Traitement symbolique de contraintes expertes en classification automatique[1]

Fabrice Rossi[2] & Frédérick Vautrain
LISE/CEREMADE (CNRS UMR 7534)
Université Paris-IX/Dauphine,
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16
e-mail : rossi@ceremade.dauphine.fr
e-mail : vautrain@ceremade.dauphine.fr

Résumé

Dans cet article, nous proposons une approche symbolique d'introduction de contraintes dans un problème de classification. Les contraintes représentent le savoir des experts du domaine, et forcent ou interdisent certains regroupements d'individus. Nous montrons qu'il est possible de traiter symboliquement les contraintes afin de tenir compte du caractère éventuellement incomplet du jeu de données étudié et d'englober différentes interprétations possibles du savoir expert. Nous montrons comment le savoir expert peut parfois être utilisé pour découvrir les limites de l'espace de description symbolique qui le représente.

Mots-clés Classification, Contraintes symboliques, Ordre symbolique, Validation symbolique, Structuration de connaissances.

1 Introduction

Dans certaines applications de classification, il est possible de s'appuyer sur le savoir d'experts du domaine étudié. Plus précisément, on peut obtenir un ensemble de contraintes que doit satisfaire toute partition du jeu de données étudié. Ces contraintes prennent deux formes : les contraintes de *fusion* indiquent que certains individus doivent impérativement être dans une même classe, alors que les contraintes d'*exclusion* indiquent au contraire que certains individus ne doivent pas être placés dans une même classe. La prise en compte simultanée de contraintes d'exclusion et de fusion enrichit le cadre habituel de la classification dite "sous contrainte" (cf. [Gordon, 1996]).

Dans cet article, nous nous intéressons de plus au cas de contraintes formulées de façon *symbolique*. Nous supposerons que les experts sont capables de fournir des descriptions symboliques de groupes d'individus et peuvent ainsi donner des contraintes entre groupes. Par symbolique, nous entendons avant tout le fait qu'un groupe sera décrit par des propriétés vérifiées par les individus qui le constituent, plutôt que par la liste de ces individus (définition en intension, cf. [Diday, 1998]).

[1]Publié dans les actes des Septièmes journées de la Société Francophone de Classification. Disponible à http://apiacoa.org/publications/1999/sfc99.pdf

[2]Les coordonnées actuelles de Fabrice Rossi sont disponibles à l'URL http://apiacoa.org/

Nous proposons d'aller plus loin qu'un simple algorithme. Comme nous le verrons dans les sections suivantes, il est relativement simple de modifier les algorithmes classiques de classification, basés sur l'utilisation d'une distance, afin qu'ils prennent en compte les contraintes d'exclusion et de fusion. Cependant, cela nécessite de passer des descriptions symboliques aux individus eux-même. Nous montrons qu'il est possible d'étudier le problème de classification sous contrainte au niveau symbolique, sans revenir aux individus.

Le traitement symbolique est intéressant pour deux raisons : du point de vue pratique, il permet de réduire le coût des algorithmes car les groupes d'individus décrits par les experts sont généralement moins nombreux que les individus eux-même. De plus, le traitement symbolique permet de prendre en compte la variabilité dans l'interprétation des descriptions données par les experts. Plus précisément, il permet de démontrer qu'une classification respectant les contraintes est possible (ou impossible) quelle que soit l'interprétation choisie (parmi une large classe d'interprétations admissibles) pour les descriptions symboliques.

2 Cadre théorique

2.1 Définitions et notations

- l'ensemble $\mathcal{A}$ désigne un *espace de description*. C'est le langage utilisé par les experts, chaque élément de $\mathcal{A}$ décrivant un groupe d'individus.
- un *jeu de données* est constitué d'un ensemble $\mathcal{O}$, sous-ensemble d'une population Ω, dont les éléments sont les individus à classer.
- ext est une fonction de $\mathcal{A}$ dans $\mathcal{P}(\mathcal{O})$, l'ensemble des parties de $\mathcal{O}$. Cette fonction est l'*interprétation* des descriptions symboliques : elle associe à une description symbolique l'ensemble des individus "décrits" par celle-ci. On l'appelle la fonction *extension*.
- F (resp. E) est une partie de $\mathcal{A}^2$, chaque élément de cet ensemble représentant une *contrainte de fusion* (resp. *d'exclusion*) : si $(a, b) \in F$ (resp. $(a, b) \in E$), toute classification de $\mathcal{O}$ doit placer dans une même classe (resp. dans des classes différentes) les éléments de $ext(a)$ et ceux de $ext(b)$.

