Distance de compression et classification prétopologique

Vincent Levorato, Thanh Van Le, Michel Lamure, Marc Bui

Distance de compression et classification prétopologique

Vincent Levorato [*] ,Thanh Van Le[*] , Michel Lamure[**] , Marc Bui [*]

[*] *Laboratoire ERIC EPHE-Sorbonne, 41 rue Gay Lussac 75005 Paris*
vincent.levorato@ephe.sorbonne.fr, than-van.le@univ-lyon1.fr, marc.bui@ephe.sorbonne.fr

[**] *Laboratoire ERIC - Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex*
michel.lamure@univ-lyon1.fr

RÉSUMÉ. *Nous présentons dans cet article des algorithmes de classification prétopologique récemment développés, ainsi que l'introduction d'un nouvel algorithme original. Les algorithmes prétopologiques,principalement basés sur la fonction d'adhérence prétopologique et le concept de fermés permettent de classifier des données complexes non représentables dans un espace métrique. Après quelques rappels sur les notions de classification, nous proposons un algorithme exploitant la notion de distance basée sur la complexité de Kolmogorov dans un modèle prétopologique afin d'introduire un indice de similarité. Nous illustrerons celui-ci par une application permettant de saisir le degré de finesse que l'on peut obtenir avec une telle approche.*

MOTS-CLÉS : *Classification, Partitionnement, Prétopologie, Modélisation, Kolmogorov.*

1. Introduction

Le problème de la classification peut être posé sous forme d'une interrogation : étant donné un ensemble d'observations dont on connaît une ou plusieurs caractéristiques, comment peut-on les regrouper en un certain nombre de groupes de manière à ce que les groupes obtenus soient constitués d'observations semblables et que les groupes soient les plus différents possible entre eux ?
Il existe de multiples méthodes de partitionnement des données, comme le regroupement hiérarchique, ou l'algorithme des k-moyennes. Nous allons nous pencher sur quelques méthodes existantes, puis nous nous focaliserons sur les récentes recherches de classification prétopologiques développées par Thanh Van Le [LE 07].
Dans la suite de l'article, nous montrerons comment il est possible d'utiliser la complexité de Kolmogorov dans un modèle prétopologique afin d'introduire un indice de similarité.

2. Classification de données par partitionnement

La classification d'objets par partitionnement en différents groupes, ou si l'on se place dans un contexte mathématique, en sous-ensembles, dépend des caractéristiques que ces objets pourraient avoir en commun, et principalement en effectuant une mesure de la distance qui les sépare. Le partitionnement de données est une technique connue de l'analyse de données en statistique que l'on utilise dans divers domaines que ce soit en fouille de données ou en reconnaissance d'images par exemple. Habituellement, on scinde les méthodes de classification en deux classes [JAR 68] : les méthodes de classification *hiérarchiques* et *non-hiérarchiques*. Mais tout d'abord, il faut poser le problème de la mesure de la distance.

2.1. *Mesure de similarité et distance*

L'objectif de l'analyse de données par partitionnement est de « séparer »un ensemble E composé de n objets en m partitions $\{C_i | i \in I\}$ tel que : $\forall i \in I, \cup C_j = E$ et $\forall (i,j) \in (I \times I), i \neq j, C_i \cap C_j = \varnothing$

Les objets d'une même partition C_i ont une forte similitude entre eux, et peu de similitude avec les objets des autres partitions. Pour mesurer cette similarité ou dissimilarité entre objets, une notion de distance, selon la nature des données, est nécessaire. Si ces objets sont exprimés par des variables numériques tel que l'âge, le nombre de sujets, ..., des distances telles que la distance Euclidienne, la distance de Manhattan (city block), la distance de Chebyshev, la distance de Minkowski, ..., sont utilisées de préférence. Cependant, pour représenter de « simples » distances entre variables de même catégorie comme les couleurs, les familles d'animaux, le sexe, ..., on se tournera vers la distance de Jaccard ou de Hamming par exemple [KAU 90].

Pourtant, dans la réalité, un objet peut être caractérisé par plusieurs types de variables. Dans ce cas, il faut utiliser une distance capable de mesurer la ressemblance entre des objets caractérisés par des variables de différents types. Pour cela, il existe le coefficient de similarité de Gower qui a été inventé pour mesurer la proximité entre données par intervalles, de manière nominale ou binaire. Il y aussi quelques distances proposées par Ralambondrainy [RAL 95] ou par Wilson et Martinez [WIL 97] permettant de traiter des variables de type quantitatif et qualitatif.

La plupart de ces distances requiert un processus de pré-traitement qui calcule la proximité entre chaque objet selon l'indice de similarité ou dissimilarité employé pour chaque variable permettant la comparaison. D'une manière différente, la distance de compression normalisée (Normalized Compression Distance) entre valeurs de différentes natures peut être calculée en se basant sur la complexité de Kolmogorov [CIL 05] :

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}$$

où $C(x)$ représente la taille binaire de x compressé, $C(y)$ représente la taille binaire de y compressé, et $C(xy)$ représente la taille binaire de la concaténation de x et y, le tout compressé. C peut utiliser des algorithmes de compression divers sans pertes comme l'algorithme GZip de Ziv et Lempelon ou BZip de Burrows et Wheeler.

Ces distances sont employées dans des méthodes de classification afin de mesurer la ressemblance des objets d'un ensemble. Le choix de ces algorithmes dépend des types de données, mais aussi de l'utilisation que l'on veut en faire. Nous pouvons résumer deux types de méthodes de classification de données : les méthodes de partitionnement hiérarchique soit construisent des amas d'objets similaires (ascendante), soit divisent un ensemble d'objets en sous-ensembles (descendante) ; et les méthodes de partitionnement non-hiérarchiques qui assignent chaque objet au sous-ensemble dont le centre est le plus proche de l'objet en question.

3. Méthodes de classification prétopologiques

3.1. *Rappels des concepts prétopologiques*

La théorie de la prétopologie est un outil mathématique de modélisation du concept de proximité. Celle-ci permet des opérations sur les ensembles telles que l'adhérence ou la fermeture [BEL 93]. Une application d'adhérence $a(.)$ de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ est appelée *adhérence* ssi $\forall A \in \mathcal{P}(E) : a(\emptyset) = \emptyset$ et $A \subseteq a(A)$. On définit soit les couples (E, a) ou soit le triplet (E, i, a) comme étant des *espaces prétopologiques*. A partir de ces propriétés, on obtient des espaces prétopologiques plus ou moins complexes, du plus général jusqu'à l'espace topologique. Les espaces les plus intéressants appartiennent au type $\mathcal{V}$, ($\forall A \in \mathcal{P}(E), \forall B \in \mathcal{P}(E), A \subseteq B \Rightarrow a(A) \subseteq a(B)$) lesquels nous utiliserons dans la suite de l'article. Le processus de dilatation généré par l'adhérence s'arrête à un instant donné et n'évolue plus. Dans ce cas, on a $a^{k+1}(A) = a^k(A)$ avec $k \in \mathbb{N}$. On nomme A comme étant un sous ensemble *fermé*.

On appelle fermé élémentaire et on note F_x la fermeture d'un singleton $\{x\}$ de E. On note $\mathcal{F}_e$, l'ensemble des fermés élémentaires de E tel que : $\mathcal{F}e(E, a) = \{F_x, x \in E\}$.

On appelle fermé minimal de E, tout élément de $\mathcal{F}_e$, minimal au sens de l'inclusion. L'ensemble des fermés minimaux est noté : $\mathcal{F}_m$.

3.2. *Méthodes MCP et MCPR*

Ces algorithmes ont été développés par Michel Lamure et Thanh Van Le, et apparaissent dans la thèse de cette dernière [LE 07], dont une écriture en pseudo-code a été faite dans [LEV 08].

L'algorithme MCPR se base l'algorithme des *k-moyennes*. Il fournit la possibilité de classifier des objets dans k classes par les k centres choisis a priori et qui sont employés comme germes afin de déterminer les k classes. Cependant, une des limites de l'algorithme *k-moyennes* est que le nombre de classes doit être prédéterminé et le procédé de choix des centres initiaux doit être effectué d'une certaine manière.

La méthode de classification MCP se fait selon deux processus :

1. *Processus de structuration :* il ressemble à celui de MCPR. Il consiste à chercher les familles des fermés et à déterminer des noyaux initiaux pour tous les fermés minimaux.

2. *Processus de partitionnement :* pour chaque élément qui n'est pas encore classé, nous choisissons une classe dans laquelle cet élément sera assigné de telle façon que deux contraintes ci-dessous soient satisfaites à la fois : la distance entre cet élément et le noyau du groupe est la plus proche ; il y a des liens entre cet élément et tous les autres du groupe.

4. Application de MCP grâce à la complexité de Kolmogorov

Dans cette partie, nous allons utiliser la méthode de classification prétopologique (MCP) en se basant sur un indice de similarité, lequel sera calculé à partir de la description élémentaire des singletons de l'espace grâce la complexité de Kolmogorov.

Voici comment est construit le modèle sur lequel nous voulons appliquer MCP : nous avons un ensemble d'auteurs, ainsi que l'intitulé des articles que ceux-ci ont écrits. L'objectif du modèle est de retrouver les auteurs en relation en fonction du titre de l'article qu'ils ont écrits, donc du sujet qu'ils ont abordé, puis de les classer selon leur domaine. Chaque auteur correspond au moins un article, et pour chaque article correspond au moins un auteur. Pour connaître la "$distance$" entre chaque article, on calcule toutes les distances de compression normalisées pour chaque paire d'articles en prenant comme mot à compresser la correspondance binaire de leur titre (pré-traitement). Soient E et F deux ensembles non vides.

