Induction Extensionnelle: définition et application à l'acquisition de concepts à partir de textes

Yves Kodratoff

CNRS, U. Paris-Sud, LRI, Bât. 490 91405 Orsay
yk@lri.fr
http://www.lri.fr/~yk/

Résumé. Lorsque des outils inductifs sont inclus dans un système d'acquisition des connaissances, on dit que l'on construit un *système apprenti*. C'est dans le but de soulager la charge de travail de l'expert du domaine que cette forme d'apprentissage comporte des outils inductifs. La difficulté tient en ce que l'énumération des connaissances expertes produit des données peu bruitées mais très incomplètes que les itérations successives d'induction vont compléter, toutefois en y ajoutant de grandes quantités de bruit. Il en résulte qu'on doit utiliser des procédures inductives spéciales, adaptées à l'apprentissage par croissance de noyaux de connaissance supervisée. En particulier, pour résoudre le problème difficile de la reconnaissance de concepts dans les textes, nous avons défini une forme d'apprentissage qui intègre *l'apprentissage à partir d'instances* et les systèmes apprentis, que nous nommons ***'Induction Extensionnelle'***, un oxymoron qui souligne que malgré l'absence de création d'un modèle explicite, une induction prend effectivement place.

1. Introduction

Parmi les nombreuses techniques utilisées pour faire réaliser un certain apprentissage par un ordinateur, deux occupent une place particulière, pour des raisons très différentes. Il s'agit d'une part des Systèmes Apprentis comme le système DISCIPLE de Tecuci [Kodratoff et Tecuci, 1987; Tecuci, 1998], particuliers à cause de leur complexité et de leur nombre réduit d'applications, et d'autre part de l'Apprentissage à partir d'Instances (souvent désigné par IBL: "Instance-Based Learning") [Kibler et Aha, 1988], très simple et très appliqué, mais particulier parce qu'il ne propose pas de généralisation des exemples.

Les Systèmes Apprentis se construisent en observant et analysant les étapes de résolution de problème introduites par les experts du domaine. Un autre exemple de tels systèmes, en plus de DISCIPLE déjà cité, est le premier de tels systèmes: LEAP [Mitchell et al., 1985].

L'Apprentissage à partir d'Instances a connu de très nombreuses applications, en particulier en linguistique. Par exemple, il a été utilisé pour l'analyse syntaxique [Cardie et Pierce, 1998], la catégorisation de textes [Riloff et Lehnert, 1994], l'étiquetage grammatical [Daelemans et al., 1996], pour la levée d'ambiguïté du sens des mots [Hoste et al., 2002]. L'apprentissage à partir d'instances repose sur une technique calculatoire appelée les k plus proches voisins. Cette technique est particulièrement bien adaptée aux techniques permettant le passage d'un apprentissage supervisé à un non supervisé: les quelques exemples classés de départ forment un noyau autour duquel on agglomère les exemples inconnus, augmentant

ainsi la taille des classes. Nous allons systématiquement utiliser cette propriété en induction extensionnelle, car nous demandons en effet à l'expert de constituer un noyau d'exemples bien classés, et nous ferons croître ces noyaux en utilisant une notion de distance.

2. L'Induction Extensionnelle

2.1 Le type de problèmes abordés

Dans l'induction extensionnelle, la présence constante d'un expert du domaine permet de résoudre les méta problèmes liés à l'évaluation à la volée des résultats intermédiaires et donc de poser de façon originale les problèmes d'apprentissage.

La contrepartie de cet avantage est que les problèmes étudiés ne peuvent pas être les problèmes jouets sur lesquels on teste habituellement les programmes d'apprentissage: le logiciel doit être immédiatement utilisable par un expert ce qui implique un développement assez lourd.

Ainsi, nous faisons l'hypothèse constante que nous traitons des problèmes qu'un expert du domaine doit résoudre dans sa vie scientifique ou industrielle.