2.2 Exemple

- Soit $\mathcal{A} = \mathcal{P}^*(Volume) \times \mathcal{P}^*(Densité) \times \mathcal{P}^*(Poids)$ où les ensembles $Volume$, $Densité$ et $Poids$ sont chacun composés des éléments $\{faible, moyen, élevé\}$. $\mathcal{P}^*(A)$ est l'ensemble des parties de A privé de $\emptyset$.
  Pour simplifier on notera $*$ les trois ensembles $Volume$, $Densité$ et $Poids$. Par exemple, $a = (*, \{faible, élevé\}, \{faible\})$ désignera donc la description $(Volume, \{faible, élevé\}, \{faible\})$
- On dispose d'un jeu de données où les individus sont décrits par trois variables quantitatives : Vol définie sur $[0, 100]$ qui indique la valeur du volume d'un individu (en dm^3), $Dens$ définie sur $[0, 1]$ qui indique la valeur de la densité d'un individu (en kg/m^3), Pds définie sur $]0, 150]$ qui indique la valeur du poids d'un individu (en kg).
- On construit une application d'extension dont on donne ici une ébauche :

$$ext\left(a\right) = \begin{cases} \{x \in \Omega \mid Vol\left(x\right) \leq 30\} & \text{si } a = (\{faible\}, *, *) \\ \{x \in \Omega \mid 30 < Vol\left(x\right) \leq 70\} & \text{si } a = (\{moyen\}, *, *) \\ \{x \in \Omega \mid Vol\left(x\right) \leq 70\} & \text{si } a = (\{faible, moyen\}, *, *) \\ \ldots \end{cases}$$

– Un exemple de contrainte de fusion (couple de F) est :
$((*, \{faible, moyen\}, \{faible\}), (*, \{faible, moyen\}, \{moyen\}))$ qui déclare que le groupe des objets de volume quelconque, de densité faible ou moyenne et de poids faible doit appartenir à la même classe que le groupe des objets de volume quelconque, de densité faible ou moyenne et de poids moyen.
Un exemple de contrainte d'exclusion (couple de E) est :
$((*, *, \{élevé\}), (\{faible\}, \{faible\}, *))$ qui déclare que le groupe des objets de volume quelconque, de densité quelconque et de poids élevé ne doit pas appartenir à la même classe que le groupe des objets de volume faible, de densité faible et de poids quelconque.

2.3 Relation compatible

Le *savoir expert* est constitué d'un couple (F, E), représentant les contraintes de fusion et les contraintes d'exclusion. Le problème de classification sous contraintes consiste à trouver une relation d'équivalence sur $\mathcal{O}$ satisfaisante (au sens de la classification automatique) et qui respecte les contraintes imposées par (F, E). On introduit la définition suivante :

Définition 1 *On dit qu'une relation r sur $\mathcal{O}$ est* compatible *avec (F, E) interprété par ext si et seulement si :*

1. *pour tout $(a, b) \in F$, pour tout $x \in ext(a)$ et tout $y \in ext(b)$, $r(x, y)$.*
2. *pour tout $(a, b) \in E$, pour tout $x \in ext(a)$ et tout $y \in ext(b)$, $\neg r(x, y)$.*

3 Solution simple

3.1 Existence d'une relation d'équivalence

Il est relativement facile de résoudre le problème suivant : trouver, si elle existe, la relation d'équivalence la plus fine compatible avec un savoir expert pour un jeu de données et une extension donnés. En effet, il suffit d'appliquer l'algorithme suivant (c'est un cas particulier de celui de [De Guio et al., 1997]) :

Algorithme 1 *Calcul de la relation la plus fine :*

1. *construire le graphe sur $\mathcal{O}$ dont les arêtes sont l'ensemble des (x, y) tels que il existe $(a, b) \in F$ avec $x \in ext(a)$ et $y \in ext(b)$;*
2. *calculer les composantes connexes de ce graphe, soit les ensembles $C_1, \ldots, C_n$;*
3. *définir la relation r sur $\mathcal{O}$ par $r(x, y)$ si et seulement si il existe $1 \leq i \leq n$ tel que $x \in C_i$ et $y \in C_i$;*
4. *vérifier que la relation 2 de la définition 1 est satisfaite :*
   (a) *si c'est le cas, alors est r est la relation d'équivalence la plus fine compatible avec le savoir expert ;*
   (b) *sinon, il n'existe pas de relation d'équivalence sur $\mathcal{O}$ compatible avec (F, E).*

Le résultat de cet algorithme n'est pas une classification. Il se contente d'indiquer s'il est *possible* de construire des relations d'équivalence compatibles. Il s'agit donc simplement d'une étape préalable à une classification.

