Document de travail

Laura Nguyen

20 juillet 2018

1 Introduction

Dans beaucoup de problèmes de classification, les valeurs des attributs et de la classe sont ordinaux. De plus, il peut exister une contrainte de monotonie : la classe d'un objet doit croître/décroître en fonction de la valeur de tout ou partie de ses attributs. A savoir, étant donné deux objets x, x', si $x \leq x'$ alors $f(x) \leq f(x')$. Les variables dépendantes, $f(x)$ et $f(x')$, sont des fonctions monotones des variables indépendantes, x et x'. On parle alors de problèmes de classification monotone, ou problèmes de classification avec contrainte de monotonie. Cette contrainte indique que les objets ayant de meilleures valeurs d'attributs ne doivent pas être assignés à de moins bonnes valeurs de classe.
L'ajout de cette contrainte de monotonie permet d'introduire des concepts sémantiques tels la préférence, la priorité, l'importance, qui nécessitent une relation d'ordre.
Il existe de nombreux domaines se prêtant à ce type de tâches, tels la prédiction du risque de faillite [5], l'analyse de la satisfaction des clients [6], le diagnostic médical [8]. L'importance de la prise en compte d'une relation graduelle entre les valeurs d'attributs et la classe a été démontrée [10] : les classifieurs auxquels sont imposés la contrainte de monotonie sont au moins aussi performants que leurs homologues classiques, et les experts sont plus enclins à utiliser les règles générées par les modèles monotones.
Afin d'extraire des règles à partir de données monotones, on décide d'utiliser les arbres de décision, dont l'efficacité et l'interprétabilité en classification a été prouvée [11]. Cependant, les algorithmes de construction d'arbres de décision standards (générés par CART [7]) ne produisent pas de classifieurs sensibles à la monotonie, même si la base utilisée est complètement monotone. En revanche, il est montré dans [3] que les classifieurs purement monotones ([2], [1], [4]) sont, en terme de taux de bonne classification, statistiquement indiscernables de leurs homologues non-monotones. Dans le même article, il est expliqué que ce phénomène est dû à la sensibilité de ces classifieurs au bruit non-monotone présent dans les données réelles.

Ce stage a pour but d'étudier la construction et l'évaluation d'arbres de décision prenant en compte une relation graduelle susceptible d'exister entre les valeurs d'attributs et la classe, tout en étant suffisamment robuste au bruit non-monotone. On reprend, en particulier, [9] pour la construction d'arbres de décision monotones paramétrés par une mesure de discrimination à rang. Une étude théorique des propriétés des mesures présentées dans le même article est également effectuée.

2 Implémentation et expérimentation de l'algorithme de construction d'arbres monotones

Dans cette partie, on implémente (à quelques variantes près) RDMT(H), l'algorithme de construction d'arbres monotones donné dans [9] et on l'évalue sur des données artificielles et réelles.

2.1 Implémentation des mesures de discrimination à rang

D'après [9], les mesures de discrimination à rang possèdent la même structure fonctionnelle : elles se décomposent en trois fonctions. Dans le même article, un modèle de construction hiérarchique de mesures de discrimination à rang est proposé. Il permet d'isoler leurs propriétés et d'en

créer de nouvelles.

Références

[1] A. Ben-David. "Monotonicity maintenance in information-theoretic machine learning algorithms". In : *Machine Learning* 19.1 (1995), p. 29-43.

[2] A. Ben-David, L. Sterling et Y.-H. Pao. "Learning and classification of monotonic ordinal concepts". In : *Computational Intelligence* 5.1 (1989), p. 45-49.

[3] A. Ben-David, L. Sterling et T. Tran. "Adding monotonicity to learning algorithms may impair their accuracy". In : *Expert Systems with Applications* 36.3 (2009), p. 6627-6634.

[4] K. Cao-Van et B. De Baets. "Consistent representation of rankings". In : *Theory and Applications of Relational Structures as Knowledge Instruments*. Springer, 2003, p. 107-123.

[5] S. Greco, B. Matarazzo et R. Slowinski. "A new rough set approach to evaluation of bankruptcy risk". In : *Operational tools in the management of financial risks*. Springer, 1998, p. 121-136.

[6] S. Greco, B. Matarazzo et R. Slowinski. "Customer satisfaction analysis based on rough set approach". In : *Zeitschrift für Betriebswirtschaft* 77.3 (2007), p. 325-339.

[7] B. Leo, J. H. Friedman, R. A. Olshen et C. J. Stone. "Classification and regression trees". In : *Wadsworth International Group* (1984).

[8] C. Marsala. "Gradual fuzzy decision trees to help medical diagnosis". In : *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*. IEEE. 2012, p. 1-6.

[9] C. Marsala et D. Petturiti. "Rank discrimination measures for enforcing monotonicity in decision tree induction". In : *Information Sciences* 291 (2015), p. 143-171.

[10] M. J. Pazzani, S. Mani et W. R. Shankle. "Acceptance of rules generated by machine learning among medical experts". In : *Methods of information in medicine* 40.05 (2001), p. 380-385.

[11] J. R. Quinlan. "Induction of decision trees". In : *Machine learning* 1.1 (1986), p. 81-106.