Méthode de regroupement par graphe de voisinage

Fabrice Muhlenbach

Université de Lyon, Université de Saint-Étienne
UMR CNRS 5516, Laboratoire Hubert Curien
18 rue du Professeur Benoît Lauras, 42000 SAINT-ÉTIENNE, FRANCE
fabrice.muhlenbach@univ-st-etienne.fr, http://labh-curien.univ-st-etienne.fr/ muhlenbach/

Résumé. Ce travail s'inscrit dans la problématique de l'apprentissage non supervisé. Dans ce cadre se retrouvent les méthodes de classification automatique non paramétriques qui reposent sur l'hypothèse que plus des individus sont proches dans l'espace de représentation, plus ils ont de chances de faire partie de la même classe. Cet article propose une nouvelle méthode de ce type qui considère la proximité à travers la structure fournie par un graphe de voisinage.

1 Introduction : classification et graphes de voisinage

Une caractéristique humaine fondamentale est la capacité que nous avons à organiser notre monde, à parvenir à effectuer des catégorisations. Ce phénomène correspond à la faculté de pouvoir regrouper dans de mêmes ensembles (c.-à-d. des classes homogènes) des éléments ayant des traits en commun. Reprise dans le domaine du traitement automatisé de l'information, cette caractéristique englobe les méthodes de classification non supervisée, appelées aussi méthodes de *clustering* (Cleuziou, 2004), une famille de méthodes d'apprentissage automatique qui, à partir des informations connues sur les données, cherchent à retrouver des groupes, à définir des amas, à construire des classes. La classification non supervisée donne lieu à de multiples applications dans le domaine de la fouille de données (en fouille de texte, en bio-informatique, dans le domaine du marketing, en vision par ordinateur, etc.)

Suivant la connaissance existant sur les données, différentes familles de méthodes de classification non supervisée pourront s'appliquer. Dans le cas où il existe a priori une hypothèse sur la distribution des données, il est possible d'employer les méthodes dites « probabilistes » (comme EM). Cependant, en absence de ce genre de connaissance, il faut se limiter aux méthodes dites « non paramétriques » qui reposent sur la seule hypothèse que plus deux individus sont proches, plus ils ont de chances de faire partie du même groupe, de la même classe. Dans ce second cas, nous distinguons principalement trois approches.

Les méthodes de la première approche produisent un partitionnement des données qui sera retenu, parmi les différents regroupements possibles, au moyen d'un critère de qualité donné. Citons parmi ces dernières la méthode des k-Means (Steinhaus, 1956) qui a pour objectif de détecter les différents groupes obtenus à partir d'une partition initiale aléatoire par la recherche des points moyens qui vont minimiser la variance intra-classe de ces différents groupes.

Les méthodes de la deuxième approche produisent une hiérarchie sur les données représentée par un arbre (appelé « dendogramme »). Dans le cas de la classification hiérarchique

ascendante, chaque individu est considéré comme étant une classe en tant que telle (il y a ainsi au départ autant de classes que d'individus), et les classes qui sont considérées comme proches sont progressivement fusionnées suivant un processus itératif qui aboutit jusqu'à la fusion totale des individus en une classe unique. Dans l'approche hiérarchique descendante, l'opération inverse est effectuée : ce sont les sous-ensembles d'une classe qui sont considérés comme trop éloignés (c'est-à-dire trop peu proches) qui sont séparés en deux classes différentes.

Une troisième approche, parfois rapprochée de la deuxième dans la littérature (Jain et Dubes, 1988), permet de réaliser des regroupements à travers un formalisme de représentation basé sur les graphes. Ainsi, dans l'algorithme CURE (Guha et al., 1998), un graphe de type *k*d-tree est employé pour retrouver la forme des différents amas en partitionnant l'espace des données. Une méthode plus ancienne (Zahn, 1971) exploite les propriétés de l'arbre recouvrant minimal, un graphe de voisinage, afin de retrouver des groupes en s'inspirant d'une approche de type psychologie gestaltiste.

