Utilisation d'amas pour la discrétisation de variables

Fabrice Muhlenbach – Ricco Rakotomalala

Laboratoire ERIC – Université Lumière – Lyon II
5 avenue Pierre Mendès-France
69676 BRON Cedex – France
{fmuhlenb, rakotoma}@univ-lyon2.fr

RÉSUMÉ. *Ce travail présente une nouvelle méthode de discrétisation des attributs continus dans le cadre de l'apprentissage supervisé. Cette méthode, appelée HyperCluster Finder, se caractérise par son fonctionnement supervisé et polythétique. Elle se base sur les graphes de voisinage. La construction d'un graphe de voisinage à partir de la base d'apprentissage permet de découvrir les amas homogènes dans l'espace de représentation. Les valeurs minimales et maximales de chacun des amas reportées sur chacune des dimensions permettent de définir des bornes afin de découper la variable continue en un ensemble d'intervalles. Les capacités de cette méthode de discrétisation sont illustrées sur une base sous forme de XOR.*

MOTS-CLÉS : *Discrétisation, Graphe de voisinage, Apprentissage supervisé, HyperCluster Finder*

1. Introduction

L'apprentissage supervisé consiste à chercher une liaison fonctionnelle entre un attribut à prédire Y (étiquette), catégoriel, et une série d'attributs prédictifs $X_j \, (j = 1, \ldots, p)$. Une grande partie des méthodes développées dans ce cadre, issues de l'intelligence artificielle, le sont exclusivement pour des attributs prédictifs catégoriels [MIC 83]. Dès lors, il est nécessaire, avant la phase d'apprentissage, de ré-encoder chaque attribut prédictif continu en attribut catégoriel constitué d'un ensemble d'intervalles disjoints : ce processus est connu sous le terme de discrétisation [DOU 95].

La discrétisation est un domaine très étudié en apprentissage supervisé. En effet, de la qualité du découpage effectué dépend en grande partie les performances du modèle que l'on va construire. On distingue généralement les méthodes de discrétisation selon deux critères : la prise en compte (discrétisation supervisée) ou non (discrétisation non supervisée) de l'attribut à prédire Y dans le découpage des attributs X_j ; le découpage des X_j de manière individuelle (monothétique) ou globalement en tenant compte des interactions (polythétique).

Dans cet article, nous présentons une nouvelle méthode de discrétisation supervisée polythétique. Elle repose sur la notion d'amas et travaille en deux phases. Dans un premier temps, à l'aide d'un algorithme de construction de voisinage, nous isolons des groupes d'individus de même étiquette, puis, dans un deuxième temps, nous procédons au découpage des attributs prédictifs X_j en projetant, sur chaque axe de l'espace de représentation, les extrémités des groupes. Par rapport aux méthodes non supervisées, cette approche a l'avantage de répondre directement à la problématique de l'apprentissage supervisé, en effet le découpage tient compte des valeurs prises par Y. Par rapport aux méthodes monothétiques, elle tient compte des éventuelles interactions entre les attributs prédictifs et permet ainsi la détection de concepts complexes tels que le « ou exclusif » (XOR).

Dans la prochaine section, nous présenterons l'approche fondée sur les graphes de voisinage que nous mettons en œuvre pour constituer des groupes homogènes. Puis nous exploiterons ces notions dans la section suivante pour décrire notre algorithme de discrétisation supervisée et polythétique. Nous illustrerons enfin le fonctionnement de l'algorithme sur un problème d'apprentissage complexe.

2. Graphes de voisinage et amas

Les graphes de voisinage sont des constructions issues de la géométrie computationnelle qui permettent de rendre compte de la proximité des données dans l'espace de représentation [ZIG 99]. Le Graphe des Voisins Relatifs (GVR) dont nous donnons la définition ci-après, le Graphe de Gabriel (GG), la Triangulation de Delaunay (TD) ou l'Arbre de Longueur Minimale (ALM) sont de telles structures de graphe planaires et connexes [PRE 85]. Dans cette approche, chaque individu de la base d'apprentissage est un point projeté dans un espace $\mathbb{R}^p$, p étant le nombre de variables prédictives. Nous utilisons ici le Graphe des Voisins Relatifs de Toussaint [TOU 80].

