Apprentissage d'espaces prétopologiques dans un cadre multi-instance pour la structuration de données

Gaëtan Caillaut, Guillaume Cleuziou

LIFO, Université d'Orléans, Rue Léonard de Vinci, 45067 Orléans Cedex 2
prénom.nom@univ-orleans.fr

Résumé. Nous présentons dans cet article une méthode supervisée de structuration (en DAG) d'un ensemble d'éléments. Étant donnés une structure cible et un ensemble de relations sur ces éléments, il s'agit d'apprendre un modèle de structuration par combinaison des relations initiales. Nous formalisons ce problème dans le cadre de la théorie de la prétopologie qui permet d'atteindre des modèles de structuration complexes.

Nous montrons que la non-idempotence de la fonction d'adhérence rentre dans le cadre du formalisme de l'apprentissage (supervisé) multi-instance et nous proposons un algorithme d'apprentissage reposant sur le dénombrement des « sacs » positifs et négatifs plutôt que sur un ensemble d'apprentissage standard.

Une première expérimentation de cette méthode est présentée dans un cadre applicatif de fouille de textes, consistant à apprendre un modèle de structuration taxonomique d'un ensemble de termes.

1 Introduction

La structuration de données et de connaissances est le sujet d'intenses recherches depuis plusieurs décennies et concerne un très large champ d'applications. Au-delà même d'une utilité applicative directe, le besoin de structurer des données, des variables ou des concepts est devenu incontournable dans de nombreux processus décisionnels.

Différentes formes de structures peuvent être envisagées parmi lesquelles on peut citer les graphes, les arbres ou encore les DAGs (graphes orientés sans cycles). Selon le contexte, deux problématiques d'apprentissage peuvent être identifiées : l'apprentissage de la structure elle-même (ex. modèle Bayésien) ou l'apprentissage d'un modèle de structuration. Nous nous intéressons dans cette étude à la seconde problématique consistant à inférer un modèle de structuration en DAG à partir d'une structure cible et d'une description sur les données.

Nous proposons dans cet article, un cadre générique d'apprentissage supervisé d'un modèle de structuration complexe d'un ensemble d'éléments E en un DAG $\mathcal{G}$ dans lequel chaque nœud représente un élément ou un sous-ensemble d'éléments. La complexité des structures ciblées par notre approche repose sur l'usage du formalisme prétopologique permettant une modélisation fine du processus de propagation de la domination entre éléments ou ensembles d'éléments. Le cadre est présenté comme *générique* dans le sens où il considère en entrée une collection de relations $\{\mathcal{R}_1, \ldots, \mathcal{R}_K\}$ binaires sur $E \times E$, plutôt qu'une description de type

vectorielle des éléments de E. Dans ce contexte multi-source, la tâche ciblée peut alors être vue comme l'apprentissage d'une fonction de combinaison (fonction d'adhérence prétopologique) des relations fournies en entrée.

2 Apprentissage de structures et prétopologie

2.1 Notions de prétopologie et structuration

Nous reprenons ici les principales notions dans le formalisme de notations proposé dans Belmandt (1993). Nous considérerons dans ce qui suit un ensemble fini E d'éléments à structurer.

On appelle espace prétopologique, un couple (E, a) où $a()$ est une application de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ telle que $a(\emptyset) = \emptyset$ et $\forall A \in \mathcal{P}(E), A \subseteq a(A)$. La fonction $a()$ est appelée *adhérence* et modélise un phénomène d'extension. On appelle fermé de A ($F(A)$) le sous-ensemble obtenu par applications successives de $a()$ sur $A \subseteq E$ jusqu'à obtention d'un point fixe.

La spécificité de l'adhérence en prétopologie est que, contrairement à la topologie traditionnelle, elle n'est pas contrainte à satisfaire l'idempotence, de sorte que l'extension d'un sous-ensemble A peut être réalisée en plusieurs étapes (e.g. $A \subseteq a(A) \subseteq a(a(A)) \subseteq \cdots \subseteq a^n(A)$) et non en une seule fois comme en topologie où l'adhérence de A désigne directement sa fermeture ($a \circ a = a$).

Dans la suite nous considérerons uniquement des espaces prétopologiques de type V, c'est-à-dire tels que l'application d'adhérence satisfait en plus la propriété d'isotonie : $\forall A \subseteq E,\ B \subseteq E,\ A \subseteq B \Rightarrow a(A) \subseteq a(B)$

Les espaces prétopologiques de type V vérifient certaines propriétés permettant de structurer l'espace selon ses fermés élémentaires [1](Largeron et Bonnevay, 2002). La Figure 1 illustre sur un ensemble $E = \{x_1, \ldots, x_5\}$ un exemple de fermés provenant d'un espace de type V ainsi que la structure de DAG obtenue.

