Classification automatique de données composites

A. Ketterlin
C.-él. : alain@dpt-info.u-strasbg.fr
LSIIT, URA CNRS 1871, Université Louis Pasteur
7, rue René Descartes, F-67084 Strasbourg Cedex

Attention : cet article est une version préliminaire.
Merci de ne pas le citer sans prendre contact avec l'auteur.

Résumé : ce résumé décrit un système de classification capable de traiter des données composites, c'est-à-dire des objets dont la description comporte d'autres objets, plus simples. Il est pour cela nécessaire de définir un algorithme de classification, ainsi qu'un procédé de généralisation de tels objets. Le résumé comporte également la description succincte d'une application dans laquelle la classification de données composites s'est révélée fructueuse, et se termine en décrivant quelques propriétés de la structure de la classification produite, qui la rapprochent de formalismes complexes de représentation des connaissances. Ces propriétés font d'un algorithme de classification un véritable système de construction automatique de bases de connaissances.
Mots-clés : classification de données composites, généralisation, construction de bases de connaissances, applications.

1 Introduction

Cet article porte sur la classification automatique d'objets composites, à savoir la répartition d'un ensemble de données en classes, selon un critère d'homogénéïté de ces classes. Des objets composites sont des données dont la forme ne se restreint pas à un vecteur de caractéristiques, mais se compose d'un ensemble de « sous-données », qui peuvent elles mêmes éventuellement être composites. On rencontre fréquemment de tels types de données dans la modélisation de bases de données, et leur gestion est donc particulièrement importante dans une optique de découverte de connaissances, ou *fouille de données*.

La classification automatique a pour but de former des classes à partir d'un ensemble d'individus [6]. Cet objectif s'accompagne dans la pratique d'un critère d'association d'un individu à une classe, qui traduit dans la plupart des cas une similarité entre l'individu et les autres membres de la classe. Si ce problème a été très étudié dans un contexte statistique [3], il l'a été également en apprentissage automatique [10, 1]. De plus, on le retrouve implicitement dans certaines méthodes de modélisation de données, principalement depuis l'apparition de formalismes complexes (par exemple orientés objets). Ces développements permettent d'envisager pour la classification une portée peut-être plus importante que celle que lui réservent les méthodes les plus anciennes. On peut en effet envisager le processus global comme une véritable méthode d'acquisition empirique de connaissances, qui permet de construire des classes (des *symboles*) à partir de leurs manifestations. C'est donc bien d'un processus de généralisation dont il s'agit.

De ce point de vue, il est important que la classification s'applique dans un grande variété de situations. Les premières avancées dans ce domaine portent essentiellement sur des données qualitatives et/ou quantitatives, la plupart du temps sur des domaines multivariés. Le résultat des algorithmes statistiques reste cependant souvent une subdivision de la population initiale (sous forme de partitionnement ou de hiérarchie), sans que les membres de cette subdivision soient qualifiés (c'est-à-dire décrits autrement que par leurs représentants). L'apprentissage automatique met quant à lui l'accent sur la capacité de qualifier les classes découvertes, en vue de les intégrer à une base de connaissances, et donc de les réutiliser.

Cet article décrit un algorithme capable de construire une hiérarchie de classes à partir d'une population d'objets composites, en fournissant pour chaque classe une description intensionnelle : un tel algorithme est brièvement décrit dans la section 2. Les objets composites sont définis de la façon suivante :

– un objet est une liste de couples attribut-valeurs ;
– un attribut est soit *primitif* (à savoir qualitatif ou quantitatif) soit *structuré*, auquel cas la valeur correspondante est un ensemble d'objets.

Cette dernière partie de la définition introduit une structuration hiérarchique des objets.

Évidemment, admettre de telles données nous oblige à définir la représentation d'une classe. Pour aider l'intuition, on peut considérer que les objets apparaissent à différents niveaux d'*abstraction* : les objets les plus abstraits « contiennent » les moins abstraits. Dans la section 3, nous décrivons un exemple mettant en jeu des objets pris à deux niveaux d'abstraction. La section 4 détaille le formalisme de représentation des classes construites.

2 Aspects algorithmiques

Notre approche se base sur un algorithme générique de construction de hiérarchies de classes à partir d'une population [4]. Cet algorithme procède de façon incrémentale, en considérant l'un après l'autre les individus de la population. Pour chacun d'eux, il l'incorpore dans la hiérarchie à partir de la racine, en le dirigeant progressivement vers les profondeurs de la subdivision. Il se base pour ce faire sur les classes existantes du niveau auquel il opère, et sur la description de l'individu « entrant ».