Définition : Le Graphe des Voisins Relatifs (GVR) est un graphe connexe dans lequel, si deux points sont reliés par une arête, alors ils vérifient la propriété ci-dessous où $d(u, v)$ est la distance euclidienne entre deux points u et v dans $\mathbb{R}^p$

$$d(a, b) \leq Max\left(d\left(a, c\right), d\left(b, c\right)\right) \forall c, c \neq a, b$$

Cette définition signifie que la *lunule* $L_{(u,v)}$ – ou zone d'influence – constituée par l'intersection des hypersphères centrées respectivement en u et en v et de rayon l'arête (u, v) ne contient aucun point. Sur la figure 1 (a), la lunule $L_{(u,v)}$ est indiquée en grisé. Le graphe complet est représenté sur la figure 1 (b).

Suivant Zighed et Sebban [ZIG 99], nous introduisons la notion d'amas pour rendre compte du fait qu'un ensemble de points proches les uns des autres sont de même étiquette. On appelle *amas* un sous-graphe connexe du graphe de voisinage dont tous les sommets sont de même étiquette. Sur la figure 1 (c), après avoir coupé 3 arêtes, nous obtenons un ensemble de 2 amas (à l'intérieur des rectangles à bouts arrondis). Pour construire les amas nécessaires à la caractérisation des structures de notre nuage de points, nous procédons comme suit :

1. construction du graphe de voisinage (GG, GVR, TD ou ALM) sur les données de la base ;
2. suppression des arêtes reliant des points de classes différentes.

Nous obtenons ainsi des sous-graphes connexes dont tous les éléments sont de la même classe.

3. Description de la méthode de discrétisation *HyperCluster Finder*

3.1. *Algorithme de discrétisation*

Notre méthode de discrétisation polythétique et supervisée procède par la recherche d'amas – ou plus exactement d'hyperamas car les points sont projetés dans un espace $\mathbb{R}$ à p dimensions. Cette méthode, appelée *HyperCluster Finder*, se déroule de la manière suivante :

1. génération d'un graphe de voisinage, ici le GVR (figures 1 (a) et 1 (b)) ;
2. coupure des arêtes reliant des points de classes différentes afin de constituer les amas ;
3. sélection des amas les plus pertinents (cf. section 3.2) ;
4. recherche des valeurs minimales et maximales de chacun des amas sur les p variables continues ;
5. utilisation de ces valeurs minimales et maximales pour définir des bornes sur chacune des p variables ;
6. pour chaque variable, le minimum et le maximum sont remplacés respectivement par moins et plus l'infini ;
7. les valeurs de ces bornes délimitent un ensemble d'intervalles sur chaque variable ;
8. les valeurs numériques des données sont ré-encodées par leur appartenance à un des intervalles obtenus.

Sur la figure 1 (c), les bornes qui ont été trouvées ($d1_{X_1}$, $d2_{X_1}$ et d_{X_2}) permettent de définir les intervalles $]-\infty; d1_{X_1}[$, $[d1_{X_1}; d2_{X_1}[$ et $[d2_{X_1}; +\infty[$ pour la variable X_1 ainsi que $]-\infty; d_{X_2}[$ et $[d_{X_2}; +\infty[$ pour la variable X_2.

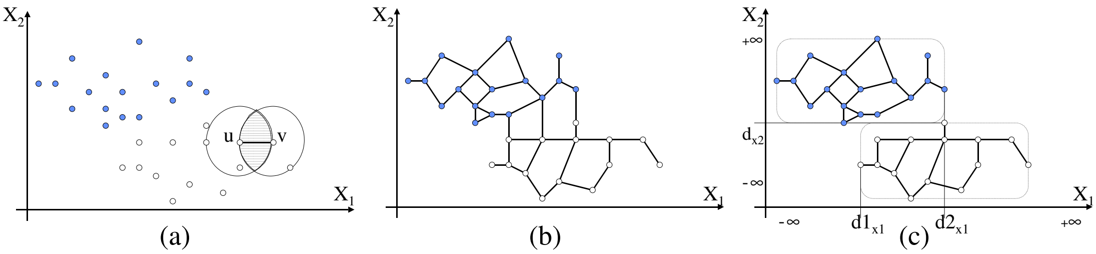

FIG. 1. *Graphe des voisins relatifs et amas : deux classes sont représentées, les points gris et les points blancs*

