
Fouille de données spatiales par arbre de décision multi-thèmes

Nadjim Chelghoum* — Karine Zeitouni* — Azedine Boulmakoul**

*Laboratoire PRISM - Université de Versailles Saint-Quentin,
45, Av. des Etats-Unis, 78035 Versailles Cedex, France, Prénom.Nom@prism.uvsq.fr

**LIST, Faculté des Sciences et Techniques de Mohammedia,
B.P. 146 Mohammedia, Maroc, boul@uh2m.ac.ma

RÉSUMÉ. La fouille de données spatiales est aujourd'hui un domaine bien identifié de la fouille de données. Cet article s'intéresse à la classification de données spatiales par arbre de décision. Cette méthode se différencie des arbres de décisions traditionnels par la prise en compte de l'organisation en couches thématiques et des relations spatiales propres aux données géographiques. Nous proposons une extension de la méthode CART dans deux directions : d'une part, l'algorithme considère plusieurs tables, rejoignant ainsi les travaux sur la fouille de données multi-relationnelles et d'autre part, il calcule les critères de discrimination en déterminant la relation de voisinage et en la combinant aux attributs des objets voisins.

ABSTRACT. Nowadays, spatial data mining is a clearly identified field of data mining. This article deals the spatial data classification using a decision tree. This method differs from conventional decision trees by considering the thematic layer structure, and the spatial relationships that characterize geographical data. We propose an extension of CART method in two directions: from one hand, the algorithm considers several tables which is similar to the multi-relational data mining works, and from the other hand, it computes the discriminating criteria by determining the neighborhood relationship and by combining it with attributes of the neighbor objects.

MOT-CLÉS : Fouille de données spatiales, règle de classement, arbre de décision, relation spatiale, base de données spatiales.

KEYWORDS: Spatial Data Mining, Classification Rules, Decision Tree, Spatial Relationship, Spatial Database.

1. Problématique

Les données géographiques sont de plus en plus utilisées dans les applications décisionnelles, surtout depuis le développement d'outils de géocodage. Seulement, la nature et le volume de données dépassent les capacités humaines en terme d'analyse. D'où l'intérêt d'appliquer des techniques d'extraction automatique de connaissances telles que la fouille de données aux bases de données géographiques. C'est le cas de l'analyse du risque d'accidents routiers dans laquelle s'inscrivent nos travaux [HUG 00].

Nos travaux visent à intégrer le caractère spatial des données et l'interaction avec l'environnement géographique. Ce qui permet, dans l'exemple de la sécurité routière, d'expliquer le risque d'accidents en tenant compte de leur contexte géographique. Ceci revient à un modèle prédictif prenant en considération, en plus des propriétés des objets à classer, les propriétés des objets voisins. Les besoins se situent à deux niveaux :

- Ces objets voisins peuvent appartenir à d'autres couches thématiques que les objets à analyser, autrement dit, à plusieurs tables. Or, les arbres de décision traditionnels se basent sur une seule table où chaque tuple constitue un exemple à classer.
- Les critères spatiaux à considérer sont nombreux. Les précédents travaux [EST 97, KOP 98] se limitent aux relations de voisinage fixées par l'utilisateur. Or, ces critères sont multiples, voir infinis (comme la distance) rendant leur choix difficile. D'où la nécessité de filtrage automatique de ces critères dans la méthode.

Après la section sur l'état de l'art permettant de positionner ce travail, cet article présente la méthode proposée et détaille l'algorithme. La section 4 décrit brièvement l'implémentation et discute les performances, suivie de la conclusion.

2. Fouille de données spatiales et arbres de décision

La fouille de données spatiales (FDS) est née du besoin d'exploitation dans un but décisionnel de données à caractère spatial produites, importées ou accumulées, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires [ZEI 99]. Dans ce domaine, des méthodes classiques comme les arbres de décision ont été adaptées. Un arbre de décision, en général, a pour but de trouver les attributs explicatifs et les critères précis sur ces attributs donnant le meilleur classement vis-à-vis d'un attribut à expliquer. L'arbre est construit par l'application successive de critères de subdivision sur une population d'apprentissage afin d'obtenir des sous-populations plus homogènes [ZIG 00]. Il existe diverses méthodes d'arbres de décision. Le critère de subdivision est déterminé au niveau de l'attribut dans ID3 [QUI 86] et au

niveau d'une valeur d'attribut dans CART [BRE 84]. Il utilise un calcul de gain informationnel pour apprécier la subdivision.

