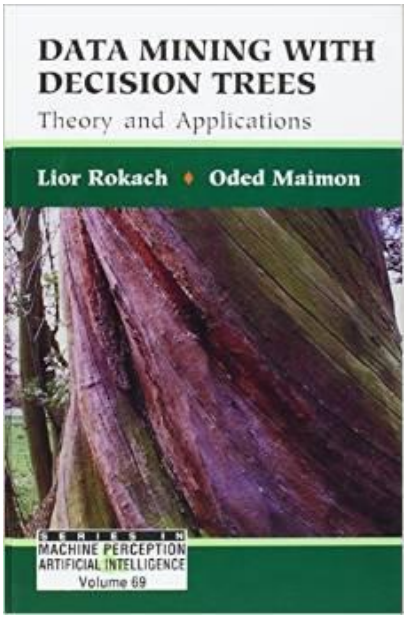
Data Mining with Decision Trees – Theory and Applications

Lior Rokach and **Oded Maimon**

Series in Machine Perception and Artificial Intelligence – Vol. 61

World Scientific Publishing, 2007



L'ouvrage de Rokach et Maimon fait un état des lieux des avancées autour de l'induction des arbres de décision. Les auteurs commencent par une description rapide de la nature du data mining et du processus de construction des arbres de décision (chapitres 1 – « Introduction to Decision Trees » et 2 « Growing Decision Trees »). On peut distinguer ensuite 3 grandes parties.

[1] La première détaille les 2 principaux éléments constitutifs de l'induction des arbres : le critère de segmentation (chapitre 4 – « Splitting Criteria ») et la détection de la bonne taille de l'arbre (chapitre 5 – « Pruning Trees »). Les innombrables variantes qui ont été publiées sur ces deux thèmes sont énumérées, un peu rapidement peut être. Des exemples didactiques sur des petits jeux de données auraient rendu l'exposé plus accessible, surtout pour les personnes peu aux faits des subtilités des techniques de construction des arbres de décision.

[2] La seconde partie concerne les développements qui ont été menées autour des arbres de décision ces dernières années. Ils touchent à différents domaines. Les différentes méthodes qui font référence sont présentées (CART, C4.5, CHAID), ainsi que les techniques destinées à améliorer la qualité des arbres (chapitre 6 – « Advanced Trees »). Le chapitre 8, « Incremental Learning of Decision Trees », décrit les enjeux et les approches pour la construction incrémentale des arbres, enjeux fort s'il en est dans un contexte « big data » où les données arrivent en profusion, avec des mises à jour fréquentes. La construction des arbres flous est abordée dans le chapitre 10 « Fuzzy Decision Trees ». L'intérêt de l'approche est connu depuis longtemps. Elle prend en compte l'imprécision inhérente à la nature et au recueil des données. Elle permet entre autres de dépasser l'arbitraire des frontières « dures » (crisp) induites par la discrétisation des descripteurs continus en introduisant une gradation dans l'appartenance des individu aux différentes zones définies par l'arbre. Le chapitre 11, « Hybridization of Decision Trees with other Techniques », expose les principes de la combinaison des autres techniques d'apprentissage statistique avec les arbres. Différentes variantes sont possibles, la plus simple étant une procédure en deux temps : (1) les feuilles de l'arbre construit classiquement définissent une partition de l'espace de représentation ; (2) dans chaque sous-population (feuille), un modèle prédictif – basé sur une autre méthode statistique (une bayésien naïf par exemple) - est élaboré à l'aide de tout ou partie des descripteurs. D'autres approches, plus sophistiquées, sont possibles. La méthode statistique que l'on combine avec l'arbre peut intervenir dans la construction même de l'arborescence, notamment lors du choix des variables de segmentation. Enfin, le chapitre 12, « Sequence Classification using Decision Trees », présente l'application des arbres aux données séquentielles. Les applications sont nombreuses, on pense au text mining et à la gestion des phrases négatives. Les auteurs font référence à l'appréhension du problème difficile de l'analyse des comptes-rendus médicaux (section 12.6, page 193).

Les auteurs adoptent le discours « survey ». Ils rentrent rarement dans les détails et se contentent de passer rapidement en revue les différents thèmes. Le principal intérêt pour nous est de se faire une idée rapide de ce qui se fait de nouveau (ou pas) dans les différents domaines. Libre à nous par la suite d'approfondir telle ou telle question en effectuant des recherches sur le web. Je m'étais fait une idée plus ou moins claire des enjeux de l'analyse des données en flux par exemple (data stream mining). Suite à la lecture de chapitre 8, je me

suis penché de manière plus attentive à la notion de dérive conceptuelle (concept drive) et des solutions pour répondre à ce type de problème.

[3] La troisième partie est plus générique et touche plutôt à l'apprentissage supervisé de manière générale. Le chapitre 3, « Evaluation of Classification Trees », présente les procédures et critères utilisés pour l'évaluation des classifieurs. Ils ne sont pas spécifiques aux arbres et s'applique à toutes méthodes supervisées. Assez détaillée pour le coup (par rapport à la partie précédente), l'exposé dépasse les indicateurs numériques et discute des aspects moins quantifiables mais tout aussi important dans l'appréhension des méthodes (scalabilité, compréhensibilité, etc.). Ce chapitre dépasse largement le cadre de l'induction par arbres. Tout lecteur intéressé par l'apprentissage supervisé et son évaluation peut en tirer profit. Le chapitre 7, « Decision Forests », fait le tour des techniques ensemblistes. Ici également, les auteurs essaiment plus large, même si les arbres constituent une des méthodes privilégiées dans ce contexte. Mais surtout, au-delà de la simple description des différentes approches (ex. bagging, wagging, random forest, etc.), les auteurs structurent le chapitre de manière à ce que l'on distingue bien les idées qui sous-tendent la création des ensembles de classifieurs (ex. comment générer de la diversité parmi les classifieurs, quelles sont les différentes stratégies de combinaison des classifieurs, etc.). La section introductive du chapitre (section 7.2) est particulièrement intéressante. Elle positionne clairement le cadre et les enjeux des techniques ensemblistes. Le chapitre 9, « Feature Selection », aborde la réduction de la dimensionnalité au moyen de la sélection de variables. Ici également, il ne s'agit pas de parler des techniques de sélection pour les arbres puisque ces derniers intègrent déjà la sélection dans le processus de modélisation. Les auteurs décrivent plutôt les différentes approches de sélection en général. Je donnerais peut être une mention particulière à la synergie entre les méthodes ensemblistes et la sélection de variables (sections 9.4 et 9.5) que l'on retrouve peu dans la littérature.

L'ouvrage de Rokach et Maimon constitue avant tout un travail bibliographique important. Le discours est un peu trop condensé pour qu'on puisse réellement s'appuyer dessus pour s'initier à des techniques. Il nous permet surtout de nous repérer par rapport aux nombreux travaux qui ont porté sur les arbres de décision ces dernières années. J'ai vu qu'une version

mise à jour a été publiée en octobre 2014 (vol. 81), avec de nouveaux chapitres (dont un qui porte sur les logiciels apparemment). On va regarder tout cela attentivement prochainement.

Enfin, les mêmes auteurs ont publié un survey « Decision Trees » dans l'ouvrage collectif « The Data Mining and Knowledge Discovery Handbook », Maimon and Rokach, Springer, 2005, chapitre 9, pages 165 à 192. L'article est accessible en ligne. Il reprend en partie des éléments du livre (http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf).