Apprentissage de grammaires formelles

13 avril 2008

1 Introduction

L'apprentissage automatique est un vaste domaine de l'informatique regroupant des approches très variées comme l'apprentissage de langages et d'automates, l'apprentissage sur des données relationnelles, l'apprentissage non supervisé, l'apprentissage par renforcement ou encore celui lié aux algorithmes d'évolution.

Dans ce cours, nous allons présenter l'apprentissage de langages dont les principales applications sont liées au traitement des textes dans le cadre de la linguistique informatique, au traitement d'images pour la reconnaissance de formes ou au traitement de données biologiques en bio-informatique. Nous nous situerons dans une branche de l'apprentissage des langages formels dont les origines remontent aux travaux de Gold en 1967.

2 Généralités

Nous commencerons par définir la théorie de l'apprentissage formel dont le but est de formaliser de manière rigoureuse l'inférence grammaticale. L'inférence grammaticale est un processus qui permet d'apprendre une grammaire à partir d'un sous-ensemble de phrases bien formées d'un langage particulier ayant justement été engendrées par la grammaire que l'on cherche à apprendre.

Si l'on se restreint uniquement à des exemples bien formés, on parle d'apprentissage ou d'inférence à partir d'exemples positifs. Nous resterons dans ce modèle qui semble être cohérent avec les modèles d'apprentissage du langage naturel par les humains.

3 Analogie avec l'apprentissage du langage humain

L'exemple classique d'inférence dans le monde réel est celui d'un enfant qui, à partir de très peu d'informations, uniquement produites par des phrases issues de son milieu culturel, est capable de produire à partir de l'âge de trois ans des phrases originales correctes dans sa langue maternelle.

A partir de cinq ans, l'enfant connaît toutes les règles grammaticales, la suite de son apprentissage n'étant plus qu'une extension de son lexique. Il semble de plus que cette acquisition des règles grammaticales de la langue se fasse essentiellement à partir de phrases correctes comme le confirment certains psycholinguistes spécialisés dans l'apprentissage chez les enfants.

Toute information permettant à l'enfant d'identifier une phrase comme non correcte est pratiquement absente lors de l'acquisition du langage, ou du moins celle-ci ne joue pas un rôle significatif dans la phase d'apprentissage. La relative pauvreté (quantitativement) de l'ensemble des exemples suffisant à l'acquisition de la grammaire chez l'enfant, liée au processus basé uniquement sur des exemples positifs, font de la théorie formelle de l'apprentissage un modèle concevable pour un traitement automatique de cet apprentissage.

4 Cadre Téorique de l'apprentissage

Nous commenceront par étudier le modèle de Gold, d'apprentissage à la limite, développé initialement pour déterminer les propriétés d'inférence des classes de grammaires associées aux langages de la hiérarchie de Chomsky, que vous avez étudié à l'EISTI en permière année lors du cours de Théorie des Langages.

L'apprentissage (au sens de Gold) dans notre contexte est un procédé symbolique pouvant être présenté comme suit. Soit $\mathcal{G}$ une classe de grammaires que nous voulons apprendre à partir d'exemples positifs.

Le but est de définir un algorithme qui, appliqué à un ensemble fini de phrases, retourne une grammaire de la classe qui a généré les exemples ; l'algorithme devant converger.

Formellement, soit $L(G)$ le langage associé à la grammaire G, et V un alphabet donné, un algorithme d'apprentissage est une fonction ϕ d'ensembles finis de mots dans V^* vers $\mathcal{G}$ telle que pour $G \in \mathcal{G}$ avec $L(G) =< e_i >_{i \in \mathbb{N}}$ il existe une grammaire $G' \in \mathcal{G}$ et il existe $n_0 \in N$ tel que : $\forall n > n_0\ \phi(\{e_1, \ldots, e_n\}) = G' \in \mathcal{G}$ avec $L(G') = L(G)$.