3.2. *Limitation du nombre d'intervalles par sélection des amas pertinents*

Afin d'éviter que des points isolés génèrent à eux seuls des amas, et donc des nouvelles bornes, il est possible de supprimer les amas considérés comme trop « petits » parmi les candidats à la constitution des intervalles. Une solution consiste à ne retenir que les amas dont le nombre d'individus est au moins égal à un nombre fixé *a priori*, ou un nombre dépendant de la taille de la base d'apprentissage ou de la population du plus grand des amas obtenus.

3.3. *Complexité algorithmique*

La méthode de discrétisation est d'une complexité algorithmique en $O(n^3)$ en raison de la création du graphe de voisinage dans $\mathbb{R}^p$ [TOU 80]. Diverses améliorations de l'algorithme existent à condition de travailler en 2 ou 3 dimensions [JAR 92] mais, dans le cas où il y a un nombre de variables prédictives $p > 3$, ce qui est le cas général, la complexité demeure en $O(n^3)$.

Dans le cas où un tel traitement n'est pas envisageable sur toute la base d'apprentissage, il est éventuellement possible de réaliser un échantillonnage sur la base à partir duquel la méthode *HyperCluster Finder* fournira les intervalles adaptés sur chacune des variables à discrétiser.

4. Illustration sur un exemple

Nous illustrons notre méthode en figure 2 à travers un exemple comprenant 100 individus décrits par 2 variables X_1 et X_2 continues. Les données, prenant des valeurs comprises entre -5 et $+5$ pour les deux variables, obéissent à une règle en XOR, le *ou* exclusif : les individus sont de la classe *Vrai* (en gris foncé) si $X_1 \geq 0$ et $X_2 < 0$ ou $X_1 < 0$ et $X_2 \geq 0$, et *Faux* (en blanc) sinon.

Sur les points projetés dans l'espace, nous construisons un graphe de voisinage (voir figure 2 (a)). Les arêtes reliant des points de classes différentes sont coupées et les amas ainsi obtenus sont recherchés. Sur cet exemple, nous ne conservons que 4 amas (deux amas d'un seul individu ne sont pas retenus sur les 6 générés). Les valeurs minimales et maximales de chaque amas sont reportées sur les axes en figure 2 (b). De cette manière, les intervalles formés pour la variable X_1 sont : $]-\infty; -0.3[$, $[-0.3; -0.1[$, $[-0.1; 0.1[$, $[0.1; 0.2[$, $[0.2; 4.7[$ et $[4.7; +\infty[$.

Pour la variable X_2 : $]-\infty; -4.9[$, $[-4.9; -0.2[$, $[-0.2; 0.0[$, $[0.0; 0.2[$, $[0.2; 4.7[$ et $[4.7; +\infty[$.

Ces intervalles, comme le montre la figure 2 (c), permettent de réaliser un pavage de l'espace. Dans notre exemple, chaque pavé obtenu ne contient que des points d'une seule classe.

Ne tenant pas compte de la valeur de la classe Y, les discrétisations non supervisées ne parviendront pas à trouver les bons intervalles, à moins de tomber par hasard sur les bornes adéquates. Quant aux discrétisations supervisées monothétiques, ne pouvant discriminer les classes sur chaque axe, elles conclurons que ni X_1 ni X_2 ne sont discrétisables sur cet exemple.

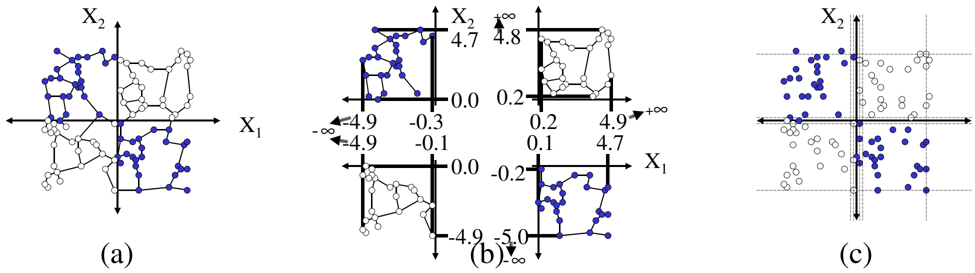

FIG. 2. *Méthode HyperCluster Finder appliquée sur une base d'apprentissage en XOR*

Après avoir remplacé les valeurs numériques des données par l'intervalle correspondant, la base peut ainsi être utilisée par des méthodes d'apprentissage supervisé nécessitant des variables prédictives catégorielles. Nous faisons toutefois remarquer que le concept du XOR ne sera retrouvé que par des méthodes qui procèdent de manière ascendante (partant des données particulières pour remonter aux concepts généraux) telles que les algorithmes *AQ* [MIC 83] ou *Data Squeezer* [MUH 01]. Avec une telle discrétisation, *Data Squeezer* fusionnera les pavés ayant des proportions de classes équivalentes (comme ceux que l'on retrouve en bordure des axes sur la figure 2 (c)), ce qui permettra à cette méthode d'apprentissage de retrouver le concept du XOR.

5. Conclusion et perspectives

Dans cet article, nous avons présenté *HyperCluster Finder*, une méthode de discrétisation procédant de manière polythétique et supervisée, ce qui permet en particulier de retrouver des intervalles adaptés au traitement des problèmes d'apprentissage comprenant des interactions entre les variables prédictives.

Dans des travaux futurs, nous envisageons de trouver les méthodes qui tirent le mieux parti de notre approche. Nous chercherons ainsi à mettre en évidence le fait que notre méthode de discrétisation ne se prête vraiment bien qu'aux méthodes d'apprentissage automatique capable de tenir compte du rôle simultané des variables lors de l'exploration des solutions. Pour ce faire, nous comptons réaliser un ensemble de tests comparatifs sur les résultats obtenus par des méthodes d'apprentissage descendantes (telles que les arbres de décision) et ascendantes (comme *AQ* ou *Data Squeezer*) après avoir procédé à des discrétisations non supervisées ou supervisées, et, dans ce dernier cas, monothétiques ou polythétiques.

6. Bibliographie

[DOU 95] DOUGHERTY J., KOHAVI R., SAHAMI M., Supervised and unsupervised discretization of continuous attributes, *Machine Learning : Proceedings of the* 12^{th} *International Conference (ICML-95)*, Morgan Kaufmann, 1995, p. 194–202.

[JAR 92] JAROMCZYK J. W., TOUSSAINT G. T., Relative Neighborhood Graphs and Their Relatives, *P-IEEE*, vol. 80, 1992, p. 1502–1517.

[MIC 83] MICHALSKI R. S., Theory and Methodology of Inductive Learning, MICHALSKI R. S., CARBONNEL J. G., MITCHELL T. M., Eds., *Machine Learning : An Artificial Intelligence Approach*, vol. 1, p. 83–134, Morgan Kaufmann, 1983.

[MUH 01] MUHLENBACH F., ZIGHED D. A., D'HONDT S., Génération de règles par compression, *ECA*, vol. 1, n^{o} 1–2, 2001, p. 93–104, Editions Hermès, Conférence EGC 2001.

[PRE 85] PREPARATA F. P., SHAMOS M. I., *Computational Geometry, an introduction*, Springer-Verlag, 1985.

[TOU 80] TOUSSAINT G. T., The Relative Neighborhood Graph of a Finite Planar Set, *Pattern Recognition*, vol. 12, 1980, p. 261–268.

[ZIG 99] ZIGHED D. A., SEBBAN M., Sélection et validation statistique de variables et de prototypes, SEBBAN M., VENTURINI G., Eds., *Apprentissage automatique*, Hermès Science, 1999.