À chaque niveau, l'algorithme doit déterminer vers quelle sous-classe le nouvel individu sera dirigé. Il va pour cela considérer diverses variantes de la partition du niveau courant (constituée des sous-classes immédiates de la classe courante), et évaluer diverses variations de cette partition incluant le nouvel individu. Parmi ces variations, l'algorithme choisit celle qui lui semble la « meilleure ». Cette sélection se base bien sûr sur un critère objectif, que nous n'aurons pas la place de détailler ici.

On peut donc résumer l'action de cet algorithme à trois opérations de base :

– *former* des classes : cette opération consiste à constituer un ensemble d'individus en une classe. C'est la structure algorithmique qui décide de l'ensemble des classes potentielles : dans le cas qui nous occupe, l'algorithme emploie une recherche par gradient dans un espace partitions potentielles. Nous ne détaillerons pas plus cet aspect ;
– *décrire* les classes : chaque classe constituée se doit d'être décrite intensionnellement. C'est une condition que nous avons admise pour justifier l'objectif d'acquisition de connaissances. Dans le cas d'attributs primitifs, une telle description est en général assez simple : paramètres d'une loi de probabilité hypothétique (par exemple moyenne et variance pour un attribut supposé normal) dans le cas quantitatif, table de fréquences des valeurs pour un attribut qualitatif. Dans le cas d'attributs composites, une telle description sera plus complexe : les paragraphes suivants y sont consacrés ;
– *évaluer* les classes : afin de décider de la modification à apporter à la classification au vu d'un nouvel individu, l'algorithme doit choisir parmi plusieurs variations potentielles. Pour ce faire, il doit disposer d'une fonction objective. Cette fonction, dans notre cas, se base uniquement sur la description des classes intervenant dans une partition à un niveau de la hiérarchie. Elle sert à la décision, c'est-à-dire au choix de la variation. Par exemple, dans le cas d'attributs quantitatifs, la variance observée peut servir a évaluer la « précision » d'une classe. Nous donnerons quelques indications plus bas sur l'évaluation d'une classe d'objets composites.

Le détail de l'algorithme ne peut être donné ici [5]. Nous nous attacherons dans le suite de cette section à détailler la deuxième opération (*décrire* les classes) dans le cas où les individus sont composites.

Avant de passer au processus de généralisation d'un ensemble d'objets composites, il nous faut décrire la stratégie générale de classification. En effet, les individus n'étant plus décrits de façon « atomique », les données apparaissent à différents niveaux d'abstraction. Prenons pour exemple un paysage dans une image de télédétection (nous reviendrons sur cet exemple dans la section 3). Un *paysage* est constitué d'un ensemble de *régions*, une région étant une zone connexe de l'image dont la « couleur » est homogène. La population est donc composée de deux types d'individus de nature différentes : les paysages et les régions, les premiers étant composites (nous pouvons considérer ici que les seconds sont décrits par des attributs primitifs uniquement). La classification ayant pour objectif de construire des classes de paysages, nous admettrons que cette tâche est vaine si l'on ne dispose pas *au préalable* de classes de régions. La stratégie globale de classification est donc de construire des classes au niveau d'abstraction le plus bas (les régions dans notre exemple) avant de construire des classes au niveau le plus haut (les paysages).

Considérons maintenant le problème de la généralisation d'objets composites. Nous avons vu que l'algorithme avait besoin de pouvoir décrire les classes qu'il construit. Nous supposerons dans la suite que chaque individu est décrit par un seul attribut, dont la valeur est un ensemble d'individus (cette hypothèse simplifie l'exposition sans impliquer de perte de généralité). Le problème est donc le suivant : étant donnés N individus (c'est-à-dire N ensembles), déterminer une description (généralisation) de ces N ensembles. Cette généralisation doit permettre de décider de l'appartenance d'un individu à la classe ainsi décrite.

Avant de donner une définition formelle de la généralisation d'ensembles, considérons la remarque suivante : nous avons dit plus haut que la classification opérait « de bas en haut » dans les différents niveaux d'abstraction. Cela signifie que les classes d'individus existent pour les composants au moment où l'algorithme construit des classes de composites. Cela signifie en particulier que l'algorithme peut s'affranchir des descriptions exactes des composants, et les remplacer par des noms de classes de composants.

L'opération de généralisation est alors définie de la façon suivante : soit N ensembles $P_i = \{R_1^i, \ldots, R_{N_i}^i\}$ (nous utilisons ici P – comme paysage – pour les individus composites et R – comme région – pour les composants). La première phase est d'identifier les classes de chaque composant : c'est immédiat, et cela nous permet de réécrire chaque composite sous la forme d'un ensemble de noms de classes de composants $P_i = \{\mathcal{R}_1^i, \ldots, \mathcal{R}_{N_i}^i\}$. On peut ensuite définir la généralisation G de ces individus composites comme étant un ensemble de noms de classes de composants $\{\mathcal{R}_1, \ldots, \mathcal{R}_L\}$ tel que :

1. chaque élément de cette description est effectivement représenté dans chaque individu :

$$\forall i \in [1,N]\ \forall \ell \in [1,L]\ \exists j \in [1,N_i] \text{ tel que } \mathcal{R}_j^i \sqsubseteq \mathcal{R}_\ell$$

   où $\mathcal{R}_1 \sqsubseteq \mathcal{R}_2$ signifie que $\mathcal{R}_1$ est sous-classe (au sens large) de $\mathcal{R}_2$;
2. chaque élément de chaque individu est représenté dans la généralisation :

$$\forall i \in [1,N]\ \forall j \in [1,N_i]\ \exists \ell \in [1,L] \text{ tel que } \mathcal{R}_j^i \sqsubseteq \mathcal{R}_\ell$$

3. les éléments de G sont les plus *spécifiques* possibles :

$$\begin{aligned} \forall \ell \in [1,L] \qquad & \nexists \mathcal{R}_* \in \mathcal{H}_R \text{ tel que} \\ & \mathcal{R}_* \sqsubseteq \mathcal{R}_\ell\ \wedge\ \{\mathcal{R}_1, \ldots, \mathcal{R}_L\} \setminus \{\mathcal{R}_\ell\} \cup \{\mathcal{R}_*\} \text{ vérifie les conditions 1 et 2} \end{aligned}$$

   c'est-à-dire qu'il n'existe pas de classe d'individus qui soit à la fois plus « précise » que celles utilisées et vérifie les mêmes critères.

Cette définition débouche sur un algorithme efficace de calcul d'une généralisation [8]. Ajoutons à cette définition une remarque quant à l'évaluation des descriptions : intuitivement, la généralisation d'un ensemble d'individus composites sera d'autant plus « précise » que les noms de classes qui la composent représentent des classes elles-même « précises ». Une façon grossière de mesurer la précision d'une classe, est par exemple, de mesurer sa profondeur dans la hiérarchie. L'algorithme utilise une approche plus fine, qui n'est pas décrite ici. Toujours est-il que cette définition d'une généralisation, alliée à une quantification de la « qualité » d'une classe, permet à l'algorithme de classification de s'appliquer à une population d'individus composites.

3 Application

Avant de détailler l'organisation des classes produites, cette section décrit brièvement une application dans laquelle l'algorithme de classification d'individus composites s'est révélé particulièrement utile. Ce domaine, déjà évoqué, est celui de l'analyse automatique d'images de télédétection par satellite. Dans une telle image, chaque pixel est décrit par trois valeurs numériques (dans le cas du satellite Spot du moins). On a donc un premier problème de classification, qui consiste à regrouper en classes les pixels de l'image. Le résultat de cette phase est une segmentation hiérarchique de l'image initiale [9]. Il est possible suite à cela de décomposer l'image initiale en régions (ou segments), chaque région étant une zone connexe (selon une définition appropriée de la connexité) dont tous les pixels sont de la même classe.

Suite à cela, on peut considérer que chaque région est au centre d'un paysage. On cherche à découvrir des classes de paysages, à savoir des régularités dans la répartition spatiale des régions radiométriquement homogènes. Chaque paysage est un individu composite, dont les éléments sont les régions apparaissant dans le voisinage de la région centrale du paysage (selon une définition appropriée du voisinage). Ces régions sont d'autre part des instances de classes de la hiérarchie construite sur des bases radiométriques. Tous les éléments sont en place pour effectuer une classification d'individus composites.

Nos expériences ont porté sur une image de taille 345 par 410. La classification des pixels a donc porté sur une population de taille 141450, ce qui ne pose pas de problème particulier, principalement du fait que l'algorithme est incrémental : il n'a donc pas besoin de conserver les individus en mémoire. Cette image a été découpée en 46430 régions (d'assez petite taille donc). Les 46430 paysages dérivés comportent entre 2 et 32 régions (avec une moyenne de 14.22 régions par paysage). L'algorithme a construit sur cette base une classification des paysages, qui a été jugée pertinente par un expert géographe [7].