5 Résultats de l'apprentissage selon Gold

L'utilisation du modèle de Gold a pendant plusieurs années semblé impossible pour les classes non triviales de langages, en conséquence du résultat initial de non-apprenabilité de Gold (1967). Le regain d'intérêt pour l'apprentissage à la limite à partir d'exemples positifs est dû aux travaux de Dana Angluin , qui détermina une première classe non triviale de grammaires apprenables.

6 Applications

De tels algorithmes d'acquisition d'une grammaire formelle sont utiles pour les applications courantes du traitement automatique des langues telles que

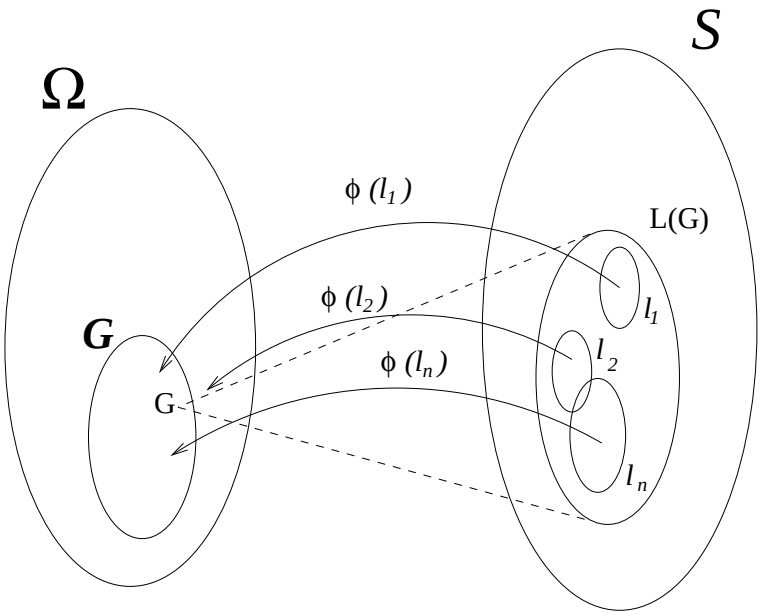


FIG. 1 – Inférence grammaticale

correction orthographique, aide à la traduction, interrogation en langage naturel.

Ces techniques interviennent également dans beaucoup d'autres domaines, comme la bio-informatique, et globalement dans l'ensemble des thématiques de l'apprentissage automatique ("machine learning").

7 Définitions et propriétés

Nous allons détailler les concepts de base de la théorie formelle de l' inférence grammaticale, formulés par Gold, ainsi que plusieurs propriétés importantes pour la compréhension des résultats que nous présenterons par la suite.

La notion de langage est à la base de la théorie formelle de l'apprentissage développée par Gold. Afin de définir l'apprentissage, il faut spécifier le domaine de recherche dans lequel on veut apprendre. Ce dernier est composé de trois éléments distincts :

- un espace d'hypothèses Ω
- un espace d'exemples $\mathcal{S}$
- une fonction L qui associe les éléments de Ω à des sous-ensembles de $\mathcal{S}$, i.e., $L : \Omega \rightarrow part(\mathcal{S})$

Ω peut être toute classe d'objets finis sur laquelle des calculs peuvent être effectués, par exemple l'ensemble des entiers naturels, les programmes d'un certain type de machines abstraites, ou les grammaires formelles d'une classe précise. Formellement, Ω peut être vu comme un ensemble récursif de chaînes sur un al-

phabet fini. Ses éléments seront donc appelés des *grammaires*. Les grammaires que nous avons définies dans la première partie de cette thèse appartiennent bien à cet ensemble, les définitions qui vont suivre s'appliquent donc aux grammaires catégorielles. Les exemples positifs appartiennent à l'ensemble $\mathcal{S}$. Les sous-ensembles de $\mathcal{S}$ sont des langages. Il est évident que la nature des éléments de $\mathcal{S}$ a une grande influence sur l'apprentissage. Nous verrons qu'effectivement, plus l'information présente dans $\mathcal{S}$ est riche, plus l'apprentissage est facilité. La fonction L associe un langage à chaque grammaire de Ω, l'appartenance d'un élément s à $L(G)$ pour $s \in \mathcal{S}$ et $G \in \Omega$ étant désignée comme le *problème d'appartenance universel* de Ω. Le triplet $\langle \Omega, \mathcal{S}, L \rangle$ est un système grammatical.

