RESUM DE LA THSE de DOCTORAT

INFORMATIQUE

Prsente par

Monsieur AUGIER Sébastien No Tl : 01.69.15.64.09

Thse prpare au Laboratoire de Recherche en Informatique

Sous la Direction de M. KODRATOFF Yves
Son directeur de Recherches.

SOUTENANCE DE THSE AYANT POUR JURY

M. Didier FAYARD
M. Yves KODRATOFF
M. Jacques NICOLAS
M. Marc SCHOENAUER
M. Gilles VENTURINI

SUJET :
Apprentissage Supervisé Relationnel par Algorithmes d'Évolution

DATE, HEURE ET LIEU DE SOUTENANCE SUR ORSAY :
19 dcembre 2000, 14h30, Salle 79, LRI, Bt 490.

Cette thèse décrit le système $\mathcal{SIAO}1$, fondé sur les algorithmes évolutionnaires, qui apprend des concepts relationnels ou propositionnels de nature symbolique ou numérique à partir d'exemples et de contre-exemples.

Le premier chapitre introduit les principaux concepts et algorithmes relatifs à l'apprentissage automatique supervisé. À cause de la multitude des approches et des travaux réalisés, il ne vise pas à l'exhaustivité mais s'attache tout d'abord à situer le domaine à la fois dans une perspective historique qui le lie à l'intelligence artificielle ainsi que dans le contexte actuel qui est caractérisé par un développement très important des applications industrielles (systèmes d'aide à la décision, *data mining*, etc.). Le corps du chapitre alterne d'une part la présentation des principales problématiques et résultats théoriques (Qu'est-ce que l'induction ? Comment formuler le problème dans le cadre de l'informatique ? Quelles sont les propriétés des objets manipulés ?), et d'autre part les principaux algorithmes d'apprentissage ainsi que leurs caractéristiques. Une dernière partie est plus spécifiquement consacrée à la programmation logique inductive (PLI), car ce domaine fournit un cadre plus précis et de nombreux résultats tant théoriques que pratiques, permettant de mieux situer notre travail.

Le second chapitre a pour ambition de présenter brièvement les algorithmes d'évolution ainsi que les principaux travaux concernant leur application dans le cadre de l'apprentissage automatique symbolique supervisé. En conséquence, une première partie est consacrée à la description des grandes classes de méthodes évolutionnaires que sont les algorithmes génétiques, les stratégies d'évolution, la programmation évolutionnaire et la programmation génétique. Toutes ces classes sont présentées comme des variations d'un même algorithme générique de simulation de l'évolution de populations d'individus. La seconde partie, dédiée à l'application de ces algorithmes aux problèmes d'apprentissage automatique symbolique supervisé, décrit les principales approches qui proviennent des algorithmes génétiques ainsi que de la programmation génétique. Elle est complétée par une présentation plus approfondie de travaux spécifiques, particulièrement importants ou représentatifs, auxquels il est fait référence au cours de cette thèse, tels que les systèmes GABIL, G-NET, REGAL et SIA.

Le chapitre 3 est dédié à l'étude du biais de langage de $\mathcal{SIAO}1$, en particulier du point de vue de la complexité du test de couverture (appelé également test de subsomption). En effet, comme l'a montré le chapitre 1, la prise en compte du dilemme coût de l'induction / richesse du langage est fondamentale dans le contexte de l'apprentissage relationnel. La structure de l'argumentation est donc divisée en deux grandes parties interdépendantes.

La première partie est centrée autour du langage de représentation de $\mathcal{SIAO}1$, noté $\mathcal{L}_h$. Elle débute par une présentation de la problématique et des résultats théoriques relatifs à la logique du premier ordre en général et au test de θ-subsomption utilisé par certaines méthodes de PLI en particulier. Une seconde section présente le langage de $\mathcal{SIAO}1$ ainsi que la relation de subsomption choisie pour structurer l'espace et notée $\vdash_2$. L'expressivité du langage ainsi que la complexité du test de subsomption sont également analysées au regard des résultats obtenus dans le cadre de la logique du premier ordre.

La seconde partie concerne l'étude et l'implantation d'un test de subsomption optimisé. Après avoir montré l'intérêt d'un tel test pour $\mathcal{SIAO}1$ dans le cas de problèmes fortement relationnels, l'architecture générale de l'algorithme de subsomption est esquissée, chacune des trois étapes (le pré-traitement, la réduction des domaines et la résolution avec propagation) étant détaillée dans une section spécifique. Finalement, les performances du test de subsomption sont validées sur un problème fortement relationnel et comparées avec les autres approches faisant référence dans le domaine, comme par exemple le test proposé dans le cadre de la PLI par T. Scheffer, qui reformule le problème de la subsomption comme un problème de recherche d'une clique dans un graphe après élagage de l'espace de recherche au moyen du contexte de graphe.

À l'issue de ce chapitre, le lecteur a pris connaissance des structures et algorithmes de base sur lesquels le système $\mathcal{SIAO}1$ est construit. En particulier, les choix qui ont été faits au niveau des biais de langage (représentation des connaissances, théorie du domaine, typage, complexité du test de subsomption) ont été explicités.