En pratique, nous n'avons abordé en détail qu'un problème, celui de la reconnaissance de concepts dans un texte, en particulier dans des textes scientifiques relatifs à la fouille de données [Fontaine et Kodratoff, 2002]. Quand on traite le texte écrit, de nombreux problèmes procèdent du Traitement Automatique de la Langue Naturelle et leur exposé complet obscurcit la description des étapes d'induction extensionnelle proprement dite, c'est pourquoi nous n'expliquerons pas ici ces problèmes, malgré leur évidente importance.

2.2 Représentation des connaissances et mesures de distance

La représentation des connaissances et les propriétés des connaissances permettant de caractériser un concept appartiennent plus ou moins directement aux Sciences Cognitives. Elles constituent un choix qui lui-même peut être vu comme une hypothèse de travail à confirmer en examinant les résultats obtenus. Notre hypothèse que l'expert se sent impliqué dans l'exploitation des résultats nous permet d'aborder ce problème, au moins de façon satisfaisante pour chaque expert. En Reconnaissance des Formes, par exemple, il est tout à fait possible que certains experts se servent intensément d'une propriété, disons la texture, et peu d'autres propriétés, comme la couleur. Dans tous les cas, si l'expert du domaine est consulté, il signalera les cas où la représentation des connaissances ne prend pas en compte sa problématique, et l'informatique doit s'adapter à ses exigences.

Cependant, et en opposition avec l'adage bien connu que "tout est dans la représentation des connaissances", nous prétendons plutôt "tout est dans une bonne adaptation des mesures de distance à la représentation des connaissances". Quelle que soit la représentation des connaissances exactement adoptée, on sait que les exemples sont représentés par une suite de symboles dont la signification fait partie de la connaissance du domaine. C'est alors à l'expert du domaine de fournir une liste de concepts intéressants pour la résolution de son problème, et de spécifier quelle(s) représentation(s) lui paraît(paraissent) la(les) meilleure(s) pour cette liste de concepts, et enfin de spécifier les relations au sein des symboles qui évoquent pour lui la présence d'un concept intéressant.

Un autre lourd travail demandé à l'expert est qu'il fournisse les noyaux autour desquels l'apprentissage prendra place. Ces noyaux sont des instances de relations existant au sein de mots décrivant les exemples, chacune de ces instances devant être rattachée par l'expert au concept qu'elle illustre.

2.3 Représentation des connaissances et mesures de distance

Dès que l'expert a fourni quelques relations qui caractérisent les concepts recherchés, plusieurs distances peuvent être définies. Pour la simplicité de l'exposé, supposons que ces relations soient de nature binaire comme (mot_1, mot_2). Il est alors évident que l'on peut effectuer de nombreuses mesures. Par exemple, on peut mesurer le nombre d'occurrences de mot_1 (resp. mot_2) dans les exemples; le nombre d'occurrences de mot_1 avec des mot_2' qui sont soit des hyperonymes, soit des hyponymes de mot_2, soit non liés à mot_2; le nombre d'occurrences de mot_1 dans les noyaux des concepts définis par l'expert; on peut créer les classes de mots liés à mot_1 (resp. mot_2) et mesurer une distance entre ces classes et les concepts définis par l'expert. Le système ASIUM [Faure et Nédellec, 1998], par exemple, utilise une combinaison de ces mesures pour définir la distance entre une instance et une classe pour réaliser des inductions strictement non supervisées.

De façon générale, on peut toujours définir deux types de distances. L'une est la distance entre un couple appartenant à un concept défini par l'expert et un couple dont la classe conceptuelle est à trouver. Elle repose sur l'existence des noyaux de base définis par l'expert. Dans notre application linguistique, nous avons nommé m_1 cette mesure. L'autre mesure est indépendante de l'expert et ne dépend que de la nature des suites de symboles représentant les exemples. Elle est basée sur la probabilité, dans l'ensemble des suites de symboles, que mot_2 se trouve lié à mot_1, pour chaque mot_1 et chaque mot_2, c'est à dire sur une évaluation de la probabilité que mot_1 et mot_2 soient couplés dans le corpus étudié.

