Comment approximer la complexité algorithmique à l'aide de la notion de réseau de concepts

Ales Bianchetti Jacques & Cornuéjols Antoine

LRI, URA 410 du CNRS, Université Paris-Sud, Orsay

bâtiment 490

91405 Orsay Cedex

(email : {ales,antoine}@lri.fr)

RÉSUMÉ.

La complexité algorithmique K, ou complexité de Kolmogorov, est reconnue depuis quelques années comme une notion fondamentale pour l'apprentissage, en particulier par ses liens avec l'approche bayésienne de l'induction, et avec la théorie de l'induction due à Vapnik. Son exploitation effective, illustrée dans plusieurs travaux, se heurte cependant à la non-calculabilité de cette mesure de complexité. Elle doit donc être approximée. Cependant, les systèmes d'induction actuels qui l'utilisent reposent trop souvent sur des choix ad hoc, peu systématisés et difficiles à interpréter.

Nous introduisons ici deux voies pour l'approximation de K, basées sur l'étude des ensembles d'hypothèses utilisables. La première vise à trouver le plus grand ensemble d'hypothèses tel que K restreinte à ce domaine soit calculable. La seconde définie une exploration de l'ensemble maximal des hypothèses à l'aide de sous-ensembles finis d'hypothèses. Pour représenter ces ensembles d'hypothèses, et leur fournir une structure, nous introduisons une formalisation des réseaux de connaissances, un outil naturel en intelligence artificielle.

Nous démontrons donc que K, restreinte à l'ensemble des fonctions récursivement énumérable est calculable, et nous conclurons en illustrant l'utilisation des réseaux de concepts, dans le cadre de l'apprentissage, pour les problèmes de compréhension et d'utilisation du raisonnement par analogie.

MOTS-CLÉS : *Analogie, Induction, Complexité de Kolmogorov, MDLp, Machines de Turing, Graphes de contrôle.*

1. Introduction

La *complexité algorithmique* ou *complexité de Kolmogorov* d'un objet, ou d'une séquence *s*, est définie par la longueur du programme le plus court permettant de l'engendrer à l'aide d'une machine de Turing universelle (voir [Li & Vitanyi,93]). Ainsi l'ensemble des séquences se trouve doté d'une mesure universelle. La notion de contenu d'information qui en résulte présente de nombreux attraits pour l'apprentissage : elle permet de caractériser un objet unique, et est directement liée à la théorie de l'induction. Ainsi, Li et Vitanyi, ont montré d'une part que la complexité de Kolmogorov permet de définir une distribution universelle sur les hypothèses candidates pour expliquer les données d'observation, mais aussi qu'effectivement le programme (ou hypothèse) le plus court permettant de rendre compte des données est celui qui fournit la meilleure explication des données [Li & Vitanyi, 95] et qui permet les meilleures prédictions (ou généralisations) [Vitanyi & Li, 97]. Rissanen a ensuite proposé une mesure le Minimum Description Length principle (MDLp) pour appliquer K dans le domaine de l'induction.

Malheureusement, la complexité algorithmique est non calculable dans le cas général. Il faut donc avoir recours à des méthodes permettant de contourner cet obstacle. On peut ainsi chercher à limiter l'ensemble des hypothèses pour trouver un sous-ensemble, maximal en taille, où la mesure K devienne calculable. Nous montrons que l'ensemble des fonctions récursives totales est un bon candidat. On peut aussi chercher à contraindre et spécialiser les hypothèses pour qu'elles soient adaptées au problème spécifique (§ figure 1). Ces deux approches peuvent se combiner. Dans la plupart des cas, la mise en pratique de la définition d'ensembles d'hypothèses est certes éclairée, mais ad hoc et difficile à systématiser. La recherche rapportée ici découle d'un effort

La formalisation à l'aide de réseaux de concepts apporte un guide pour une mise en pratique aussi systématique que possible partant des spécifications du problème et pouvant tirer parti des connaissances expertes à la fois sur les descripteurs et concepts intéressants mais aussi sur les connaissances a priori du système.