Exemple 7.1 (Angluin) *La première classe de grammaires non triviale apprenable dans le modèle de Gold correspond au système grammatical suivant : soit Σ un alphabet fini, et Var un ensemble infini de variables, disjoint de Σ. Un motif sur Σ est un élément de $(\Sigma \cup Var)^+$. Soit Pat l'ensemble des motifs sur Σ. Pour tout $p \in Pat$, soit $L(p)$ l'ensemble des chaînes pouvant être obtenues à partir de p en remplaçant uniformément chaque variable x présente dans p par une chaîne $w \in \Sigma^+$. Par exemple, si $p = \mathtt{a}x\mathtt{b}x$, avec $\mathtt{a},\mathtt{b} \in \Sigma$ et $x \in Var$, alors $L(p) = \{\mathtt{a}w\mathtt{b}w \mid w \in \Sigma^+\}$. Un ensemble $L \subseteq \Sigma^+$ est un langage de motif s'il existe un motif p tel que $L = L(p)$. Le triplet $\langle Pat, \Sigma^+, L \rangle$ est* un système grammatical.

La seconde définition nécessaire au processus d'inférence grammaticale permet de formuler des hypothèses à partir des exemples bien formés d'un langage particulier sur la grammaire qui les a engendrés. Il s'agit de la fonction d'apprentissage.

Définition 7.2 *Soit $\langle \Omega, \mathcal{S}, L \rangle$ un système grammatical. Une* fonction d'apprentissage *est une fonction partielle qui associe aux ensembles finis non vides d'exemples positifs une grammaire,*

$$\varphi : \bigcup_{k \geq 1} \mathcal{S}^k \to \Omega$$

En proposant une fonction d'apprentissage, on conjecture l'existence d'un ensemble fini d'exemples positifs appartenant au langage engendré par une grammaire particulière. La fonction étant partielle, on ne peut a priori pas déterminer à quel moment du parcours de l'ensemble d'exemples, on a inféré la grammaire ayant engendré ces exemples. Le succès du processus d'inférence dépend donc de l'existence de cette étape (atteinte après un nombre fini d'exemples) à partir de laquelle la grammaire engendrée sur la base des exemples positifs reste identique. Formellement, cette notion est définie comme la convergence de la fonction d'apprentissage.

Définition 7.3 *Soit $\langle \Omega, \mathcal{S}, L \rangle$ un système grammatical, $\langle s_i \rangle_{i \in \mathbb{N}} = \langle s_0, s_1, s_2, \cdots \rangle$ une séquence infinie d'exemples positifs appartenant à $\mathcal{S}$ et φ une fonction d'apprentissage définissant une grammaire $G_i = \varphi(\langle s_0, \cdots, s_i \rangle)$ pour tout $i \in \mathbb{N}$ tel*

que φ soit défini sur $\langle s_0, \cdots, s_i \rangle$.[1] *La fonction φ* converge *vers G sur $\langle s_i \rangle_{i \in \mathbb{N}}$ si il existe $n \in \mathbb{N}$, tel que pour tout $i \geq n$, G_i est défini et $G_i = G$.*

La nature infinie du processus d'inférence ne permet pas de déterminer à un instant précis si l'exemple suivant de l'énumération va changer l'hypothèse (la grammaire) ou non. La notion de classe de grammaires apprenable dépend donc de l'existence d'une fonction d'apprentissage pour chaque langage engendré par une grammaire de cette classe, qui converge vers cette grammaire à partir de toute énumération d'exemples positifs de ce langage.

