Commentaires sur une histoire de discrétisation

Jean-Louis Golmard

Dépt de Biomathématiques CHU Pitié-Salpêtrière
91, Bd de l'hôpital 75634 Paris cedex 13

L'article de Gilles Celeux et Claudine Robert contient, à mon sens, deux "messages", l'un principal et l'autre secondaire.

Le message principal est qu'il peut être dangereux de discrétiser d'emblée les variables continues. Sur ce point, non seulement je suis en plein accord avec cette idée, mais, si je peux me permettre une note personnelle, je suis très heureux de pouvoir citer une référence illustrant cette affirmation. C'est vrai que la discrétisation des variables est trop souvent un réflexe, et cet article est donc utile dans la mesure où il constitue une mise en garde contre cette pratique. Je suis tellement en accord avec la phrase ".... notre pratique nous a montré qu ... il était souvent plus avantageux de garder les variables dans leur forme originelle .... " que, poussant cette idée plus loin, j'ajouterai "Pourquoi ne pas essayer de garder continues les variables continues ?" Dans ce cadre, on peut citer, par exemple, les travaux de Lauritzen et Wermuth (1,2) ou Edwards (3), sur les modèles graphiques comprenant à la fois des variables continues et discrètes et généralisant plusieurs modèles multivariables usuels en statistique (modèles "Conditionnel-Gaussien"). Ce type de modèle est certainement appelé à devenir très utilisé dans les prochaines années.

Le message secondaire de l'article est une sorte de guide pratique de l'analyse de données, avec une référence explicite aux applications médicales. Je me sens beaucoup moins en accord avec ce guide, qui semble pourtant issu du simple bon sens. L'idée principale, ici, est qu'il faut soigneusement analyser les données avant de les discrétiser éventuellement. On peut difficilement être opposé à une telle affirmation, aussi ma critique porte plus sur un élément oublié que sur les éléments présents dans l'article : il me semble que les auteurs n'ont pas accordé suffisamment d'importance à la source d'information principale des systèmes experts, c'est-à-dire aux connaissances du domaine. Cet "oubli" est particulièrement regrettable quand le domaine est médical, car la Médecine est souvent considérée comme un domaine à connaissances "fortes". C'est vrai que les premiers systèmes experts ne mettant l'accent que sur "la modélisation des experts" ont montré leurs limites, mais il me semble qu'il ne faut pas aller trop loin dans l'autre sens et ne mettre en avant qu'une vision "statisticienne" de l'aide à la décision. Les

modèles statistiques finalement retenus (car c'est de modélisation statistique qu'il s'agit, que l'analyseur de données en soit conscient ou non) doivent tenir compte, à mon sens, à la fois des connaissances du domaine et des informations apportées par les échantillons (et mises en évidence par les statisticiens). On ne pourra jamais, à partir d'un échantillon de taille moyenne (quelques centaines d'individus) retrouver la totalité des connaissances médicales concernant les variables observées. Il doit donc exister une sorte de compromis entre les informations d'origine "humaine" et les informations provenant de l'échantillon ("statistiques"). Il est possible, et même probable, que cette dernière affirmation soit une évidence et qu'elle ait été implicitement admise dans l'article; ma critique, dans ce cas, serait une critique sur la forme et non sur le fond.

Références

(1) Lauritzen SL, and Wermuth N (1989) Graphical models for association between variables, some of which are qualitative and some quantitative. Annals of Statistics; **17**: 31-57.

(2) Lauritzen SL (1992) Propagation of probabilities, means and variances in mixed graphical association models. Journal of the American Statistical Association; **87**: 1098-1108.

(3) Edwards DE (1990) Hierarchical interaction models (with discussion). Journal of the Royal Statistical Society, Series B; **52**: 3-20.