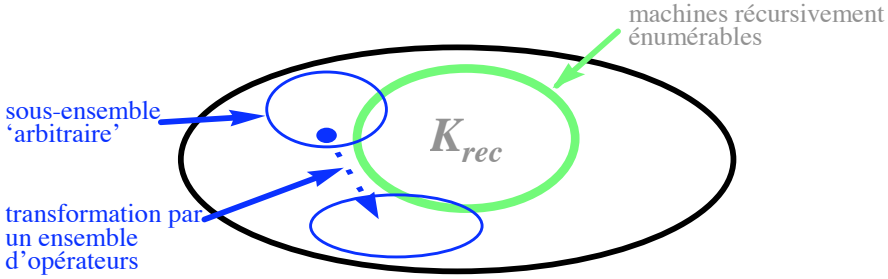

Figure 1 : Deux stratégies d'approximation de K

2. Définition d'un réseau de concepts

Dans le cadre de la complexité algorithmique, une hypothèse est un algorithme qui reconnaît l'occurrence du concept dans la séquence dont on mesure la complexité. Ces algorithmes peuvent être représentés par des noeuds étiquetés par le nom du concept qu'ils

reconnaissent. Un arc entre deux noeuds représente l'utilisation dans un algorithme de reconnaissance, d'un concept déjà défini ailleurs.

3. Une restriction calculable : un biais de représentation

La non calculabilité provenant de la taille infinie des graphes, il est nécessaire de trouver une restriction de l'espace d'hypothèses qui préserve la propriété de convergence vers le calcul de K.

Théorème : *La complexité de Kolmogorov est calculable dans l'ensemble des machines de Turing associé à l'ensemble des fonctions récursives totales.*

La démonstration se fait par exploration du graphe restreint aux fonctions récursivement énumérables. Nous pouvons toujours fixer une borne sur $K_{REC}(s)$; par exemple le calcul de la fonction Identité appliquée à *s*. En évaluant le coût de cette solution, on obtient bien une borne notée B_{Id} sur $K_{REC}(s)$, puisque l'on a bien définit un programme qui produit *s*. Considérons maintenant les graphes qui permettent de produire *s*. Cet ensemble est infini, mais on peut élaguer le limiter à la borne B_{Id}. Nous obtenons ainsi un ensemble fini.

La recherche du programme de longueur minimale sur cet ensemble fini est alors calculable. K_{REC} est donc calculable.

Il est intéressant de comparer la restriction à l'ensemble des *fonctions récursives totales* examiné ici avec un espace classiquement étudié en apprentissage : les clauses de Horn. Celles-ci correspondent aux *fonctions co-récursives partielles*, c'est-à-dire aux formules du premier ordre quantifiées universellement. Or les *fonctions récursives totales* sont à l'intersection des *fonctions co-récursives partielles* et des *fonctions récursives partielles*, c'est-à-dire les formules du premier ordre quantifiées existentiellement (§ figure 2). Ces fonctions ne couvrent donc pas tout l'espace des fonctions usuellement considérées en Programmation Logique Inductive, mais il s'agit cependant d'un sous-ensemble respectable capable de satisfaire bien des domaines d'application.

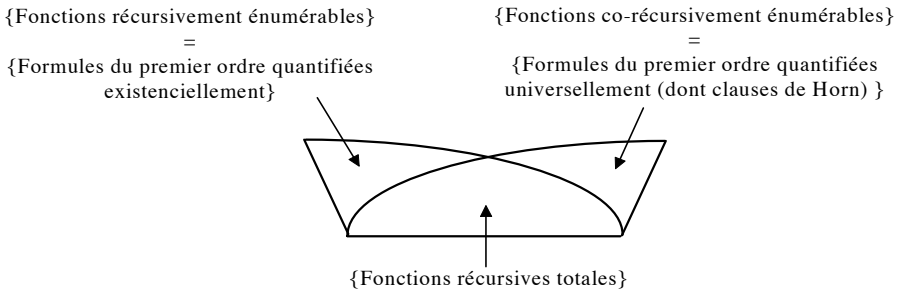

Figure 2 : Hiérarchie des fonctions récursives [De Rougemont, 96]