— $F_{x_1} = E$
— $F_{x_2} = F_{x_3} = \{x_2, x_3, x_5\}$
— $F_{x_4} = \{x_4, x_5\}$
— $F_{x_5} = \{x_5\}$

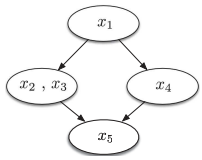

FIG. 1 – *Exemple de fermés élémentaires (à gauche) et structure de DAG induite (à droite).*

2.2 Apprentissage d'espaces prétopologiques : formalisation multi-critère et méthodologies

La prétopologie présente un formalisme particulièrement adapté à l'analyse multi-critère. Cleuziou (2015) montre en effet que, dans le contexte qui nous intéresse - où l'on dispose en entrée d'une collection $\{\mathcal{R}_1, \ldots, \mathcal{R}_K\}$ de relations binaires réflexives sur E - il est possible de définir une fonction d'adhérence engendrant un espace de type V.

1. Un fermé élémentaire correspond à la fermeture d'un singleton, on le note F_x (ou $F(\{x\})$).

Cleuziou et Dias (2015) proposent la méthode LPS (*Learning Pretopological Spaces*) comme cadre générique d'apprentissage semi-supervisé d'un espace prétopologique. Le principe de LPS consiste à rechercher une pondération sur chacune des relations initiales ; cette pondération est exploitée dans une fonction d'adhérence paramétrée qui contrôle le processus de structuration. Le principe d'adhérence paramétrée est ensuite étendu dans un formalisme logique cette fois, par Cleuziou (2015) qui définit une nouvelle classe d'espaces prétopologiques, engendrés de manière logique par la fonction d'adhérence suivante : $a(A) = \{x \in E \mid Q(A, x)\}$ avec $Q(A, x)$ une formule logique définie sur le langage des K fonctions propositionnelles suivantes : $q_k(A, x) = (B_k(x) \cap A \neq \emptyset)$, et $B_k(x) = \{y \in E | x \mathcal{R}_k y\}$. On peut alors montrer que toute formule logique Q en forme normale disjonctive et sans négation (DNF positive) définit une adhérence satisfaisant la propriété d'isotonie ; l'espace prétopologique engendré est alors de type V.

LPS repose sur un processus évolutionnaire pour sélectionner les meilleures formules Q dans une population et les faire évoluer par croisements et mutations (Cleuziou et Dias, 2015; Cleuziou, 2015). Cela nécessite une population de grande taille pour espérer explorer au mieux l'espace des solutions, ce qui implique de répéter le processus de structuration un nombre considérable de fois.

Pour améliorer l'efficacité de la méthode LPS, nous proposons dans cet article de reconsidérer le problème d'apprentissage dans un cadre supervisé cette fois, et en utilisant une heuristique de construction gloutonne de la formule Q centrée sur la fonction d'adhérence directement, plutôt que sur la structure finale induite.

2.3 Vers une formulation multi-instance du problème

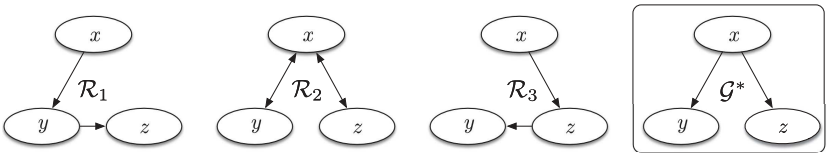

FIG. 2 – *Exemple d'une collection de 3 relations et d'une structure cible.*

Nous formalisons le problème d'apprentissage de la fonction d'adhérence comme une tâche de classification binaire sur un ensemble d'exemples, constitué de tous les couples (A, x) sur $\mathcal{P}(E) \times E$ tels que $x \notin A$. Chaque exemple est décrit par un vecteur booléen sur $\{0, 1\}^K$ où chaque composante vaut 0 ou 1 selon que la fonction propositionnelle $q_k(A, x)$ est satisfaite ou non par l'exemple.

Nous présentons en Table 1, l'ensemble des instances issues de l'exemple de la Figure 2. À chaque exemple est associée une description booléenne. Les points d'interrogation reportés dans la colonne identifiant la classe (+ ou -) de l'exemple indiquent l'ambiguïté due à la non-idempotence de la fonction d'adhérence. Il serait en effet faux d'imposer une étiquette positive à l'exemple $(\{x\}, y)$ au risque de se priver de la possibilité d'apprendre d'autres modèles satisfaisant, de même pour l'exemple $(\{x\}, z)$; en revanche, il est indispensable qu'au moins l'un de ces deux exemples soit positif car, pour espérer satisfaire le fermé cible F^*_x, l'ensemble $\{x\}$ doit nécessairement être étendu à y et/ou à z par application de l'adhérence.