Ces deux mesures sont combinées pour réaliser deux effets opposés.

Si les deux mesures dépassent une certaine valeur (typiquement: l'une est supérieure à la moyenne et l'autre supérieure à la moyenne plus l'écart-type) alors la distance entre deux couples de mots est suffisamment faible pour qu'ils soient affectés à la même classe ou au même concept.

Si un couple de mots déjà attribué à un concept (soit par l'expert, soit par une itération précédente) se trouve beaucoup plus près d'un autre concept alors le système propose à l'expert de modifier l'attribution de ce couple. Cette mesure permet de contredire des attributions déjà faites et d'assurer une sorte de cohérence par rapport aux exemples, dans la façon dont elles sont faites.

En conclusion, l'induction extensionnelle est caractérisée par le fait qu'elle crée un modèle en extension et non pas en intention, tout comme l'apprentissage à partir d'instances, et par cette propriété capitale d'utiliser plusieurs mesures, définies en accord avec la représentation des connaissances expertes, d'optimisation de la recherche dans l'espace des solutions et de les combiner pour soit accepter de nouvelles attributions, soit en rejeter des anciennes. Cette dernière propriété pourrait bien entendu être celle d'autres systèmes d'induction.

3. Un exemple: la reconnaissance de concepts dans les textes

Induction Extensionnelle

On trouvera des revues de l'état de l'art sur le rôle de l'apprentissage en linguistique calculatoire dans [Daelemans, 1999; Cardie et Mooney, 1999]. Le travail le plus proche du nôtre est celui de [Yarowsky, 1995], relatif à la levée de l'ambiguïté du sens des mots. Ce sujet de recherche est très proche du notre, puisque reconnaître qu'un mot appartient à un concept permet (ou exige) de lever l'ambiguïté quant au sens de ce mot.
La recherche sur ce dernier thème, cependant, est concernée par des mots définis dans les dictionnaires [Pedersen, 2002] et non pas par des mots décrivant des concepts de spécialité pour un problème particulier. Ces mots sont donc définis pour un petit nombre de spécialistes qui, de plus, peuvent même ne pas s'être encore mis d'accord sur le sens des mots!
Comme nous, Yarowsky utilise un noyau autour duquel il agglomère les exemples, et donc il utilise la même technique de passage du supervisé au non supervisé que la notre, bien qu'il affirme travailler en apprentissage non supervisé . Cependant, l'algorithme d'apprentissage supervisé qu'il utilise construit un modèle à partir des exemples. Il est, comme nous l'avons signalé plus haut, donc mal adapté au fait de partir d'un petit noyau.
Une dernière ressemblance avec le travail de Yarowsky est que nous utilisons, comme lui, la collocation comme méthode de détection des relations entre mots. Ces relations sont alors considérées comme des traces linguistiques de la présence d'un concept de spécialité dans le texte. Ils font partie de la connaissance nécessaire à la résolution du problème spécifique que l'expert du domaine désire résoudre.

Les collocations ne sont évidemment qu'une forme possible de traces linguistiques, et c'est une hypothèse de travail de considérer qu'elles suffisent à reconnaître les concepts. Pour une définition linguistiquement fondée des collocations, on se rapportera aux travaux de Halliday, par exemple le chapitre 6, *Relations Lexicales*, dans l'édition de ses articles choisis [Halliday, 1976]. Notre définition orientée vers les calculs repose sur la notion de dépendances lexicales nominales, c'est-à-dire que la collocation doit comporter au moins un nom ou groupe nominal.

