Un système cellulaire neuro-symbolique pour l'extraction et la gestion des connaissances

Baghdad Atmani*, Mohamed Benamina*, Bouziane Beldjilali

*Equipe de recherche Simulation, Intégration et Fouille de données « SIF »
Laboratoire d'Informatique d'Oran « LIO »
Département Informatique, Faculté des Sciences, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie
atmani.baghdad@gmail.com, benamina.mohamed@gmail.com, bouzianebeldjilali@yahoo.fr

Résumé. Le CNSS – Cellular Neuro-Symbolic System – est un système hybride ralliant conjointement le neuro-symbolique et le cellulaire. CNSS permet, à partir d'une base de cas pratique, de faire coopérer un réseau de neurones, un graphe d'induction et un automate cellulaire pour la construction d'un modèle de prédiction. En détectant et en éliminant les individus non applicables et les variables non pertinentes, le réseau de neurones optimise la base d'apprentissage. Le résultat ainsi obtenu est affiné par un processus d'apprentissage symbolique à base de graphe d'induction. Ce raffinement se fait par une modélisation booléenne qui va assister l'apprentissage symbolique à optimiser le graphe d'induction et va assurer, par la suite, la représentation et la génération des règles de classification sous forme conjonctives avant d'entamer la phase de déduction par un moteur d'inférence cellulaire. CNSS a été testé sur plusieurs applications en utilisant des problèmes académiques et réels. Les résultats montrent que le système CNSS a des performances supérieures et de nombreux avantages.

1 Introduction

La réalisation de systèmes hybrides est une démarche courante qui permet de combiner les points forts de deux ou plusieurs approches, et d'obtenir ainsi des performances plus élevées et/ou un champ d'application plus large. En s'inspirant du système INSS – Incremental Neuro-Symbolic System – (Osorio, 1998) nous avons développé des recherches sur les systèmes d'apprentissage automatique numériques et symboliques, et en particulier sur l'acquisition automatique incrémentale de règles de production à partir de connaissances empiriques (exemples). Un nouveau système hybride, nommé CNSS, a été étudié et réalisé. Ce système permet, à partir d'une base de cas pratiques, de faire coopérer un réseau de neurones, un graphe d'induction et un automate cellulaire pour la construction d'un modèle de prédiction. En détectant et en éliminant les individus non applicables et les variables non pertinentes, le réseau de neurones optimise l'échantillon d'apprentissage. Le résultat du réseau de neurones ainsi obtenu, est affiné par un processus d'apprentissage automatique symbolique à base de graphe d'induction. L'optimisation du graphe d'induction généré par le principe de la méthode SIPINA – Système Interactif Pour l'Induction Non Arborescente – (Zighed et Rakotomalala, 2000) se fait par l'automate cellulaire qui va générer un graphe d'induction cellulaire et assurer, par la suite, la représentation et la génération des règles de

production sous formes conjonctives avant d'entamer la phase de validation par un système expert cellulaire.

Cet article est structuré comme suit. La section 2 est consacrée à la présentation du système CNSS où nous détaillons son architecture et son mode de fonctionnement. La section 3 est dédiée aux résultats expérimentaux, une conclusion et quelques perspectives.

2 Le système CNSS

Le système CNSS, a eu comme point de départ la machine cellulaire ACSIR – Automate Cellulaire pour des Systèmes d'Inférences à base de Règles – développée dans le cadre d'un mémoire de Magister au sein de l'université d'Oran (Atmani, 1996). Depuis, nous avons essayé d'améliorer les points faibles de cette machine cellulaire en ajoutant de nouvelles propriétés. Les principales améliorations du CNSS par rapport à la machine cellulaire ACSIR, sont la sélection des variables exogènes par apprentissage connexionniste, la sélection des individus non applicables par validation connexionniste, l'acquisition automatique de connaissances par graphe d'induction cellulaire et la validation des nouvelles connaissances (règles) acquises par moteur d'inférence cellulaire.

Le système CNSS constitue une nouvelle approche de système d'apprentissage automatique par induction. L'apprentissage automatique par induction (Quinlan, 1986) est un apprentissage empirique qui tend à produire des règles générales à partir d'une série d'observations. Ce processus d'induction peut s'insérer dans des démarches plus générales d'extraction et de gestion de connaissances, ou de prédiction. On distingue principalement l'extraction de connaissances à partir de données et l'alimentation des Systèmes Experts.