4. Exploration des graphes de contrôle finis : une biais de contrôle

Une autre manière d'aborder le problème de la non calculabilité de K, consiste à explorer le graphe, en partant d'un ensemble d'hypothèses donné, fini. Il faut alors s'assurer qu'une *convergence* est garantie, c'est-à-dire que n'importe quel sous-graphe peut être

obtenu par exploration à partir de n'importe quel graphe initial. Pour travailler avec ces sous-graphe, nous définissons des opérateurs : un opérateur d'ajout de noeud (au plus un par noeud possible), un opérateur générique de suppression de noeud, et un opérateur qui modifie les coûts associés aux noeuds. A la limite, on obtient une convergence vers l'ensemble de toutes les hypothèses possibles.

L'ensemble maximal d'opérateurs, que nous appellerons *ensemble saturé*, n'est pas satisfaisant du fait de son indépendance vis à vis de l'état courant, et de la possibilité qu'il crée de toujours parfaitement adapter le réseau à la production de la plus arbitraire des séquences *s* : C'est le problème du *surapprentissage*. Nous adaptons donc la mesure de Vapnik [Vapnik, 95] pour contrôler cette capacité d'adaptation.

5. Conclusion : application à l'analogie

La recherche dÈcrite ici a ÈtÈ stimulÈe par l'étude du raisonnement par analogie, particulièrement dans la lignée des travaux de Hofstadter [Hofstadter ..]. [Cornuéjols,'96] a proposé une mesure fondamentale de la qualité d'une analogie, reposant sur la minimisation d'une certaine quantité d'information associée. Afin de pouvoir la mettre en œuvre, dans un contexte où la notion de réseau sémantique est naturelle, nous avons introduit la notion de réseau de concepts [Ales&Cornuéjols,'98]. la recherche du meilleur graphe s'impose alors.

Nos objectif actuels sont d'une part de réaliser un système de raisonnement par analogie à partir de ce idées, et , d'autre part, d'explorer les liens avec la théorie de l'induction de Vapnik.

Références

[Ales&Cornuéjols,'96] Ales Bianchetti J. Cornuéjols A., C*ognitive hints for a computational model of analogy*, Workshop ECML'98, Springer Verlag, 1998

[Kolmogorov 65] Kolmogorov A, *Three approaches for defining the concept of information quantity*. *Information transmission*, 1, 3-11, 1965

[Vitanyi&Li 97] Vitanyi P. Li M., *On prediction by data compression,* In Proc. of the European Conference on Machine Learning (ECML-97), LNAI, 1224, 14-30, Springer-Verlag, 1997.

[Vitanyi&Li 95] Vitanyi P. Li M, *Computational machine learning in theory and praxis,* Computer Science Today, J. van Leeuwen, Ed., LNCS, 1000, 518-525, Springer-Verlag, 1995.

[Vitanyi&Li 93] Vitanyi P. Li M, *Introduction to Kolmogorov complexity and its applications,* Springer-Verlag, 1993.

[Mitchell 93] Mitchell M., *Analogy-Making as Perception*, MIT Press, 1993.

[Rissanen 89] Rissanen, *Stochastic Complexity in Statistical Inquiry,* World Scientific, 1989.

[Rissanen 95] Rissanen, *Stochastic complexity in learning,* In proc. of the EUROCOLT'95 Conference, LNAI-904, 196-210, Springer-Verlag, 1995.

[De Rougemont 96] de Rougemont M. Lassaigne R., *logique et complexité*, ed Hermes, 1996

[Vapnik 95] Vapnik V., *The nature of statistical learning theory,* Springer-Verlag, 1995.