(A, x)	$q_1(A, x)$	$q_2(A, x)$	$q_3(A, x)$	$x \in a(A)$	$x \in F^*(A)$
$(\{x\}, y)$	1	1	0	?	+
$(\{x\}, z)$	0	1	1	?	+
$(\{y\}, x)$	0	1	0	-	-
$(\{y\}, z)$	1	0	0	-	-
$(\{z\}, x)$	0	1	0	-	-
$(\{z\}, y)$	0	0	1	-	-
$(\{x, y\}, z)$	1	1	1	+	+
$(\{x, z\}, y)$	1	1	1	+	+

TAB. 1 – *Ensemble d'exemples décrits sur* $\{0, 1\}^3$ *et étiquettes de classe associées.*

Cette problématique de classification pour laquelle les étiquettes sont attribuées à des groupes (ou sacs) d'instances plutôt qu'aux instances elles-mêmes correspond au cadre de l'apprentissage multi-instance (Dietterich et al., 1997).

Dans notre contexte particulier d'apprentissage, il est important de remarquer que la taille (exponentielle) de l'ensemble des exemples rend impossible toute tentative de processus d'apprentissage qui nécessiterait de générer explicitement tous les exemples. Nous montrons alors qu'il est possible de dénombrer les sacs positifs et négatifs couverts par une DNF Q et ainsi d'en dériver un algorithme d'apprentissage multi-instance utilisant une heuristique basée sur une estimation de la quantité de sacs couverts/rejetés par Q et qui s'abstrait de toute génération explicite d'exemples.

2.4 Dénombrement des sacs positifs/négatifs couverts par un modèle de structuration prétopologique

Nous considérons dans la suite l'heuristique d'apprentissage consistant à construire une règle de classification de manière gloutonne guidée par le principe de maximisation du nombre de sacs positifs couverts et de sacs négatifs rejetés. Cette heuristique nécessite de connaître, au fur et à mesure de la construction du modèle, le nombre total de sacs positifs/négatifs à couvrir et au moins une estimation du nombre de sacs positifs/négatifs couverts par un modèle Q en construction.

Nombre total de sacs positifs. Pour connaître le nombre total de sacs positifs, on s'appuie sur la structure cible $\mathcal{G}^*$. On partitionne les éléments de E en classes d'équivalences $\{E_1^*, \ldots, E_p^*\}$ ($p \leq |E|$) relativement à leurs fermés élémentaires, notés $\{F_1^*, \ldots, F_p^*\}$. Chaque sous-ensemble E_k génère autant de sacs positifs qu'il existe de parties de E dans le sous-treillis (privé de sa borne supérieure), ayant E_k^* comme plus petits éléments et F_k^* comme plus grand élément : on utilisera la notation $T_{E_k^*}^{F_k^*}$ pour désigner ce sous-treillis des parties de E.

Nombre total de sacs négatifs. De façon analogue au calcul précédent, le nombre total de sacs négatifs induits par la structure cible $\mathcal{G}^*$ est obtenu en observant les sous-treillis $\{T_{E_k^*}^{F_k^*}\}$ mais cette fois sans considérer le fermé élémentaire maximal ($F_k^* = E$) qui n'induit aucun sac négatif, et en conservant les bornes supérieures pour les autres.

Nombre de sacs positifs/négatifs couverts par un modèle Q. Le nombre de sacs positifs couverts par un modèle Q en construction est estimé (au rabais) en considérant pour chaque

éléments $x \in E_k^*$ le nombre de sacs positifs à couvrir ($|T_x^{F_k^*}| - 1$) privé des sacs positifs qui ne sont pas couverts par Q : $|T^{F_k^*}_{F_k^* \cap F_k^Q}|$. Le principe de calcul reste identique dans le cas des sacs négatifs couverts par le modèle Q.

3 Application à l'acquisition de taxonomies lexicales

Nous présentons une expérimentation dont l'objectif est de construire un modèle de structuration pour la taxonomie de *wagons* (en anglais). Cette structure provient de WordNet et les relations fournies comme entrées sont extraites de corpus textuelle (Cleuziou et Dias, 2015).

Nous comparons les résultats obtenus par la nouvelle méthode (LPS multi-instance) proposée avec ceux obtenus par "LPS évolutionnaire" (Cleuziou et Dias, 2015) ainsi qu'une variante gloutonne "LPS glouton".