Définition 7.4 *Soit $\langle \Omega, \mathcal{S}, L \rangle$ un système grammatical et $\mathcal{G} \subseteq \Omega$ une classe de grammaires. La fonction d'apprentissage φ* apprend *$\mathcal{G}$ si*
– *pour tout langage $L \in L(\mathcal{G}) = \{L(G) \mid \ \in \mathcal{G}\}$,*
– *et pour toute séquence infinie $\langle s_i \rangle_{i \in \mathbb{N}}$ énumérant L (i.e, $\{s_i \mid i \in \mathbb{N}\} = L$) il existe $G \in \mathcal{G}$ tel que $L(G) = L$ et φ converge vers G sur $\langle s_i \rangle_{i \in \mathbb{N}}$.*

Définition 7.5 *La classe de grammaire $\mathcal{G}$ est* apprenable *s'il existe une fonction d'apprentissage calculable qui apprend $\mathcal{G}$.*

Théorème 7.6 *Les grammaires régulières (type 3) ou algébrique (type 2) ne sont pas apprenables.*

8 Langages k-réversibles

Nous allons maintenant étudier une sous-classe de grammaires régulière apprenable. Nous ne nous interesserons pas à prouver ces propriétés mais à appliquer les algorithmes d'apprentissage afin de bien comprendre leurs principes sous-jacent.

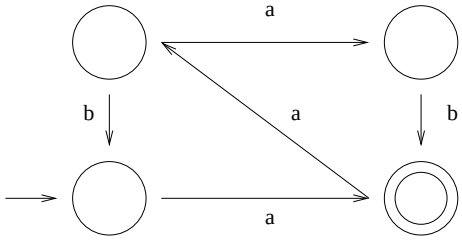
La classe de grammaire que nous allons étudier est celle des grammaires réversibles.

Les différents algorithmes que nous allons voir, manipulent les automates d'états finis (AFD) au lieu des grammaires régulières qui leur sont équivalentes. En effet toute grammaire régulière peut être simulée par un AFD. Il est très facile de retrouver la grammaire régulière à partir de l'AFD et vice-versa (cf cours de théorie des langages ING1 EISTI).

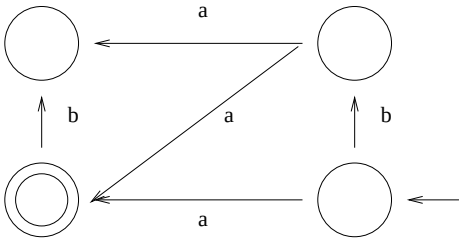
Définition 8.1 *Un automate d'états finis admet un automate inverse, que l'on obtient en inversant les transitions entre les états de l'automate, et en remplaçant les états finaux en états initiaux (et vice -versa).*

Exemple 8.2 *Soit l'automate déterministe suivant :*

[1] $\langle s_0, \cdots, s_i \rangle$ désigne une séquence finie non vide. Si $i = 0$, on la remplace par $\langle s_0 \rangle$

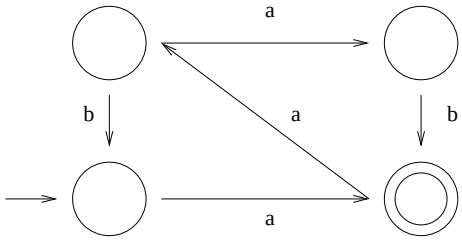


Son automate inverse est le suivant :

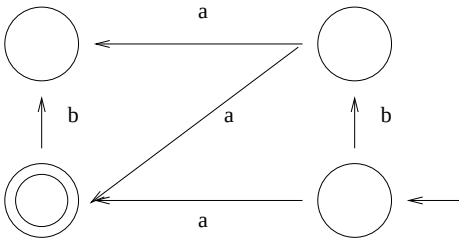


Définition 8.3 *Un automate d'états finis est k-réversible ssi il est déterministe et l'automate inverse est également déterministe avec anticipation k :*
$\forall q, q' \in Q, q \neq q', (q, q' \in Q_0) \vee (q, q' \in \delta(q, a)) \Rightarrow \neg\exists u \in \Sigma^k : (\delta(q, u) \neq \emptyset) \wedge (\delta(q', u) \neq \emptyset))$.

