT	Arbres de décision	0,8437	Données administratives + spécialité du médecin
Test_026_04_MNN	Réseaux neuronaux	0,7668	Données administratives + spécialité du médecin

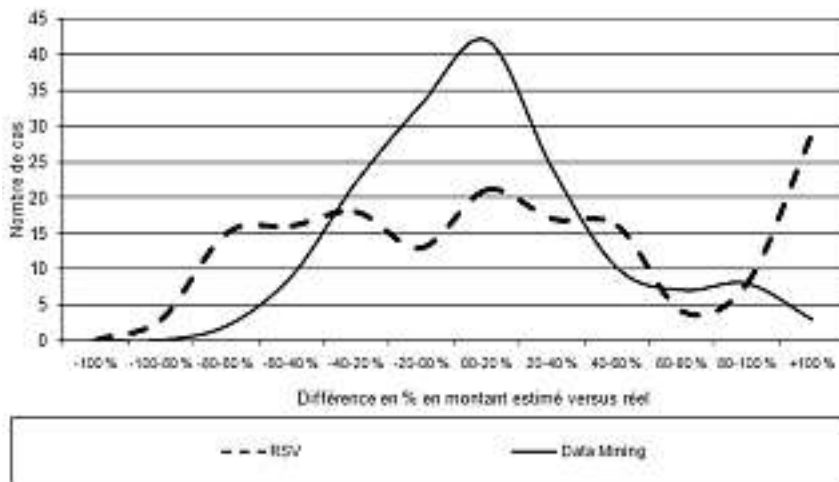


Figure 1

Différence relative entre les montants estimés et réels par tranche de 20%.

Conclusion

Cette étude illustre d'une part que les technologies de Data Mining sont désormais aisées à mettre en œuvre et d'autre part, elle montre que l'issue «médico-économique» d'un séjour hospitalier peut facilement être estimée avec les outils de Data Mining grâce à de simples données administratives. On peut dès lors espérer une utilisation plus large de ces technologies pour estimer cette fois-ci l'issue «médicale» d'un séjour comme par exemple le pronostic vital ou les complications éventuelles.

Références

- 1 Gnaegi A. ICHV – Application Infoval [Internet]. Applications Infoval. [cité 2008 Mar 30] Available from: <http://www.ichv.ch/default.asp?contentID=775>
- 2 Tang Z, MacLennan J. Data Mining with SQL Server 2005. Indianapolis: Wiley; 2005.
- 3 Schenker L. TAR APDRG 2006 Principes et règles de financement et de facturation par APDRG [Internet]. 2006 Jan; [cité 2008 Mar 28] Available from: http://www.apdrgsuisse.ch/public/fr/o_tar_apdrg_2006-f2.pdf