

Conception et implémentation d'une nouvelle technique cellulaire de discrétisation : intégration dans TANAGRA

Baghdad Atmani et Mohamed Benamina

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

1 Introduction

Dans le domaine de l'Extraction de Connaissances à partir de Données (ECD), beaucoup de méthodes de fouille de données telles que les règles d'association, les réseaux bayésiens ou les graphes d'induction ne peuvent manipuler que des attributs discrets. Discrétiser un attribut numérique consiste à découper son domaine de valeurs en un nombre fini d'intervalles, qui seront identifiés par un code (Dougherty et al., 1995). Dans ce papier nous proposons une nouvelle technique de discrétisation, baptisée DSynchro, qui utilise pour choisir les points de coupure du domaine de valeurs le principe de Mazoyer adopté pour la synchronisation par automate cellulaire. Le but est d'optimiser l'indicateur de qualité du partitionnement et réduire le temps de calcul. Les résultats obtenus par DSynchro sont comparés à ceux des techniques de discrétisation, déjà intégrées dans la plateforme de fouille de données TANAGRA (Rakotomalala, 2005), et s'avèrent être plus satisfaisants.

2 Principe de DSynchro

La synchronisation par automate cellulaire a été utilisée par Yunes (1993) pour étudier le problème de la synchronisation d'une ligne de fusiliers, et notamment, le problème ouvert concernant l'existence ou non d'un automate de synchronisation dont l'ensemble des états ne contiendrait que cinq éléments. Deux approches sont énoncées: 1) une première (descendante et constructive) montre comment, après avoir construit un automate avec treize états, il est possible de diminuer la cardinalité de l'ensemble par des codages successifs. Le procédé de Minsky (divide-and-conquer) est la base de ces solutions; 2) une deuxième (ascendante et analytique) étudie certains comportements observés dans quelques automates n'ayant qu'un nombre très réduit d'états. Parmi les différentes méthodes de synchronisations proposées dans ce domaine (Yunes, 1993) nous avons expérimenté la méthode de Mazoyer. La synchronisation obtenue par cette méthode est une synchronisation en temps minimale. La différence principale par rapport aux autres solutions de Minsky et de Waskman-Blazer et qu'il n'y a pas d'image miroir. En fait il s'agit de couper aux $2/3$, puis $4/9$, $8/27$, etc.

Pour déterminer les meilleurs points de coupure pour les variables continues DSynchro utilise la synchronisation par automate cellulaire. Le principe de cette méthode est basé sur les résultats de découpage obtenu par la méthode de synchronisation Mazoyer où les cellules sont représentées par un nouvel attribut virtuel, obtenu à partir de la plus petite et la plus grande valeur de la variable continue à discrétiser, avec un pas de 1.