Une dépendance lexicale nominale est constituée de deux entités, que nous appelons un pivot et un terme. Un terme est un terme nominal constitué d'un nom associé à un adjectif, un adverbe ou un autre nom. Par cette définition, un terme ne contient jamais un groupe verbal. Le pivot peut être un verbe, un adverbe, un nom, un adjectif ou un autre terme nominal. Quand un des deux composants de la dépendance nominale contient un groupe verbal, ou ne contient pas de nom, alors c'est le pivot. Par contre, si les deux composants contiennent un nom, alors l'utilisateur a le choix entre deux options: 'le premier est le pivot', ou 'le second est le pivot'. La première option est par défaut en français, la seconde en anglais.

Le premier travail d'induction est la constitution d'une liste exhaustive de tous les groupes, de taille au moins n, de collocations pour tous les pivots possibles, dans un corpus contenant N mots différents. Nous avons construit un algorithme qui construit ces groupes en temps N^2. Une recherche exhaustive est alors toujours possible en quelques heures de temps du calcul pour n>3 et sur un ordinateur qui travaille au moins à 1 giga hertz. Par exemple, cette procédure permet d'obtenir le groupe 2229, tel que n = 4, donné en italiques ci-dessous.

Trois des collocations du groupe 2229 sont contenues dans la classe 'implication' définie par l'expert du domaine. Une des collocations de ce groupe n'appartient pas à 'implication'. Ainsi, 75% de 2229 appartient à 'implication'. Plus généralement, mesurer m_1, c'est évaluer combien le groupe défini par l'expert et le groupe obtenu automatiquement se recouvrent.

**(part:Nom,de:Preposition,motivation:Nom)*	2229
(facteur:Nom,de:Preposition,motivation:Nom)	2229
(surcroît:Nom,de:Preposition,motivation:Nom)	2229
(élément:Nom,de:Preposition,motivation:Nom)	2229
(facteur:Nom,de:Preposition,motivation:Nom)	implication
(surcroît:Nom,de:Preposition,motivation:Nom)	implication
(élément:Nom,de:Preposition,motivation:Nom)	implication

La valeur de m_2 évalue la fréquence de cooccurrence dans tous les groupes entre un nom et les autres noms observés dans le même groupe. Dans l'exemple ci-dessus, m_2 évalue donc combien de fois le nom 'part' se retrouve dans un même groupe (appartenant à la liste exhaustive des groupes) avec 'facteur', 'surcroît' et ' élément'. Si m_1 et m_2 sont assez élevées, alors la collocation (part:Nom,de:Preposition,motivation:Nom) sera ajoutée à la liste des traces linguistiques du concept 'implication'.

Cette induction est exécutée sans consulter l'expert. Ceci est un trait avantageux en termes de temps de travail de l'expert, mais peut conduire à une catastrophe quand on ajoute de nombreuses collocations erronées. Pour éviter ce défaut, nous détectons les dépendances qui contredisent les définitions existantes des concepts et nous les présentons à l'expert.

Pour éliminer une dépendance d'un concept déjà défini (par l'expert humain, ou par des pas d'induction antérieurs), et si une collocation est proposée comme trace de plusieurs concepts, nous définissons une mesure m_3, combinaison de m_1 et m_2, qui mesure la "force d'attraction" et la "force de répulsion" de cette collocation vis-à-vis des concepts. Cette mesure permet de décider à quel concept attribuer la collocation.

4. Conclusion

Nous avons défini une combinaison de deux techniques classiques en Apprentissage Automatique que sont les Systèmes Apprentis et l'Apprentissage à partir d'Instances. Cette nouvelle technique d'apprentissage permet d'aborder le problème de l'acquisition de connaissances expertes en extension. Cette approche se limite aux cas où l'expert du domaine est suffisamment intéressé par le problème à résoudre pour qu'il accepte de fournir les noyaux de connaissance à partir desquelles l'induction peut prendre place. Une autre limite de cette approche est qu'elle nécessite une analyse détaillée de la façon dont les concepts du domaine sont reconnus afin de définir efficacement les trois mesures présentées plus haut.