4 Représentation des connaissances

L'application décrite dans la section précédente a servi à valider empiriquement la classification d'individus composites. Étant donnée la complexité des données manipulées et des résultats produits, on peut s'interroger sur la structure du résultat. Dans l'exemple précédent, le système a construit des classes de pixels (donc de régions) et des classes de paysages. Il est intéressant d'examiner de quelle façon ces classes sont reliées. On peut pour cela partir d'une remarque simple : supposons deux classes de paysages $\mathcal{P}$ et $\mathcal{P}'$ telles que $\mathcal{P}' \sqsubseteq \mathcal{P}$ ($\mathcal{P}'$ est sous-classe – directe ou non – de $\mathcal{P}$). On a alors la propriété suivante : les classes utilisées dans la description de $\mathcal{P}'$ sont des sous-classes de celles utilisées dans le description de $\mathcal{P}$.

On peut faire plusieurs remarques sur ce dernier point :

– l'algorithme détermine lui-même le nombre de noms de classes à utiliser dans la description d'une classe d'individus composites. Il effectue donc une forme d'induction constructive : on peut considérer que chaque élément de la description est un nouvel attribut (déterminé automatiquement) ;

– les éléments de la description de deux classes non disjointes ne sont pas indépendants, mais organisés en une hiérarchie.

Ce dernier point est à rapprocher de structures utilisés dans certains formalismes de représentation des connaissances, pour lesquels cette capacité est appelée *différenciation* de descripteurs : si une classe est décrite par un attribut A, une de ses sous-classes peut être décrite par deux – ou plusieurs – attributs A_1 et A_2, de telle façon que les valeurs de A dans la classe « de base » soient réparties entre A_1 et A_2 dans la sous-classe. C'est exactement le phénomène que l'on observe dans la structure de classes créée par l'algorithme dont nous venons de parler. Dans un langage tel que KL-ONE, cette structure est appelée *différenciation de rôle*[2]. Dans d'autres formalismes, tels que les graphes conceptuels, elle apparait sous la forme d'une hiérarchie de *liens*, qui évolue parallèlement à deux portions de la hiérarchie de noeuds. La classification de données composites telle qu'elle a été décrite ici est donc un véritable système de construction automatique de bases de connaissances.

5 Conclusion

On peut tirer plusieurs conclusions de ce qui a été exposé dans ce résumé :

– La classification est un processus général de découverte de connaissances. Elle permet de construire un ensemble de classes qui représentent de façon compacte un grand volume de donnees. La puissance de ce processus n'est pas tant dans la stratégie qu'il utilise que dans la variété des types de données qu'il est capable de manipuler.

– La classification de données composites rapproche les systèmes automatiques des formalismes utilisés dans le cas où la classification est effectué « manuellement ». Le résultat de la classification atteint un degré de complexité proche de celui des formalismes pour lesquels il n'existe pas *a priori* de technique automatique d'acquisition. De tels mécanismes augmentent donc la portée du processus de classification.

On peut donc voir la classification comme une méthode de construction automatique de bases de connaissances.

Références

[1] G. Bisson. *Induction de bases de connaissances en logique des prédicats.* Thèse de doctorat, Université de Paris-Sud, 1993.

[2] R. J. Brachman and J. G. Schmolze. An overview of the KL-ONE representation system. *Cognitive science*, 9 :171–216, 1985.

[3] E. Diday, J. Lemaire, J. Pouget, and F. Testu. *Eléments d'analyse de données.* Dunod, 1982.

[4] J. H. Gennari, P. Langley, and D. H. Fisher. Models of incremental concept formation. *Artificial intelligence*, 40 :11–61, 1989.

[5] A. Ketterlin. *Découverte de concepts structurés dans les bases de données.* Thèse de doctorat, Université Louis Pasteur, Strasbourg I, 1995.

[6] A. Ketterlin. Classification et apprentissage : la formation de concepts structurés. In J.-F. Mari and A. Napoli, editors, *Aspects de la classification*, pages 51–60. Centre de Recherche en Informatique de Nancy, Rapport de recherche 96-R-072, 1996.

[7] A. Ketterlin, D. Blamont, and J. J. Korczak. Unsupervised learning of spatial regularities. In J. Desachy, editor, *Image and signal processing for remote sensing – II*, volume 2579 of *SPIE Proceedings Series*, Paris, France, 1995.

[8] A. Ketterlin, P. Gançarski, and J. J. Korczak. Conceptual clustering in structured databases : a practical approach. In U. M Fayyad and R. Uthurusamy, editors, *Proceedings of the first international conference on knowledge discovery in databases*, Montreal, Canada, 1995. AAAI/MIT Press.

[9] J. J. Korczak, D. Blamont, and A. Ketterlin. Thematic image segmentation by a concept formation algorithm. In J. Desachy, editor, *Image and signal processing for remote sensing*, volume 2315 of *SPIE Proceedings Series*, Rome, Italie, 1994.

[10] R. E. Stepp and R. S. Michalski. Conceptual clustering of structured objects : a goal-oriented approach. *Artificial intelligence*, 28 :43–69, 1986.