Le chapitre 4 décrit le système $\mathcal{SIAO}1$ qui apprend, à partir d'exemples, des définitions exprimées dans le langage relationnel $\mathcal{L}_h$.

La première section, qui s'appuie sur les résultats mentionnés dans le chapitre 3, présente l'architecture générale du système ainsi que la structure des chromosomes codant les formules. Les différentes composantes de la fonction de qualité fondée sur l'estimateur de Laplace ainsi que sur des biais d'intelligibilité et de concision sont ensuite explicitées, ce qui permet d'analyser le biais d'évaluation de l'algorithme et de montrer son adéquation à la stratégie de recherche.

La section suivante est consacrée aux opérateurs d'évolution (mutation et croisement), et souligne les différences importantes qui, combinées à la stratégie de recherche adoptée, distinguent nettement le système proposé des autres approches évolutionnaires : le biais de recherche de $\mathcal{SIAO}1$ est en effet caractérisé au niveau des opérateurs d'évolution par deux grands principes :

Les opérateurs sont ascendants et procèdent par petits pas . Une première caractéristique de $\mathcal{SIAO}1$ est l'exploration dirigée du treillis défini par la relation de généralité $\vdash_2$. Cette prise en compte de la structure de l'espace de recherche, associée au schéma de sélection, assure une exploration efficace puisque les cycles et les redondances sont rendus impossibles. Par contre, cette même propriété suppose que les opérateurs puissent parfois générer des individus suffisamment proches de leurs parents afin de fournir une solution de bonne qualité : puisqu'aucun retour en arrière n'est possible, l'algorithme doit être capable de se rapprocher de la solution par petits pas ascendants dans le treillis.

Les opérateurs sont dirigés par les données. Une seconde caractéristique de $\mathcal{SIAO}1$ vient de l'utilisation, par les opérateurs, de données provenant de la base d'apprentissage, qu'il s'agisse de la structure relationnelle des exemples ou de la distribution des symboles en fonction des classes des instances dans lesquelles ils apparaissent.

En particulier, la structure des hypothèses n'est pas limitée à la forme des exemples ou à un modèle fourni par l'utilisateur.

Les deux sections suivantes sont consacrées d'une part au post-traitement et à la classification de nouvelles instances à l'aide des règles obtenues, et d'autre part aux connaissances du domaine d'ordre symbolique ou numérique pouvant être prises en compte par le système afin de fournir des définitions plus précises et plus intelligibles. Finalement, une dernière partie présente les résultats obtenus sur différentes bases d'apprentissage de référence (*Adult, Census-income, Iris, Mushrooms, Mutagenesis, 1000-Trains*). Les résultats obtenus montrent que $\mathcal{SIAO}1$ se compare favorablement aux autres systèmes. De plus, $\mathcal{SIAO}1$ est polyvalent et sollicite peu l'utilisateur en ce qui concerne des questions assez techniques telles que les biais de recherche ou d'évaluation.

Le chapitre 5 est relatif à l'application des algorithmes d'apprentissage (en particulier évolutionnaires) à des problèmes de grande taille. Cette problématique, particulièrement importante pour l'ECD (extraction de connaissances à partir des données) est abordée à la fois du point de vue du changement d'échelle en temps (apprendre plus vite sur les données) et en espace (apprendre sur des volumes de données plus importants). Dans ce contexte, deux architectures parallèles implantées afin de permettre à $\mathcal{SIAO}1$ de changer d'échelle tout en conservant le fonctionnement de l'algorithme sont validées.

Dans le cas où l'accroissement de la complexité est dû à une forte composante relationnelle du problème, ou lorsque la base d'apprentissage est de taille moyenne (500 à 20 000 exemples), l'architecture de type maître-esclave doit être préférée. En effet, d'une part son asynchronisme lui permet de s'adapter à la variation des durées d'évaluation, et d'autre part il a été montré qu'elle est plus performante dans ce contexte d'utilisation.

Par contre, lorsqu'on se place dans une perspective de montée en charge afin de traiter de grandes bases d'apprentissage, dont la taille est éventuellement supérieure à la mémoire disponible sur une machine, on utilisera l'architecture basée sur un pipeline puisqu'elle répartit les données sur les processeurs. De plus, les temps d'évaluation devenant très importants, le problème du débit du réseau n'est plus un obstacle et les performances obtenues approchent la limite théorique prédite par la loi d'Amdahl.

Finalement, un modèle simple et précis des performances des architectures est proposé. Sa mise en œuvre permet d'évaluer plus précisément les capacités de changement d'échelle de $\mathcal{SIAO}1$ par le biais du comportement de l'iso-efficacité.

Le dernier chapitre résume les apports de la thèse et présente les perspectives possibles du travail. Les annexes qui suivent sont consacrées aux bases d'apprentissage utilisées ainsi qu'à des notions non rappellées dans le corps du manuscrit (logique relationnelle, treillis et ordres partiels, récapitulatif des notations utilisées, etc.).