Commentaires sur "Une Histoire de discrétisation"

Yves Lechevallier

INRIA-Rocquencourt

S'il est vrai que la conclusion d'une étude statistique est de nature discrète je suis surpris de voir, actuellement, promouvoir la discrétisation des variables en entrée de réseaux de neurones. Cette pratique n'a aucune raison d'être; par exemple l'algorithme de rétropropagation doit être appliqué sur des variables continues car algorithme permettant l'approximation numérique de la solution de son problème d'optimisation est la méthode des approximations successives. ("The multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function " D.W. Ruck, S. Rogers, ... ,IEEE Trans. on Neural Networks, Vol 1 , Dec 90).

Avant d'aborder le problème de discrétisation il faut analyser l'ACP effectuée sur z1, z2, z3 et z4. Le premier plan factoriel montre qu'il y a trois coefficients de proportionnalité différents entre ces 4 variables ce qui engendre trois formes linéaires. Par contre les papillons 22 et 6 possèdent des coefficients différents des autres. Il est alors évident que le fait de créer les variables zi/zj va nous permettent de "voir" ces trois groupes (Fig. 5) cependant le choix "innocent" des auteurs de prendre z4 comme diviseur isole le papillon 6 car c'est sur z4 qu'il possède une valeur pas très "normale". Cet effet serait réduit si les variables z2 ou z3 auraient été prises comme diviseur.

Dans ce problème la discrimination "naturelle" entre ces 3 groupes se fait sur les coefficients de proportionnalités entre variables. De ce fait la discrétisation casse cette relation et nous avons une multitude de petits groupes car les auteurs essaient d'approcher une fonction linéaire par une fonction constante par morceaux (fonction en escalier).

Cependant la discrétisation par le codage additif est, dans ce cas, beaucoup plus informative car il permet de tenir compte de l'ordre des variables. Ce groupe (8, 16, 2, 9, 22, et 15) visible sur la figure 10 du codage additif est aussi visible sur l'ACP (Fig. 3) sauf qu'il ne correspond pas sur ce plan à un ordre issu d'une fonction linéaire facilement reconnaissable par l'oeil (les auteurs ont plutôt associé les papillons 4, 14 aux papillons 9, 2 ,16, 8).

En conclusion le codage disjonctif complet et le codage additif montrent que la discrétisation entraîne deux formes de perte d'information sur les variables continues. Une information d'approximation ou de précision de la mesure en remplaçant par une seule valeur les valeurs mises dans la même classe; une information de structure qui, ici, correspond à l'ordre induit par la variable continue. Cette dernière est partiellement conservée par le codage additif ce qui permet de retrouver certains "spaghettis" de l'ACP de la figure 3.

Cette histoire de discrétisation me laisse perplexe. C'est plutôt une enquête policière qu'une analyse statistique car les auteurs, via l'ACP de la figure 3, ont trouvé un indice important puis, grâce à leur intelligence de statisticien, ont résolu l'énigme (Fig. 5). Et puis ils ont joué les naïfs en proposant l'apprentissage par l'exemple de l'intelligence artificielle avec des méthodes de discrétisation, d'où l'échec observé. Alors que la discrétisation des variables zi/zj c'est-à-dire des coefficients de proportionnalités des variables résout le problème statistique et permet de générer des règles de production intelligibles aux personnes de l'intelligence artificielle.