Une des principales composantes d'un Système Expert est la base de connaissances, que l'on assimile souvent à une base de règles. Traditionnellement, ces bases sont construites par un Expert du domaine qui, partant de son expérience et de ses connaissances, propose des règles de production. L'introduction de l'apprentissage supervisé, et notamment de l'induction des règles à partir d'exemples, a permis d'atténuer deux contraintes procédurales (deux goulots d'étranglement) dans la construction des Systèmes Experts (Osorio, 1998) : d'un coté l'accélération du processus d'acquisition des connaissances et, d'un autre coté une meilleure fiabilité face aux règles d'Experts, surtout une fiabilité qu'on a l'avantage de quantifier.

Le système CNSS est composé de deux modules principaux, Neuro-IG (Atmani et Beldjilali, 2007a) et CASI – Cellular Automata for Symbolic Induction – (Atmani et Beldjilali, 2007b) et, de trois modules secondaires de conversion et de transfert de connaissances (voir figure 1).

Réseau de neurones. C'est le module connexionniste, proprement dit, du système CNSS. Il est responsable, après élagage et validation, de la sélection des variables exogènes pertinentes et de l'élimination des individus non applicables.

Convertisseur neuro-symbolique –CSN–. Il est responsable de la communication avec le réseau de neurones. Il assure la présentation des échantillons d'apprentissage Ω_{A}^{RN} et de test Ω_{T}^{RN} au réseau de neurones et la récupération de l'échantillon des individus non classés Ω_{E}^{RN}. D'autre part, il permet aussi la proposition des variables non pertinentes.

Convertisseur cellulo-graphe –CGIC–. Il est responsable de la communication avec la machine cellulaire. Il assure la codification et la présentation des échantillons d'apprentissage Ω_A^{SS} et de test Ω_T^{SS} à la machine cellulaire et la récupération des résultats.

La machine cellulaire. C'est le module cellulaire, proprement dit, du système CNSS. Il est responsable, après apprentissage automatique cellulo-symbolique (principe de la méthode SIPINA), de l'optimisation du graphe cellulaire et de la génération des règles conjonctives par chaînage arrière. D'autre part, il permet la validation du modèle cellulaire (ensemble de règles conjonctives) par déduction en chaînage avant.

Convertisseur cellulo-symbolique –CCS–. Il sert d'interface de communication entre l'utilisateur ou l'expert et la machine cellulaire.

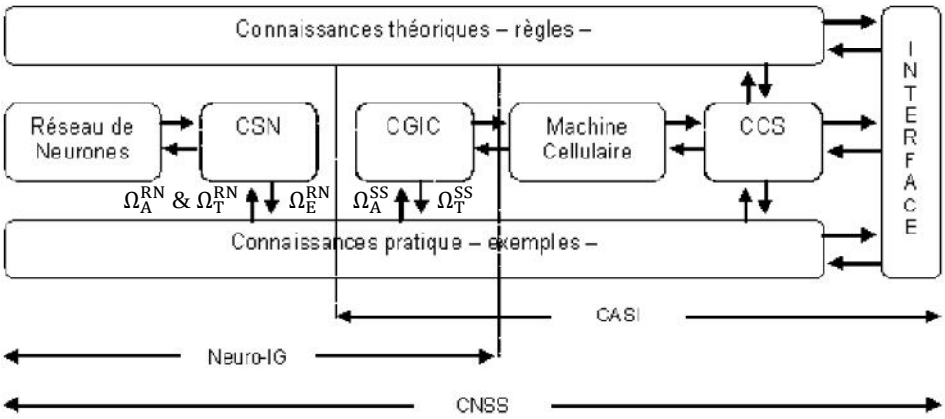

FIG. 1 – *Schéma général du système hybride CNSS.*

2.1 Apprentissage supervisé par induction

Le processus général d'apprentissage automatique par induction que CNSS applique à une population Ω est organisé selon les étapes suivantes :

- Acquisition, transformation et préparation des données par l'interface ;
- Présentation des échantillons d'apprentissage Ω_A^{RN} et de test Ω_T^{RN} au réseau de neurones en utilisant le module CSN ;
- Elaboration d'un modèle de classification numérique et prétraitement des données par réseau de neurones (sélection des variables et des individus) ;
- Récupération de l'échantillon des individus non classés Ω_E^{RN} et proposition des variables non pertinentes en utilisant toujours le module CSN ;
- Codification et présentation des échantillons d'apprentissage Ω_A^{SS} et de test Ω_T^{SS} à la machine cellulaire en utilisant le module CGIC ;
- Elaboration d'un modèle de classification symbolique à base de graphe d'induction (SIPINA) ;
- Récupération et interprétation des règles par le module CGIC ;
- Extraction et validation des règles symboliques par le CCS ;
- Interprétation et généralisation par le CCS.