Le modèle de structuration appris, illustré par la Figure 3, est utilisé pour construire la taxonomie du terme *vehicles* (toujours en anglais). Pour cette seconde structuration, nous avons obtenu des scores de précision, rappel et F-mesure de, respectivement, 0.71/0.37/0.49. Pour cette même expérimentation, LPS évolutionnaire obtient 0.74/0.36/0.49 ; donc des résultats en tous points comparables. LPS glouton obtient, lui les scores 0.75/0.24/0.37. Ces résultats, confirmés par des expérimentations complémentaires sur d'autres taxonomies de WordNet, nous incitent à penser que l'approche multi-instance gloutonne de notre méthode est justifiée.

Bien que les résultats obtenus par LPS multi-instance ne soient pas nécessairement meilleurs que ceux de LPS évolutionnaire, ils sont dans la plupart des cas comparables pour une complexité bien moindre. En terme de nombre de structurations, LPS évolutionnaire a une complexité en $O(n \cdot p)$ avec n le nombre d'itérations et p la taille de la population, contre une complexité en $O(k \cdot |\mathcal{R}|^2)$, avec k le nombre de clauses de la DNF, pour LPS multi-instance et glouton. Les résultats présentés dans cette section on été obtenus avec plus de 5000 structurations dans le cas de LPS évolutionnaire contre moins de 500 pour LPS multi-instance ($|\mathcal{R}|^2 \ll p$).

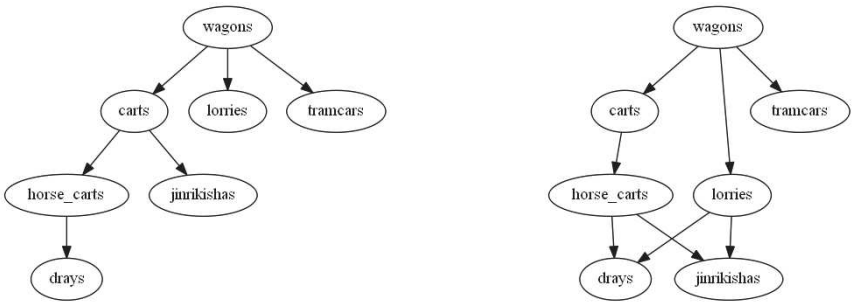

FIG. 3 – *Structure cible (à gauche) et structure apprise (à droite) du terme wagons.*

4 Conclusion

Nous avons proposé une nouvelle approche pour l'apprentissage supervisé de modèles de structuration sur un ensemble d'éléments et montré que le cadre d'apprentissage multi-instance est particulièrement adapté à notre problème du fait du caractère non-idempotent de la fonction d'adhérence en prétopologie. Dans ce cadre nous avons présenté le principe de dénombrement des « sacs » positifs et négatifs engendrés par une structure cible à reconstruire sur lequel repose notre algorithme.

Une preuve de concept a été proposée sur la tâche de reconstruction de taxonomies lexicales. Cette expérimentation préliminaire nécessite d'être approfondie en terme de diversité (et de taille) des jeux de données utilisés. Néanmoins elle valide la faisabilité de l'ensemble du processus d'apprentissage dans un contexte applicatif réel.

Cette étude offre de nombreuses pistes de travail parmi lesquelles l'amélioration de l'algorithme d'apprentissage multi-instance (prise en compte de la taille des sacs, meilleure estimation du nombre de sacs couverts), raffinement du formalisme logique utilisé (ex. passage à la logique du premier ordre). Enfin, il s'agira d'exploiter cette approche dans des domaines d'application plus variés (ex. réseaux sociaux, réseaux biologiques) et des contextes particuliers (incrémentalité, semi-supervision, etc).

Références

Belmandt, Z. (1993). Manuel de prétopologie et ses applications. *Hermes, Paris 472*.

Cleuziou, G. (2015). *Structuration de données par apprentissage non-supervisé : applications aux données textuelles*. Habilitation à Diriger des Recherches, Université d'Orléans.

Cleuziou, G. et G. Dias (2015). Learning pretopological spaces for lexical taxonomy acquisition. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015*. Springer.

Dietterich, T. G., R. H. Lathrop, et T. Lozano-Pérez (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence 89*(1), 31–71.

Largeron, C. et S. Bonnevay (2002). A pretopological approach for structural analysis. *Information Sciences 144*, 169–185.

Summary

This paper proposes an original supervised method for learning a structuring model from a set of elements described by a collection of relations (multi-view context). It uses the theory of pretopology (and the crucial pseudo-closure operator) that offers a powerful formalism leading to complex structuring models. The pseudo-closure operator being non-idempotent, we show that the underlying binary classification problem matches with the well known multi-instance learning framework. We propose a multi-instance learning algorithm based on the enumeration of the positive and negative bags of instances rather than the instances themselves. Finally a proof of concept is proposed for the whole methodology that performs the task of lexical-taxonomy reconstruction.