Classification cas ostéoporotique - témoin par Modélisation Graphique.

Makrem Djebali, Dhafer Malouche, Sylvie Sevestre-Ghalila

Classification cas ostéoporotique - témoin par Modélisation Graphique.

Makram Jebali, Dhafer Malouche, Sylvie Sevestre-Ghalila

Unité Signaux et Systèmes, Ecole Nationale d'Ingénieurs de Tunis, Université El Manar-Tunis, Tunisie,
Laboratoire MAP5, Université Paris Descartes, Paris 5, France.

Résumé : *Pour détecter l'ostéoporose à partir de sa probabilité de présence, on s'intéresse à estimer cette dernière à partir d'un jeu de données où certaines observations peuvent être manquantes au moment du calcul de la probabilité. Nous avons ici recours aux outils de modélisation graphique pour la sélection de variables dans le modèle de prédiction. Pour ce faire, on estime le modèle graphique CG par ajustement à nos données qui sont à ce stade complètes.Quand il s'agit d'estimer la robabilité de présence pour un patient pour lequel quelques variables peuvent être manquantes, une version modifiée du modèle graphique sélectionné sur l'ensemble des variables est alors estimé sur la base cette fois privée des variables manquantes. On compare cette approche à celle où le modèle est à la fois sélectionné et estimé sur la base privée des variables en question. Quelques soient ces variables, on constate une nette amélioration de la classification résultante.*

Abstract : The aim of this paper is the estimation of the probability of osteoporosis presence from a given dataset where some observations are missing. For this purpose we use the Conditional-Gaussian Graphical. The model selection is obtained by estimate the its graph from the whole complete dataset. To estimate the probability of presence for a patient for which observation of some variables is missing, a modified version of the graphic model previously selected on the whole variables is estimated on the database deprived of the mentioned missing variables at this step. We compare the performance of our estimation with the approach where both selection and estimation of the model are deduced from the incomplete database. Whatever are the missing variables, we notice a clear improvement of resulting osteoporosis-control classification.

Mots clés: Modèles Graphiques, Lois Conditionnelles-Gaussiennes, Ostéoporose, données manquantes.

Pour détecter l'ostéoporose à partir de sa probabilité de présence, on s'intéresse à estimer cette dernière à partir d'un jeu de données contenant, au moment de l'estimation, certains observations manquantes. Pour ce faire, nous avons recours aux outils de modélisation graphique pour la sélection des variables dans le modèle de prédiction.

Les modèles graphiques sont ceux générés par distributions de probabilités Conditionnelles-Gaussiennes (CG) (voir Lauritzen and Wermuth (1989), Lauritzen (1996), Edwards (2000)).

D'abord, on suppose que nos observations sont la réalisation d'un vecteur aléatoire $\mathbf{X} = (Y, X_v,\, v \in V)'$ où Y est la variable discrète réponse qui indique l'état de l'individu (ostéoropotique ou non) et les variables X_v, $v \in V$ sont les variables explicatives continues. Le vecteur aléatoire $\mathbf{X}$ suit une loi Contionnelle-Gaussienne si pour tout valeur possible y de Y, la loi du vecteur $(X_v,\, v \in V)'$ conditionnelle à $Y = y$ suit une loi Gaussienne multivariée de vecteur moyenne $\mu_y \in \mathbb{R}^{|V|}$ et de matrice covariance Σ_y.

Ensuite, on considère la partie de la base complète, on estime le modèle graphique CG ajustant nos données. Ceci sera fait utilisant les algorithmes d'estimations implémentées dans le logiciel `MIM`[1] (voir Edwards (2000)). Ces algorithmes sont basés sur l'estimation du maximum de vraisemblance de chaque matrice covariance Σ_y pour tout y valeur possible de Y. Le graphe obtenu G, de sommets Y et toutes les autres variables X_v, constitue alors un *a priori* sur le graphe exprimant les dépendances entre d'abord les variables X_v et entre Y et les variables X_v.

Pour maintenant estimer la probabilité de présence de l'ostéoporose pour un individu pour lequel la variable X_{u_0} n'est pas observé, on propose de remplacer G par un autre graphe, noté $G \setminus \{u_0\}$ de sommets Y et X_v, pour $v \neq u_0$, tel que chaque chemin de type $X_v \sim X_{u_0} \sim Y$ est remplacé par l'arête $X_v \sim Y$ dans ce nouveau graphe $G \setminus \{u_0\}$.

D'abord, on montre que sous une hypothèse de *fidélité* de la distribution de probabilité au graphe G, le nouveau graphe $G \setminus \{u_0\}$ représente aussi l'ensemble des indépendances conditionnelles dans la distribution du nouveau vecteur aléatoire $\mathbf{X} \setminus \{u_0\} = (Y, X_v,\, v \in V \setminus \{u_0\})'$. Ce résultat nous permet alors d'estimer tous les paramètres non-nuls dans le modèle graphique CG généré par $G \setminus \{u_0\}$ utilisant toujours le logiciel MIM.

On a ainsi appliquée cette méthode sur la base recueillie au Centre Hospitalier Régional d'ORLEANS par l'équipe INSERM (Institut National de la Santé Et de la Recherche Médicale). On a observé ainsi sur 264 personnes 21 variables continues dont 8 concernent l'état clinique du patient et 13 sont issues du programme de traitement d'image. On a comparé le taux de bon classement obtenu avec la méthode décrite ci dessus avec celle où le modèle en éliminant la variable en question est à nouveau estimer. Quelque soient les variables manquantes au moment de l'estimation de la prédiction, on constate que ce taux de bon classement s'améliore.

[1] http://www.hypergraph.dk

On comprend l'attrait de cette démarche dans le cas d'absence de la densité minérale osseuse qui explique plus de 50% de la présence de la maladie. En effet, on peut constater dans table (1) que le modèle estimé en absence de cette dernière est clairement peu performant comparé à celui correspondant au modèle estimé après modification du graphe. L'estimation du modèle résultant des deux étapes : sélection du modèle, correspondant à l'estimation du graphe, et celle de l'estimation des paramètres du modèle sélectioné. Dans la démarche proposée, seul la sélection du modèle se fait avec l'ensemble des données. On peut donc certainement expliquer la différence de performance par le fait que l'estimation du graphe à partir de l'ensemble des données est certainement plus performante que celle en l'absence de la variable fondamentale qu'est la densité minérale osseuse.

Ainsi, cette technique qui nous affranchit de l'étape de substitution de la donnée manquante, peut s'avérer attrayante pour d'autre classification où certaines données peuvent s'avérer difficiles à obtenir pour certains individus à classer.

Variables manquantes	Méthode par modification du graphe	Estimation du modèle à variables manquantes
Taille	86.7%	74.4%
Age	85.6%	80%
Age, Taille	81.11%	79.3%
Densité Minérale Osseuse	80.3%	67.7%

Table 1: Taux de bonne prédiction des deux méthodes,

Bibliographie

[1] S. Sevestre-Ghalila, N. Mellouli, A. Ricordeau, A. Benazza, C. Chappard, C.L. Benhamou, 2005, Descripteurs 2D de micro-architecture osseuse : aide à la détection de l'ostéoporose , p:105-112, TAIMA'05, Hammamet.
[2] Lauritzen, S.L. and Wermuth, N. 1989. Graphical Models for associations between variables some of which are qualitative and some quantitatives. Ann. Stat. 17: 31-57.
[3] Lauritzen, S. L., 1996. Graphical Models. New York : Oxford University Press.
[4] Edwards, D., 2000. Introduction to graphical modelling. Springer texts in statistics.