(1) On appelle multiapplication ou correspondance de E dans F toute application de E dans $\mathcal{P}(F)$, ensemble des parties de F. Si X est une multiapplication de E dans F on notera :$X \ : E \rightarrow \mathcal{P}(F)$ dans ce cas, pour tout $x \in E$, $X(x)$ est une partie de F. (2) Soit X une multiapplication de E dans F. Si $\forall x \in E$, $X(x) \neq \emptyset$, on dit que X est une multiapplication à valeurs non vide ; si $\forall x \in E$, $X(x)$ est constante on dit que X est une multiapplication constante. (3) Soit X une multiapplication de E dans F, et A une partie de E ; on appelle image de A par X la partie de F notée $X(A)$ définie par : $X(A) = \bigcup_{x \in A} X(x)$.

On définit l'ensemble E comme l'ensemble des auteurs et l'ensemble G comme l'ensemble des articles. Comment savoir si deux articles sont "$proches$" ou pas ? En utilisant la définition de la distance de compression normalisée, on construit un espace prétopologique avec des relations valuées (à la manière d'un graphe complet), ces relations représentant la distance entre deux articles (calculée par la distance de compression normalisée appliquée au titre de l'article). Ainsi, les articles qui contiennent des mots similaires auront une distance qui les séparent plus faible que des articles qui ne possèdent aucun terme en commun, même s'il existe un risque de biais. Voici comment on le formalise :

$$Soit \ \ G \times G \rightarrow \mathbb{R}_+^*, (x, y) \rightarrow v(x, y)$$

$$\forall B \in \mathcal{P}(G), s \in \mathbb{R}_+^*, a_G(B) = \{y \in G - B, \sum_{x \in B} v(x, y) \leq s\} \cup B$$

L'application adhérence $a_G(.)$ nous permet d'exprimer la proximité entre articles. Le couple (G, a_g) correspond donc à l'espace prétopologique des articles. Nous allons nous servir de cette adhérence pour construire celle de l'espace prétopologique des auteurs. Définissons tout d'abord les deux multiapplications nous permettant de passer de E à G et de G à E :

Soit f une multiapplication de E dans G, A une partie de E, on définit l'image de A par la multiapplication f la partie de G notée $f(A)$ par $f(A) = \{y \in G; x \in E, y = f(x)\}$. Soit g une multiapplication G dans E, B une partie de G, on définit l'image de B par la multiapplication g la partie de E notée $g(B)$ par $g(B) = \{y \in E; x \in G, y = g(x)\}$.

En somme, f nous permet de déterminer quels articles ont été écrits par un ensemble d'auteurs et g quels auteurs ont écrits un ensemble d'articles. Nous pouvons à présent construire l'adhérence sur E :

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E, g(a_G(f(A)))\}$$

Le couple (E, a) correspond à l'espace prétopologique qui nous intéresse, c'est-à-dire celui portant sur les auteurs. On peut remarquer que celui-ci est de type $\mathcal{V}$.

On peut constater par cet exemple de modélisation que la prétopologie peut aisément concevoir des représentations de réseaux complexes. Les méthodes de classifications vues plus haut s'intègrent parfaitement dans ce type de modèle, étant basé sur l'adhérence.

5. Conclusion

Dans cet article, on a pu constater que des méthodes de classification pouvaient être utilisées dans les espaces prétopologiques, et qu'elles pouvaient être appliquées sur tout type de données. Etant basées sur l'adhérence comme la plupart des algorithmes prétopologiques car définissant la proximité, on a montré qu'il était possible d'utiliser la définition de cette adhérence pour définir différentes sortes de similarité, permettant de donner des résultats de classification différents selon le modèle étudié, faisant de la prétopologie un outil complet pour construire des modèles au comportement complexe. La modélisation d'auteurs présentée comme un cas particulier peut être utilisée de manière générique pour d'autres types de problèmes. Une des perspectives envisageables serait une comparaison de l'aspect théorique à l'aspect qualitatif d'une telle modélisation, ouvrant une direction vers des recherches intéressantes.

6. Bibliographie

[BEL 93] BELMANDT Z., *Manuel de prétopologie et ses applications : Sciences humaines et sociales, réseaux, jeux, reconnaissance des formes, processus et modèles, classification, imagerie, mathématiques*, Hermes Sciences Publications, 1993.

[CIL 05] CILIBRASI R., VITANYI P., Clustering by compression, *IEEE Transactions on information theory*, vol. 51, n^o 4, 2005.

[JAR 68] JARDINE N., SIBSON R., The construction of hierarchic and non-hierarchic classifications, *The Computer Journal*, vol. 11, n^o 2, 1968.

[KAU 90] KAUFMAN L., ROUSSEEUW P. J., *Finding groups in data : An introduction to cluster analysis*, WILEY-Interscience, 1990.

[LE 07] LE T. V., Classification prétopologique des données. Application à l'analyse des trajectoires patients, PhD thesis, Université Lyon 1, 2007.

[LEV 08] LEVORATO V., Contributions à la Modélisation des Réseaux Complexes : Prétopologie et Applications, PhD thesis, Université de Paris 8, 2008.

[RAL 95] RALAMBONDRAINY H., A conceptual version of the k-means algorithm, *Pattern Recognition Letters*, vol. 16, Elsevier Science, 1995, p. 1147-1157.

[WIL 97] WILSON D. R., MARTINEZ T. R., Improved heterogenous distance function, *Journal of Artificial Intelligence Researche*, vol. 6, 1997.