Les graphes de voisinage précédemment cités sont des outils issus de la géométrie computationnelle qui ont été à de nombreuses reprises appliqués dans le domaine de l'apprentissage automatique. Par définition, un graphe de voisinage associé à un ensemble Ω de n individus décrits par un vecteur $X = \{x_1, \ldots, x_i, \ldots, x_p\}$ est un graphe dont les sommets sont les différents individus $\omega_1, \ldots, \omega_n$ composant l'ensemble Ω. Notons que nous appellerons aussi « points » les n individus de Ω puisqu'ils sont projetés dans un espace de représentation $\mathbb{R}^p$. Dans un tel graphe, deux individus α et β sont reliés par une arête lorsqu'ils sont voisins au sens d'une structure de voisinage à définir. Cette structure de voisinage peut par exemple être les k plus proches voisins, l'arbre recouvrant minimal, le graphe de Gabriel, la structure des voisins relatifs, les polyhèdres de Delaunay, etc. Dans la suite de cet article, nous considérerons plus particulièrement le graphe des voisins relatifs (Toussaint, 1980) qui est un graphe connexe dans lequel, si deux points sont reliés par une arête, alors ils vérifient la propriété suivante : $d(\alpha, \beta) \leq \max\left(d\left(\alpha, \gamma\right), d\left(\beta, \gamma\right)\right) \forall \gamma, \gamma \neq \alpha, \beta$, où $d(\alpha, \beta)$ est la distance euclidienne (ou une autre distance à définir) entre deux points α et β dans $\mathbb{R}^p$, avec $\alpha, \beta, \gamma \in \Omega$.

En apprentissage automatique supervisé, les graphes des voisins relatifs ont déjà été employés, et ceci avec un certain succès (Muhlenbach et Rakotomalala, 2002; Zighed et al., 2005; Toussaint, 2005). Dans la plupart de ces travaux, l'information sur la variable à prédire Y est utilisée afin de distinguer les arêtes du graphe de voisinage reliant des individus de la même classe de celles reliant des individus de classes différentes. En l'absence de Y, ces notions associées aux arêtes du graphe ne peuvent exister, ce qui peut expliquer pourquoi l'usage des graphes de voisinage dans le domaine de l'apprentissage automatique non supervisé est beaucoup plus marginal. Nous allons néanmoins indiquer dans la section suivante comment un graphe de voisinage peut servir à réaliser un partitionnement des données.

2 Algorithme de regroupement par graphe de voisinage

2.1 Motivation de la méthode

Dans les travaux énoncés précédemment dans le domaine de l'apprentissage automatique supervisé, les auteurs ont observé que les graphes de voisinage – et les graphes de voisins relatifs en particulier – rendent bien compte de la dispersion d'un ensemble d'individus au sein de l'espace de représentation. Pour des méthodes de classification non paramétriques, en

l'absence de toute autre information sur le comportement des données, la notion qui devient prépondérante est la détermination de la *proximité*, or cette proximité peut justement être retrouvée au moyen d'un graphe de voisinage comme le graphe des voisins relatifs qui, comparé à l'arbre recouvrant minimal, comporte plus d'arêtes, ce qui le rend plus « robuste ».

2.2 Description de l'algorithme

La méthode de regroupement par graphe de voisinage procède selon les étapes suivantes :
1. construction du graphe des voisins relatifs (Toussaint, 1980) ;
2. tri des arêtes du graphe par ordre décroissant ;
3. suppression de la plus grande arête du graphe (ou du sous-graphe) ;
4. vérification que le graphe (ou sous-graphe) obtenu est toujours connexe ;
5. création de nouveaux amas s'il y a eu perte de connexité ;
6. retour à l'étape [3] en prenant l'arête suivante jusqu'à satisfaire un critère d'arrêt donné ;
7. les classes sont obtenues à partir des k sous-graphes.

Dans cet algorithme, le processus peut s'arrêter à l'étape [6] quand le nombre k d'amas a été trouvé, comme pour la méthode des k-Means. Pour cela, il est nécessaire d'indiquer au préalable un nombre de groupes désiré, ce qui n'est pas toujours facile à déterminer, mais nous montrons dans la suite de cet article que d'autres critères peuvent être appliqués.

Notons que la complexité algorithmique de la méthode est bornée par la construction du graphe des voisins relatifs qui s'effectue en $O(n^3)$, la construction de la liste des arêtes, le tri de cette liste ou le test de la connexité d'un graphe s'effectuant dans une complexité moindre.

3 Résultats expérimentaux et discussion

Nous illustrons le principe de fonctionnement de notre méthode de regroupement sur une structure composée d'éléments imbriqués l'un dans l'autre. Il s'agit d'une base comportant un ensemble de 2000 points (les coordonnées des pixels noirs) extraits d'une image binaire du symbole chinois *Yin* et *Yang*. Sur la figure 1, nous présentons, de gauche à droite, les données en *Yin* et *Yang*, le graphe des voisins relatifs appliqué aux données, et les 2 classes obtenues par la méthode de regroupement présentée après suppression de 4 arêtes. Les résultats obtenus sur cet exemple sont repris dans le tableau 1 qui indique les tailles des $\#i = \{1 \dots 4\}$ premiers amas trouvés en fonction du nombre d'amas obtenus noté k, du nombre d'arêtes supprimées dans le graphe pour passer de l'état $k-1$ à k et noté a_sp, de la somme des longueurs d'arêtes supprimées (en pourcentage de l'ensemble des longueurs d'arêtes du graphe) notée $\Sigma(l)$ et de la longueur moyenne d'arêtes supprimées et notée μ_l (exprimée en pourcentage).