Le principal défaut de cette approche est que le résultat obtenu dépend à la fois de $\mathcal{O}$ et de ext. De ce fait, il ne caractérise pas le savoir expert **en lui-même**, mais **une** interprétation de celui-ci, sur **un** jeu de données particulier.

3.2 Classification

Il est possible d'adapter les algorithmes de classification basés sur une métrique à notre problème. En effet, on peut considérer la contrainte de fusion comme une mesure de proximité absolue, alors que la contrainte d'exclusion induit au contraire une similarité nulle ([De Guio et al., 1997]).

De ce fait, il est possible de modifier une similarité entre les individus d'un jeu de données pour inclure les contraintes. Si deux individus (x, y) sont tels qu'il existe $(a, b) \in F$ avec $x \in ext(a)$ et $y \in ext(b)$, on posera $s(x, y) = \infty$. Dans le cas des contraintes d'exclusion, si on a $(a, b) \in E$ avec $x \in ext(a)$ et $y \in ext(b)$, on posera $s(x, y) = 0$.

4 Une approche symbolique

4.1 Introduction

Le but de l'approche symbolique est de travailler directement sur les contraintes expertes, afin d'en extraire de l'information, qui sera valable pour *toute* interprétation admissible des descriptions symboliques de groupes d'individus.

4.2 Notion d'ordre

Pour atteindre notre but, il est nécessaire d'ajouter quelques hypothèses au cadre théorique présenté à la section 2.

Définition 2 *Un* espace de description ordonné *est un ensemble ordonné* $(\mathcal{A}, \leq)$ *dans lequel un sous-ensemble totalement ordonné possède au moins un minorant. On note* $\min(B)$ *l'ensemble des minorants d'une partie* B *de* $\mathcal{A}$.

La relation d'ordre sur l'espace de description correspond à la notion de généralité. Dire que $a \leq b$ signifie que a est moins générale que b, ou encore plus précise.

Définition 3 *Soit* $(\mathcal{A}, \leq)$ *un espace de description ordonné et* $\mathcal{O}$ *un jeu de données. On dit qu'une application* ext *de* $(\mathcal{A}, \leq)$ *dans* $(\mathcal{P}(\mathcal{O}), \subseteq)$ *est* admissible *si elle vérifie les propriétés suivantes :*

1. ext *est une fonction croissante ;*
2. *pour tout couple* (a, b) *d'éléments de* $\mathcal{A}$, $ext(a) \cap ext(b) \neq \emptyset$ *implique* $\min(a, b) \neq \emptyset$.

La première condition donne à la relation d'ordre sur les descriptions le sens de la relation de subsomption [Napoli, 1997] : $a \leq b$ (c'est-à-dire a plus précis que b) implique que pour toute interprétation admissible, l'extension de a est contenue dans l'extension de b.

La seconde condition demande à l'application ext une certaine "complétude" : on suppose que si deux descriptions décrivent les mêmes individus, c'est qu'il existe au moins une autre description plus précise que les deux premières.

4.3 Clôture d'une relation

Soit R une relation sur un ensemble $\mathcal{T}$. Nous rappelons les définitions et notations suivantes :

1. ${}^tR$ est la transposée de R, obtenue par ${}^tR = \{(x,y) \in \mathcal{T}^2 \mid (y,x) \in R\}$;
2. R^s est la clôture de R par symétrie, i.e., $R^s = R \cup {}^tR$;
3. Si S est une autre relation sur $\mathcal{T}$, le produit RS est défini par : $RS = \{(x,y) \in \mathcal{T}^2 \mid \exists z \in \mathcal{T},\ (x,z) \in R \text{ et } (z,y) \in S\}$;
4. R^k est la puissance k-ième de R, définie par récurrence par $R^1 = R$ et $R^k = R^{k-1}R$;
5. R^+ est la clôture de R par transitivité, définie par $R^+ = \bigcup_{k\geq 1} R^k$;
6. $\mathcal{S}(R)$ est le support de R, i.e., l'ensemble $\mathcal{S}(R) = \{x \in \mathcal{T} \mid \exists y \in \mathcal{T},\ (x,y) \in R^s\}$.

