Commentaires sur une histoire de discrétisation

Olivier Gascuel

Département d'Informatique Fondamentale, LIRMM
161 rue Ada, 34392 - Montpellier Cedex 5.

L'article de Gilles Celeux et de Claudine Robert montre bien ce qu'il ne faut pas faire, et ce qu'il faut faire, en matière de traitement et de discrétisation de données quantitatives.

Ce qu'il ne faut pas faire.

- Oublier le sens des données. Sachant que les classes de papillons sont invariantes par homothétie (Dieu, Th. 6.8), l'approche de normalisation qui consiste à considérer les variables *(z1/z4, z2/z4, z3/z4)* au lieu de *(z1, z2, z3, z4)* est de toute évidence préférable. C'est d'ailleurs la seule approche qui donne des résultats vraiment convaincants, les résultats de l'ACP non-normée ne faisant pas apparaître si clairement que cela l'existence de trois classes.

- Perdre la structure ordinale des données en utilisant, par exemple, une discrétisation puis le codage disjonctif complet. Cela explique à la fois l'insuccès de la première approche, mais aussi celui de la dernière dans laquelle on considère chaque variable quantitative comme une variable nominale prenant autant de valeurs que la variable quantitative en prend. En ce qui concerne le codage additif, il est clair que l'aspect ordinal est en partie préservé. Mais je me demande dans quelle mesure il est pris en compte par l'ACM et dans quelle mesure également les relations "logiques" qui lient les variables binaires issues d'une même variable quantitative ne perturbent pas l'ensemble du processus ? Je ne suis pas sur non plus que les variables aient été découpées assez finement. Quels auraient été les résultats avec un découpage en 4 ou 5 intervalles ?

Existe-t-il un lien entre ces erreurs et l'Intelligence Artificielle, comme cela est suggéré ?

- En ce qui concerne les réseaux de neurones, il n'y a à mon avis aucun lien possible. Le modèle le plus employé, le "perceptron multicouche", fonctionne avec des entrées quantitatives et l'ensemble des traitements qu'il réalise est de type numérique.

- En ce qui concerne les systèmes experts, il est exact que la tendance est à la discrétisation des variables quantitatives. Ceci parce que le schéma de base est du type

Si (ET Proposition1 Proposition2 .....) Alors Conclusion, les propositions étant de type logique. Chaque variable booléenne est alors une proposition possible. Cependant, il est inexact de dire que les variables quantitatives (comme la fièvre) ne sont jamais utilisées. De plus, ce schéma de base n'est pas le seul possible. Divers formalismes issus de la logique flou permettent d'exprimer des règles du genre *Plus le malade a de la fièvre, plus il est vraisemblable qu'il développe une maladie infectieuse*. Le flou n'est généralement pas apprécié parmi les membres de la communauté Statistique. En fait il correspond parfaitement à la situation à laquelle s'adressent les systèmes experts (en particulier médicaux) à savoir que l'on cherche à construire un système de décision à partir des seules connaissances d'un expert, en ne disposant ni de données ni de statistiques. Ces connaissances sont de nature qualitative, plutôt que quantitative. Le formalisme flou, réducteur sans doute mais simple à mettre en oeuvre, s'avère alors bien adapté. Par ailleurs lorsque l'on dispose de données, rien ne prouve que l'Analyse de Données et l'approche Système Expert se complètent bien. Des approches du type segmentation qui sont explicatives comme les systèmes experts, ou d'autres, issus de l'Apprentissage par exemple, sont peut-être plus pertinentes. Enfin, rien n'interdit en Intelligence Artificielle de discrétiser intelligemment les variables. Ceci m'amène au dernier point.

Ce qu'il faut faire.

Gilles Celeux et Claudine Robert donnent une solution pour les papillons qui est sûrement applicable à d'autres jeux de données : "discrétiser quelques composantes principales ou discriminantes". Il me semble qu'une méthode de segmentation comme celle de Williams & Lambert, associée à un codage additif fin (pourquoi pas autant d'intervalles qu'il y a de valeurs, comme cela est fait dans CART ?), permettrait à la fois de disposer de discrétisation des variables et en même temps de représentations explicites des classes. Bien sûr cela reste à montrer. Pour conclure, il me semble surtout que la discrétisation doit tenir compte du but poursuivi et sur ce point je rejoins tout à fait les auteurs.