La figure 2 présente le même type de résultats obtenus pour une base de données artificielle de 402 individus constituée de 2 amas de 200 points et de 2 points isolés (des *outliers*). La méthode retrouve bien (partie droite de la figure) les 2 amas et les *outliers*. Le tableau 2 reprend les résultats obtenus par la méthode sur ces données.

Sur les deux tableaux, nous remarquons que pour passer d'un état à $k-1$ amas à un état à k amas, afin d'arriver au nombre d'amas pertinent (2 pour la figure 1, 4 pour la figure 2), il existe une différence importante dans la longueur moyenne des arêtes supprimées (colonne μ_l). Dans un ensemble de données générées de manière aléatoire (résultats non présentés ici), ce saut

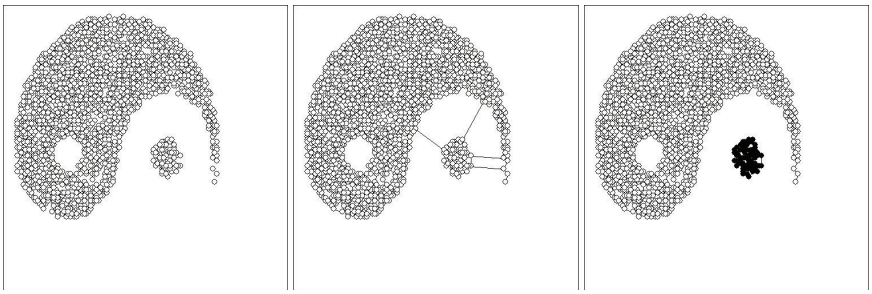

FIG. 1 – *Résultats sur les données Yin et Yang.*

k	a_sp	$\Sigma(l)$	μ_l	#1	#2	#3	#4
1	0	0	0	2000			
2	4	15,86	3,96	73	1927		
3	3	0,68	0,23	73	3	1924	
4	7	1,08	0,15	73	1924	1	2

TAB. 1 – *Taille des #$i = \{1 \dots 4\}$ amas obtenus sur la base Yin et Yang en fonction du nombre d'amas obtenus (k), des arêtes supprimées (a_sp) pour passer de l'état $k-1$ à k, de la somme des longueurs d'arêtes supprimées ($\Sigma(l)$) et de la longueur moyenne de ces arêtes (μ_l).*

quantitatif n'existe pas. En effet, pour obtenir un nouvel amas, un faible nombre d'arêtes coupées associé à une grande longueur d'arêtes (ou somme de longueur d'arêtes) est une indication pertinente pour considérer que cette nouvelle séparation est appropriée.

Même si cette méthode de regroupement par graphe des voisins relatifs est limitée au cas du *clustering* sans recouvrement, ses points forts sont ses capacités à (1) fonctionner pour des tailles d'amas très différentes, (2) ne pas nécessiter de phase d'initialisation préalable (contrairement à des méthodes comme les *k*-Means), (3) ne pas être sensible à la différence de densité des amas, (4) parvenir à retrouver des amas non sphériques (puisque le graphe parvient à « coller » à la forme des données) et (5) être robuste à la présence d'*outliers* (détectés comme des amas qu'il est pertinent de séparer du reste des donnés mais qui ne sont constitués que d'un seul individu). En outre, cette méthode est assimilable à une classification hiérarchique descendante (les tableaux présentés pouvant aisément se traduire en dendogramme) pour une complexité algorithmique bien moindre : dans un tel algorithme divisif, pour un amas initial de taille n, il y a $2^{n-1} - 1$ possibilités de le diviser en 2 sous-amas, soit un coût algorithmique prohibitif.

4 Conclusion et perspectives

La méthode de regroupement par graphe de voisinage que nous proposons s'avère être, au final, une méthode qui présente un certain nombre d'avantages comparées aux autres méthodes de classification automatique non paramétriques. Non sensible à une quelconque initialisation