2.2 Classification numérique et minimisation des connexions

Pour l'élaboration du modèle connexionniste nous avons utilisé un perceptron à trois couches et la méthode de descente de gradient à segment nul proposée par Atmani et Beldjilali (2007a). Généralement, la phase d'apprentissage d'un réseau de neurones perceptron multicouches commence par un premier ensemble de poids (w, v) initialisé aléatoirement en utilisant un générateur, et met à jour itérativement ces poids pour réduire au minimum la fonction globale $E_{app}(w,v) + E_{min}(w,v)$. L'apprentissage du réseau est terminé quand le gradient de la fonction est suffisamment petit. Pour produire le modèle de classification numérique nous avons adopté les étapes proposées par Setiono (1996).Une première phase d'apprentissage permet de déterminer les poids de raccordements d'un réseau possédant seulement une couche cachée avec un nombre arbitraire de neurones. Par une méthode de minimisation, le réseau obtenu est simplifié en éliminant les connexions avec les plus petits poids, tout en restant dans un bassin de solutions. Un nouvel apprentissage est lancé pour les raccordements utiles restants. Le processus d'optimisation s'arrête selon un ou plusieurs critères de satisfaction guidés par l'expérimentation et fixés, en général, par l'utilisateur. Enfin, une validation, après élagage, du réseau optimal qui contient seulement les raccordements estimés pertinents.

2.3 Classification cellulo-symbolique

La machine cellulaire CASI – Cellular Automata for Symbolic Induction – est composée de trois modules : COG (Cellular Optimization and Generation), CIE (Cellular Inference Engine) et CV (Cellular Validation) (Atmani et Beldjilali, 2007b). Le module CIE, cœur de la machine cellulaire CASI, simule le fonctionnement du cycle de base d'un moteur d'inférence en utilisant deux couches finies d'automates finis. La première couche, CELFACT, pour la base des faits et la deuxième couche, CELRULE, pour la base de règles. Chaque cellule au temps *t+1* ne dépend que de l'état des ses voisines et du sien au temps *t*. Dans chaque couche, le contenu d'une cellule détermine si et comment elle participe à chaque étape d'inférence : à chaque étape, une cellule peut être active (1) ou passive (0), c'est-à-dire participe ou non à l'inférence.

Supposons que les règles générées par apprentissage symbolique (SIPINA) sont : R1) *A et B*$\rightarrow$ *C* ; R2) *F et D*$\rightarrow$ *A* ; R3) *D et E*$\rightarrow$ B ; R4) *B et D*$\rightarrow$ *F* ; R5) *E et F*$\rightarrow$ *D* ; R6) *E et F*$\rightarrow$ *B* ; R7) *B et F*$\rightarrow$ *G* ; A partir de cette base de connaissances le module CIE initialise les deux couches CELFACT et CELRULE comme le montre la figure 2 :

CELFACT(Faits)	*EF*	*IF*	*SF*
A	0	1	0
B	0	1	0
C	0	1	0
D	0	1	0
E	1	1	0
F	1	1	0
G	0	1	0

CELRULE (Règles)	*ER*	*IR*	*SR*
R_1	0	1	1
R_2	0	1	1
R_3	0	1	1
R_4	0	1	1
R_5	0	1	1
R_6	0	1	1
R_7	0	1	1

FIG. 2 – *Représentation cellulaire de la Base des connaissances des 7 règles.*

De même le voisinage de CIE est défini par les matrices d'incidence d'entrée R_E et de sortie R_S illustrées par la figure 3.

$\boldsymbol{R_E}$	R_1	R_2	R_3	R_4	R_5	R_6	R_7
A	1						
B	1			1			1
C							
D		1	1	1			
E			1		1	1	
F		1			1	1	1
G							

$\boldsymbol{R_E}$	R_1	R_2	R_3	R_4	R_5	R_6	R_7
A		1					
B			1			1	
C	1						
D					1		
E							
F				1			
G							1

FIG. 3 – *Les matrices d'incidence d'entrée R_E et de sortie R_S des 7 règles.*

La dynamique du moteur d'inférence cellulaire *CIE* utilise deux fonctions de transitions δ_{fact} et δ_{rule}, où δ_{fact} correspond à la phase d'*évaluation*, de *sélection* et de *filtrage*, et δ_{rule} correspond à la phase d'*exécution*.

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, \boldsymbol{EF}, \boldsymbol{ER} + (\boldsymbol{R_E^T} \cdot \boldsymbol{EF}), IR, SR)$$

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (\boldsymbol{EF} + (\boldsymbol{R_S} \cdot \boldsymbol{ER}), IF, SF, ER, IR, \overline{\boldsymbol{ER}})$$

Où la matrice R_E^T désigne la transposée de la matrice R_E.