Exemple 8.4 *Reprenont l'automate précédant :*



Son automate inverse est déterministe avec anticipation 1 *:*



Définition 8.5 *Une grammaire régulière G est k-réversible si l'automate la reconnaissant est réversible.*

Théorème 8.6 *Les grammaire k-réversibles sont apprenables*

8.1 Méthode d'agrégation des suffixes

L'ensemble des suffixes depuis un état de l'automate est l'ensemble des chaîne entre cet état et l'état final.

Cette méthode d'inférence de grammaires réversibles calcule la distance d'inclusion entre deux ensembles de suffixes A et B :

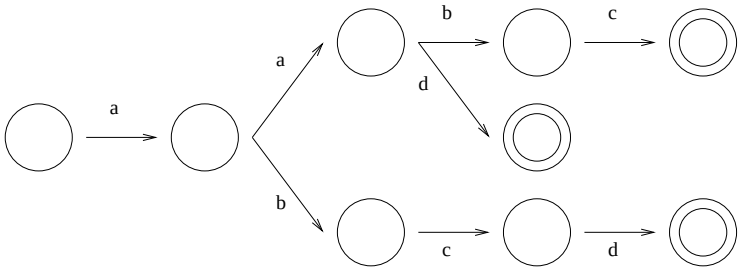
– si $A \cap B \neq \emptyset$, $d(A, B) = min(|A|, |B|) - |A \cap B|$

– si $A \cap B = \emptyset$, $d(A, B) = \infty$

Afin d'écrire l'algorithme d'apprentissage par la méthode d'aggrégation des suffixes, il nous faut une dernière définition, celle de l'automate canonique minimal.

Définition 8.7 *L'automate canonique minimal construit à partir d'un ensemble d'exemples, s'obtient en créant autant de chemins qu'il y a d'exemples à partir d' un état initial, en regroupant les préfixes communs sur un unique chemin.*

Exemple 8.8 *A partir des exemples (aabc aad abcd), on obtient l'automate canonique suivant :*



Algorithme :

1. A partir de l'ensemble d'exemple, on construit l'automate canonique minimal.
2. Construction du tableau des distances entre ensemble de suffixes pour tous les états pris 2 à 2
3. Réduction du tableau en fusionnant les états ayant des distances finies
4. Répétition de l'étape précédente tant qu'il reste des distances infinies

8.2 Méthode des k suffixes

Soit $z \in T$ une chaîne telle que $zw \in R$ pour $w \in T$. L'ensemble des k-suffixes de z par rapport à R constitue l'ensemble des sous-chaînes de longueur $\leq k$ ayant z comme préfixe dans R : $h(z, R, k) = \{w \mid zw \in R \text{ et } |w| \leq k\}$

Soit $Q = \{q \mid h(z, R, k) \text{ pour } z \in T^*\}$ et pour chaque $a \in T$, $\delta(q, a) = \{q' \in Q \mid q' = h(za, R, k) \text{ où } q = h(z, R, k)\}$

L'automate construit à partir de R a pour états des sous-ensembles de l'ensemble de toutes les k-suffixes pouvant être construits à partir de R.

9 Exercices

1. Les langages suivants sont-ils 0-réversibles, 1-réversible :
   $\Sigma^*, 1^*01, 0^*1^+, 11^*$
2. Trouver un langage non 1-réversible
3. Soit $R = \{abcd, abcbcd\}$.
   Trouver une grammaire qui reconnait toutes les chaînes de R en utilisant la méthode d'agrégation des suffixes
4. Soit $R = \{a, ab, abb\}$.
   Trouver une grammaire qui reconnait toutes les chaînes de R en utilisant la méthode des k-suffixes avec $(k = 1, k = 2)$.
5. Soit $R = \{a^5b^4, a^6b^3, a^4b^5\}$.
   Appliquer les méthodes suivantes pour l'inférence grammaticale
   (a) k-suffixes $(k = 1, k = 5, k = 9)$
   (b) agregation des suffixes