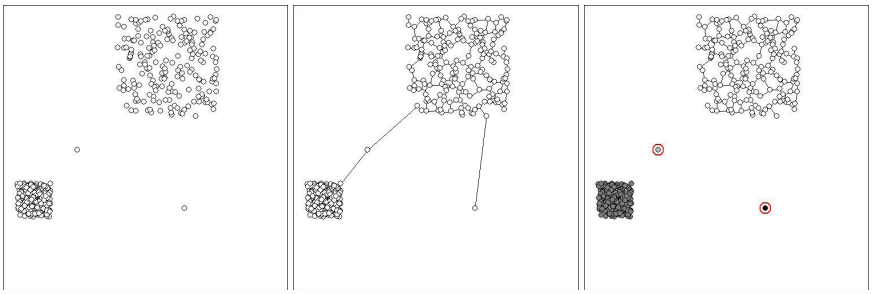

FIG. 2 – *Résultats sur une base artificielle de 402 individus : 2 amas et 2 outliers.*

k	a_sp	$\Sigma(l)$	μ_l	#1	#2	#3	#4	#5
1	0	0	0	402				
2	1	33,58	33,58	401	1			
3	1	16,95	16,95	1	200	201		
4	1	7,27	7,27	1	200	200	1	
5	6	3,92	0,65	1	200	1	199	1

TAB. 2 – *Taille des 5 premiers amas obtenus sur la base des 402 individus en fonction du nombre d'amas obtenus (k), des arêtes supprimées (a_sp) pour passer de l'état $k-1$ à k, de la somme des longueurs d'arêtes supprimées et de la longueur moyenne des arêtes supprimées.*

et pouvant s'appliquer au problème d'imbrication de classes, notre méthode de regroupement, comparée aux méthodes de classification hiérarchique, reste d'une complexité équivalente à l'approche ascendante même si elle procède d'une façon plus proche de la classification hiérarchique descendante alors que cette dernière demande un temps de calcul trop important sur des données nombreuses. De ce fait, notre méthode cumule l'ensemble des avantages des deux classifications hiérarchiques : à travers le graphe de voisinage, une approche descendante devient possible en un temps raisonnable, or celle-ci présente l'intérêt de fournir des typologies dont l'interprétation est plus claire que celles produites par les méthodes ascendantes. En effet, pour la perception visuelle humaine, adaptée surtout à la représentation plane (malgré la vision stéréoscopique), il est plus facile de retrouver les grandes distances séparant de gros ensembles de données (approche descendante) que d'apparier dans de mêmes ensembles des données qui sont considérées comme proches (approche ascendante).

Enfin, en plus d'étudier la piste de la détection des *outliers* pour laquelle nous n'avons encore fait que des hypothèses et d'aborder la problématique de l'évaluation en *clustering* afin de valider objectivement l'apport de cette méthode, nous avons l'intention de poursuivre nos travaux dans diverses voies qui nous permettront de résoudre quelques problèmes associés à notre méthode de regroupement. Au premier plan, celui qui nous paraît critique pour le passage à l'échelle – opération essentielle pour un processus d'extraction de connaissances – est le frein que constitue la complexité d'ordre cubique, celle-ci étant associée à la construction du graphe des voisins relatifs. Plusieurs pistes nous semblent envisageables, telles que la construction du

graphe à partir d'un sous-échantillon représentatif des données suivi d'un processus d'ajout local, à la suite des travaux entrepris par Hacid et Zighed (2005) par exemple.

Remerciements

Ce travail a été réalisé avec les soutiens du projet SATTIC (*Strings and Trees for Thumbnail Images Classification*) de l'Agence Nationale pour la Recherche et du Centre Mutualiste d'Alcoologie de Saint Galmier (nous remercions en particulier le docteur Christian Digonnet, directeur du centre, et Catherine Pons, directrice administrative).

Références

Cleuziou, G. (2004). *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information*. Thèse de doctorat, Université d'Orléans.

Guha, S., R. Rastogi, et K. Shim (1998). CURE : an efficient clustering algorithm for large databases. In *SIGMOD 1998*, pp. 73–84.

Hacid, H. et D. A. Zighed (2005). An effective method for locally neighborhood graphs updating. In *DEXA 2005*, Volume 3588 of *LNCS*, pp. 930–939. Springer.

Jain, A. K. et R. C. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice Hall.

Muhlenbach, F. et R. Rakotomalala (2002). Multivariate supervised discretization, a neighborhood graph approach. In *ICDM 2002*, pp. 314–321. IEEE Computer Society.

Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences C1 III*(4), 801–804.

Toussaint, G. T. (1980). The relative neighbourhood graph of a finite planar set. *Pattern Recognition 12*(4), 261–268.

Toussaint, G. T. (2005). Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *International Journal of Computational Geometry and Applications 15*(2), 101–150.

Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing Gestalt clusters. *IEEE Transactions on Computers C*(20), 68–86.

Zighed, D. A., S. Lallich, et F. Muhlenbach (2005). A statistical approach to class separability. *Applied Stochastic Models in Business and Industry 22*(2), 187–197.

Summary

This work is related to the unsupervised machine learning problem. Some clustering methods, which are part of this research area, are based on the following hypothesis: the more two individuals are close in the representation space, the more they have a chance to belong in the same class. This paper presents a new clustering method that considers the proximity through the structure of a geometric neighbourhood graph.