Définition 4 *Soit deux relations R et S sur un même ensemble :*

1. *on dit que R est stable par S, si et seulement si, $RSR \subset R$;*
2. *R_S désigne la clôture de R par S, i.e., la plus petite relation contenant R et stable par S.*

La démonstration de la proposition suivante (et des autres résultats présentés) est disponible dans [Rossi and Vautrain, 1999].

Proposition 1 *Soit R et S, deux relations sur un même ensemble. On définit $R_S^0 = R$, et par récurrence, $R_S^k = R_S^{k-1}SR$. On a alors :*

1. *$R_S = \bigcup_{k\geq 0} R_S^k$;*
2. *si R et S sont symétriques, il en est de même de R_S ;*
3. *si R est transitive, R_S l'est aussi.*

4.4 Cohérence symbolique

Définition 5 *Soit $(\mathcal{A}, \leq)$ un espace de description ordonné et F une relation sur $\mathcal{A}$. On note $\widetilde{F}$, la clôture par S de $(F^s)^+$, où S désigne la relation suivante : $S = \{(a,b) \in \mathcal{A}^2 \mid \min(a,b) \neq \emptyset\}$.*

$\widetilde{F}$ est alors symétrique et transitive. De plus, pour tout $a \in \mathcal{S}(F)$, $(a,a) \in \widetilde{F}$.

Définition 6 *Soit (F,E) un savoir expert sur $(\mathcal{A}, \leq)$ un espace de description ordonné. (F,E) est dit* symboliquement cohérent *sur $(\mathcal{A}, \leq)$ (et on note $F \propto E$) si et seulement si :*

1. *pour tout $(a,b) \in E$, $\min(a,b) = \emptyset$;*
2. *pour tout $(a,b) \in F$ et tout $(c,d) \in E$, $\min(a,c) = \emptyset$ ou $\min(b,d) = \emptyset$.*

Théorème 1 *Soit (F,E) un savoir expert sur $(\mathcal{A}, \leq)$ un espace de description ordonné. Les deux propositions suivantes sont équivalentes :*

1. *$\widetilde{F} \propto E$;*
2. *pour tout jeu de données $\mathcal{O}$ et pour toute extension admissible ext, il existe une relation d'équivalence r sur $\mathcal{O}$* compatible *avec (F,E) interprété par ext.*

De plus, on obtient la relation la plus fine en traduisant $\widetilde{F}$ sur $\mathcal{O}$ par l'intermédiaire de ext.

Le théorème précédent montre qu'il est donc possible de déterminer par un calcul purement symbolique si le savoir expert est cohérent, c'est-à-dire si il est possible de construire une relation d'équivalence sur un jeu de données qui respecte ce savoir. Les conséquences pratiques sont importantes :

1. la classification la plus fine est calculée symboliquement, ce qui sera en général plus rapide qu'un calcul sur la population (dans les deux cas, on calcule les composantes connexes d'un graphe, qui est en général plus petit dans le cas symbolique) ;
2. la relation $\widetilde{F}$ peut servir de base à une classification symbolique/numérique ;
3. si un savoir expert est symboliquement cohérent, on a la certitude qu'il est possible de construire une partition. Au contraire, obtenir une relation d'équivalence compatible sur un jeu de données ne permet pas d'assurer l'existence d'une telle relation sur un autre jeu de données (par exemple si le jeu de données de départ n'est pas significatif car il comporte trop peu d'individus) ;
4. l'interprétation précise de la connaissance symbolique (i.e., la fonction *ext*) n'a pas besoin d'être connue, ce qui permet d'étudier globalement différentes interprétations possibles.

5 Traitement de contraintes incohérentes

5.1 Position du problème

Dans la pratique, on dispose d'un espace de description $(\mathcal{A}, \leq)$, d'un savoir expert (F, E), d'une interprétation *ext* des descriptions et d'un jeu de données $\mathcal{O}$. Trois situations sont alors envisageables :

1. *le savoir expert est symboliquement cohérent* : dans ce cas, on est sûr de pouvoir construire une relation d'équivalence sur n'importe quel jeu de données ;
2. *le savoir expert n'est pas symboliquement cohérent* et *il n'existe pas de relation d'équivalence compatible* sur $\mathcal{O}$: dans ce cas, si le jeu de données et l'interprétation représentent bien le problème réel étudié, on sait qu'il n'est pas possible de satisfaire les contraintes. On peut donc considérer que le savoir expert est incohérent ;
3. *le savoir expert n'est pas symboliquement cohérent*, mais **il existe une relation d'équivalence compatible** sur $\mathcal{O}$: dans ce cas, il est difficile de conclure, car on ne sait pas *a priori* si le problème vient du savoir expert ou du jeu de données.

