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.

Algorithme de DSynchro.

1. Initialisation en classant les exemples d'apprentissage par ordre croissant des valeurs de l'attribut X à discrétiser ;
2. Sélectionner la plus petite et la plus grande valeur de l'attribut X ;
3. Construire un nouvel attribut virtuel, en allant de la plus petite valeur jusqu'à la plus grande valeur avec un pas de 1. On obtient alors un attribut dont les valeurs se succèdent, et sa taille est égale a N (N = plus grande valeur – plus petite valeur + 1) ;
4. On sélectionne les points de coupure de l'attribut virtuel aux points 2/3 puis 4/9, 8/27…jusqu'à ce qu'il ne reste plus de possibilité de découpage. Ainsi on obtient des intervalles avec les points de coupure trouvés ;
5. On code les valeurs de l'attribut X, selon leurs appartenances aux intervalles obtenus auparavant.

Pour évaluer la méthode DSynchro nous avons utilisé la plateforme TANAGRA, en particulier différentes méthodes à base d'arbres de décision pour l'induction symbolique. DSynchro a été testé sur plusieurs applications en utilisant différentes bases d'exemples. Les résultats obtenus ont montré que la discrétisation par automate cellulaire possède des propriétés intéressantes et de nombreux avantages par rapport aux autres techniques de discrétisation. Le tableau 1 résume les résultats obtenus avec une validation croisée.

Bases d'exemples	Nbr Var Continue	MDLPC	EqFreq	EqWidth	DSynchro
adult	6	0.1401	0.1661	0.1670	0.1525
autos	15	0.3707	0.5463	0.5463	0.4976
heart	6	0.2481	0.2481	0.2481	0.2481
iris	4	0.6667	0.6667	0.6667	0.6667
waveform	21	0.2488	0.2576	0.2370	0.2408
breast	9	0.0758	0.0873	0.0687	0.0672
weather	2	0.3571	0.3571	0.3571	0.3571
wine_quality	4	0.6471	0.6471	0.6471	0.6471

TAB. 1 – *Evaluation de la méthode DSynchro.*

Références

Dougherty, J., Kohavi, R., Sahami, M. (1995). Supervised and unsupervised discretization of continuous attributes. In Morgan Kaufmann, editor, Machine Learning: Proceedings of the Twelth International Conference (ICML-95), pages 194-202.

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

Yunes, J-B. (1993). Synchronisation et automates cellulaires ; la ligne de fusiliers, Thèse de Doctorat, Université de Paris 07, Paris, France.

Summary

In this paper we propose a new discretization technique, called DSynchro, which uses to select the break points in the domain of values the Mazoyer principle based on synchronization by cellular automata. The goal is to optimize the quality indicator of partitioning and reduce the computation time.