Nous avons appliqué cette approche à la détection des concepts dans des textes scientifiques en anglais relatifs à la fouille de données, et à la biologie moléculaire, et dans des textes en français dédiés à la psychologie dans l'entreprise, appartenant à la société PerformanSe. L'exposé de nos résultats exige une description complète des étapes préalables à la phase d'induction extensionnelle, il sera fait en détail ailleurs. Des exposés partiels ont déjà été fournis dans [Kodratoff et al., 2003].

Références

[Cardie et Pierce, 1998] Cardie C., Pierce, D. Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification. COLING, pp. 218-224, 1998.

[Cardie et Mooney, 1999] Cardie C., R. J. Mooney Guest Editors, Introduction: Machine Learning and Natural Language. Machine Learning Journal 34, 1-5, 1999.

[Daelemans, 1999] Daelemans, W. Memory-Based Language Processing. Introduction to the Special Issue. Journal of Experimental and Theoretical AI 11, 1999.

[Daelemans et al., 1996] Daelemans, W., Zavrel J., Berck P., Gillis S. MBT: A Memory-Based Part of Speech Tagger-Generator. Proceedings of the Fourth Workshop on Very Large Corpora, pp. 14-27, 1996.

[Faure et Nédellec, 1998] Faure D., Nédellec C. ASIUM: Learning subcategorization frames and restrictions of selection. In Proceedings of the Text Mining workshop, 10th European Conference on Machine Learning (ECML 98), Kodratoff Y. (Ed.), 1998.

[Fontaine et Kodratoff, 2002] Fontaine, L. Kodratoff Y. La notion de 'concept' dans les textes spécialisés: une étude comparative entre la progression thématique et la texture des concepts. ASp 37-38, 59-83, 2002.

[Halliday, 1976] Halliday M. A. K., Halliday: System and function in language, selected papers edited by G. Kress, Oxford University Press, London 1976.

[Hoste et al., 2002] Hoste V., Hendrickx I., Daelemans W., Van den Bosch A. Parameter Optimization for Machine-Learning of Word Sense Disambiguation. Natural Language Engineering 8, 311-325, 2002.

Kibler, D., Aha, D. W. Comparing instance-averaging with instance-filtering learning algorithms. Proceedings of the Third European Working Session on Learning, Pitman, London, pp. 63-80, 1988.

[Kodratoff et al., 2003] Kodratoff Y., Azé J., Roche M., Matte-Tailliez O. Des textes aux associations entre les concepts qu'ils contiennent. Actes des XXXVièmes Journées de Statistique (résumé) Volume 2, pp. 599-602 Version complète dans *RNTI* **1**, 171-182, 2003a.

[Kodratoff et Tecuci, 1987] Kodratoff, Y., Tecuci, G., "Techniques of Design and DISCIPLE Learning Apprentice," International J. of Expert Systems 1, 39-66, 1987.

[Mitchell et al., 1985] Mitchell T., Mahadevan S., Steinberg L.. LEAP: A learning apprentice for VLSI design. In Proceedings of the Ninth International Joint Conference on Artificial Intelligence, Los Angeles, California, 1985. Morgan Kaufmann.

[Pedersen, 2002] Pedersen T. A Baseline Methodology for Word Sense Disambiguation. Lecture Notes in Computer Science 2276, p. 126, Springer-Verlag Heidelberg, 2002.

[Riloff et Lehnert, 1994] Riloff E., Lehnert W. Information extraction as a basis for high-precision text classification. ACM Trans. on Information Systems 12, 296-333, 1994.

[Tecuci, 1998] Tecuci G., *Building Intelligent Agents*, Academic Press, 1998

[Yarowsky, 1995] Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189-196, Cambridge, MA.

Summary

This paper explains how a combination of the Apprentice Systems, together with Instance-Based Learning, that we call 'Extensional Induction' enables a partially supervised learning. It starts with a (small) set of examples of the classes of interest and, using a technique similar to the one of the k-nearest neighbors, agglomerates new instances of the class in order to complete an extensional definition of the class.