3 Résultats expérimentaux, conclusion et perspectives

Pour évaluer le système CNSS nous avons utilisé la plateforme TANAGRA (Rakotomalala, 2005), en particulier le couplage de la méthode de sélection MIFS (Battiti, 1994) avec plusieurs différentes méthodes à base d'arbres de décision pour l'induction symbolique. Le système CNSS a été testé sur plusieurs applications, en utilisant différentes bases d'exemples : Crédit, Ulcère, Diabète, Cancer du sein, Titanic, etc. Les résultats obtenus ont montré que CNSS possède des propriétés intéressantes et de nombreux avantages par rapport aux autres approches neuro-symboliques testées séparément : une classification avec un taux de succès de 93%, une optimisation de la base d'exemples de plus de 17%, une réduction du nombre de variables descriptives (exogènes), une réduction de la taille du graphe d'induction, généré par SIPINA, de plus de 15%, et une amélioration du temps de réponse (validation par déduction cellulaire).

Ces résultats forts probants nous permettent de souligner que le système CNSS possède plusieurs points avantageux qui sont :

- la capacité d'intégrer des connaissances théoriques (règles) et des connaissances empiriques (exemples) avec l'utilisation d'algorithmes très performants ;
- la capacité de sélectionner et de proposer les variables exogènes non pertinentes ;
- la capacité de proposer à l'élimination les exemples (individus) non applicables ;
- la capacité d'acquérir par apprentissage automatique (principe SIPINA) d'extraire et de valider les nouvelles connaissances sur un problème réel très complexe, avec une très bonne performance (modélisation booléenne) ;
- la capacité de coder des règles de haut niveau (règles d'ordre 0+) dans la machine cellulaire CASI ;

- la capacité de traiter les valeurs d'entrée discrètes ainsi que les valeurs continues ;

Dans cette étude nous avons mis l'accent sur l'apport intrinsèque d'une forte collaboration de trois systèmes largement éprouvés. Par rapport à ses antagonistes, le CNSS se présente, aujourd'hui, comme un produit assez performant pouvant apporter des solutions intéressantes d'un point de vue optimisation des performances et de coûts de réalisation des systèmes experts. CNSS est utilisé actuellement pour la construction de la base de connaissance du système d'information décisionnel de vaccination. L'objectif est la réalisation d'une architecture orientée service pour le programme élargi de vaccination.

4 Références

Atmani, B. (1996). Automate Cellulaire pour des Systèmes d'Inférence à base de Règles (ACSIR), Thèse de Magister en Informatique. Université Es-Senia, Oran, 1996.

Atmani, B., Beldjilali, B. (2007a). Neuro-IG : A Hybrid System for Selection and Elimination of Predictor Variables and non Relevant Individuals. *Informatica, Journal International*, Vol. 18, N°2 163-186.

Atmani, B., Beldjilali, B. (2007b). Knowledge Discovery in Database: Induction Graph and Cellular Automaton. *Computing and Informatics Journal*, Vol.26, N°2 171-197.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning, IEEETrans. Neural Networks, 5, 4, 1994, 537-550.

Clark, P. (1989). Knowledge representation in machine learning, Machine and Human Learning, Eds : Y. Kodratoff and A. Hutchison, London.

Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning.

Osorio, F.S., Amy, B. (1999). INSS : A hybrid system for constructive machine learning, Neurocomputing, Vol. 28, 1, Oct 1999, 59-67.

Setiono, R. (1996). Extraction rules from pruned neural networks for breast cancer diagnosis, 8, 1, February 1996, 37-51.

Rakotomalala, R. (2005). TANAGRA : Une Plate-Forme d'Expérimentation pour la Fouille de Données, Revue MODULAD, 32, 70-85, 2005.

Zighed, D.A., Rakotomalala, R. (2000). Graphs of induction, Training and Data Mining, Hermes Science Publication, 21-23.

Summary

The CNSS - Cellular Neuro-Symbolic System - is a hybrid system jointly endorsing the neuro-symbolic and cellular automata. CNSS enables, from a practical database, to cooperate a neural network, an induction graph and a cellular automata to build a predictive model. By detecting and eliminating not applicable individuals and irrelevant variables, the neural network optimizes the learning base. The result thus obtained is refined by a symbolic learning process based induction graph. This refinement is done by Boolean modeling will attend the symbolic learning to optimize the graph and, subsequently, the representation and generation of classification rules as conjunctive before entering the phase of deduction by an cellular inference engine. CNSS has been tested on several applications using real and academic problems. The results show that the system CNSS has superior performance and many advantages.