Cette dernière situation est très intéressante : si on fait l'hypothèse que la population et l'interprétation sont "raisonnables", le problème est donc de comprendre comment le savoir expert peut ne pas être symboliquement cohérent.

5.2 Éléments de solution

Une approche possible consiste à dire que l'espace de description est trop expressif par rapport aux populations envisageables. Revenons à l'exemple de la section 2.2. On munit $\mathcal{A}$ de l'ordre induit par l'inclusion ensembliste, i.e., $a = (a_1, a_2, a_3) \leq b = (b_1, b_2, b_3)$ si et seulement si $a_i \subseteq b_i$ pour tout i.

Il est clair que la description $(\{faible\}, \{faible\}, \{élevé\})$ n'est pas vraiment cohérente : on ne voit pas comment un objet peu dense et de faible volume peut avoir un poids élevé.

Or, un savoir expert peut par exemple dire que les objets de poids élevé ne doivent pas être placés avec les objets de densité et de volume faibles (techniquement, le savoir expert s'écrit alors $F = \emptyset$ et $E = \{((*, *, \{élevé\}), (\{faible\}, \{faible\}, *))\}$). Symboliquement,

$(\{faible\}, \{faible\}, \{élevé\})$ est plus précis que $(*, *, \{élevé\})$ (les objets de poids élevé) et que $(\{faible\}, \{faible\}, *)$ (les objets de densité et de volume faibles). Techniquement, on peut montrer que la contrainte qui vient d'être énoncée n'est pas symboliquement cohérente.

Pour donner une réponse à ce problème, une approche consiste à supprimer de l'espace de description, les descriptions qui ne décrivent pas d'individus "pertinents". Dans l'exemple qui précède, on supprimerait $(\{faible\}, \{faible\}, \{élevé\})$.

Théorème 2 *Soit (F, E) un savoir expert sur $(\mathcal{A}, \leq)$ un espace de description ordonné. On suppose que $(\widetilde{F}, E)$ n'est pas cohérent. On suppose qu'il existe $\mathcal{O}$ un jeu de données, ext une interprétation admissible et r une relation d'équivalence sur $\mathcal{O}$ compatible avec (F, E) interprété par ext. Alors il existe un ensemble $\mathcal{B} \subset \mathcal{A}$ maximal tel que (F', E') la restriction à $\mathcal{B}$ de (F, E) vérifie $\widetilde{F'} \propto E'$.*

Ce théorème montre qu'il est donc toujours possible de supprimer certaines descriptions d'un espace de description afin d'en former un nouveau sur lequel le savoir expert est cohérent. Le savoir expert remplit donc ainsi deux rôles : il permet de contraindre les classifications acceptables, mais il permet aussi de restreindre l'espace de description considéré.

6 Conclusion

Dans cet article, nous proposons une approche symbolique de prise en compte d'un savoir expert dans un problème de classification. Cette approche a de nombreux avantages comparée à l'approche "numérique" esquissée à la section 3. Elle est en général plus efficace. Elle permet d'obtenir des résultats indépendants du jeu de données et de l'interprétation des descriptions symboliques. Enfin, elle permet d'obtenir des renseignements sur l'espace de description lui-même.

La principale direction d'extension de l'approche proposée est une prise en compte plus souple des contraintes. En effet, les contraintes sont pour l'instant satisfaites ou non. Or, on peut imaginer une approche moins binaire qui tiennent compte des effectifs des groupes décrits symboliquement par les contraintes expertes. De la même façon, il serait intéressant d'envisager des fonctions d'extension floues.

Références

[De Guio et al., 1997] De Guio, R., Erbeja, T., and Laget, V. (1997). A clustering approach for GT family formation problems. In *1st international conference on Engineering Design and Automation*, pages 18–21.

[Diday, 1998] Diday, E. (1998). L'analyse des données symboliques : un cadre théorique et des outils. Technical Report 9821, LISE/CEREMADE (CNRS UMR 7534), Université Paris-IX/Dauphine.

[Gordon, 1996] Gordon, A. D. (1996). A survey of constrained classification. *Computational Statistics & Data Analysis*, 21 :17–29.

[Napoli, 1997] Napoli, A. (1997). Une introduction aux logiques de descriptions. Technical Report 3314, Projet SYCO, INRIA Lorraine.

[Rossi and Vautrain, 1999] Rossi, F. and Vautrain, F. (1999). Constrained classification. Technical report, LISE/CEREMADE (CNRS UMR 7534), Université Paris-IX/Dauphine.