Ester et al. [EST 97] ont proposé une méthode de classification spatiale basée sur ID3 et utilisant le concept de graphe de voisinage. Le grand défaut de cette méthode est qu'elle ne garantit pas une segmentation correcte, car les critères spatiaux ne sont pas discriminants. Cette méthode est de plus limitée à une seule relation de voisinage. Enfin, elle ne fait pas de distinction entre les thèmes. Une autre méthode est proposée dans [KOP 98]. Après une généralisation des données, elle transforme les "attribut = valeur" en prédicats. Les inconvénients sont d'une part, le coût du pré-traitement induit par la transformation en prédicats et d'autre part, l'absence de choix dynamique de la distance dans les critères de discrimination. Comme souligné plus haut, un arbre de décision spatial exploite des données provenant de plusieurs tables. Ceci a été traité dans [KNO 99] parmi les travaux de fouille de données dite multi-relationnelle. Mais, cette méthode ne permet pas de résoudre la détermination des relations spatiales.

3. La méthode d'arbre de décision spatial proposée

L'importance des relations spatiales en FDS nous a amenés à proposer dans [ZEI 00] une structure secondaire dite *index de jointure spatiale* qui pré-calculé la relation spatiale exacte entre les objets spatiaux de deux collections les stocke dans une table relationnelle (objet1, objet2, distance). La méthode proposée se base sur deux idées. La première est l'exploitation de l'index de jointure spatiale et la seconde l'adaptation des méthodes d'arbre de décision multi-relationnelle. Ainsi, grâce à l'index de jointure spatiale, la classification peut se baser désormais sur une représentation directement en relationnel. En effet, la méthode utilise une table à analyser, des tables de correspondances que sont les index de jointures, et d'autres tables décrivant les propriétés d'autres thèmes.

L'algorithme proposé est une extension de la méthode CART. Il s'appuie sur l'indice de Twoing car c'est celui préconisé dans les cas de deux classes [ZIG 00]. Dans cette extension, la partition d'un nœud se base sur le test d'existence d'un objet voisin combiné avec sa relation spatiale avec l'objet à classer. Le fils droit d'un nœud est le complément du fils gauche (ce qui explique le choix d'un arbre binaire telle que CART). Par exemple, le fait qu'il existe à la fois une école et un commerce à distance de 100m de l'accident *a1* et que la condition de segmentation est *qu'il existe une école à distance $\leq 200m$* , alors l'accident *a1* sera affecté au nœud gauche. La seconde originalité de notre méthode est de qualifier précisément les relations de voisinage avec l'objet à classer. Ainsi, le calcul du gain informationnel utilise la combinaison d'une valeur d'attribut de la table liée et de la « relation spatiale » avec l'objet à classer.

Algorithme de construction de l'arbre de décision spatial

L'algorithme prend en entrée trois tables : la première - Table_cible - contient les objets à classer ; la deuxième - Table_voisin - contient les objets voisins des objets à classer ; la troisième représente l'index de jointure spatial.

1. Au fur et à mesure du classement, les tuples de la Table_cible seront attribués¹ à une feuille courante de l'arbre. Initialement, tous les tuples sont affectés au nœud racine de numéro 1.
2. Pour chaque attribut exogène, calculer le meilleur gain informationnel. A ce niveau, on adapte la formule du gain informationnel lorsque l'attribut provient de Table_voisin. On détermine alors la valeur de cet attribut et la relation spatiale R telles que "l'existence de voisins vérifiant la condition *attribut = (ou ≤) valeur et en relation R avec l'objet à classer*" donne le meilleur gain pour l'attribut. La valeur de l'attribut de segmentation est celle maximisant le gain informationnel tout attribut confondu.
3. Si la feuille courante n'est pas saturée, affecter les objets de la feuille courante aux fils gauche ou au fils droit selon qu'ils vérifient ou non la condition de segmentation. A noter que la saturation peut être déclarée par différents critères : seuil minimum d'occupation du nœud, une profondeur maximale de l'arbre ou une valeur plancher du gain informationnel.
4. Itérer l'étape 2 pour le nœud suivant s'il existe. Sinon, l'algorithme s'arrête.

4. Implémentation et discussion

Cette méthode a été mise en œuvre avec Oracle 8i et testée sur des données réelles relatives à la sécurité routière. Un exemple de résultat est donné dans la figure 1. La première condition de segmentation est la proximité d'un marché (à distance de 100m). Elle est liée aux accidents plutôt piétons. Le fils droit de la racine correspond aux autres accidents que ceux à proximité de marchés. Il est segmenté à son tour en une partie proche des écoles où le taux d'accidents piétons est plus fort et le taux d'accidents autres — ici véhicules — est proportionnellement plus faible et inversement pour le fils droit. Au troisième niveau de découpage, on constate que l'attribut le plus discriminant provient de la table cible. La dernière feuille en bas de l'arbre veut dire simplement que si on n'est prêt ni des marchés, ni des écoles, ni des administrations, alors on n'est moins confronté aux accidents impliquant les catégories vulnérables piétons ou 2 roues.

¹ Les numéros des nœuds sont défini récursivement par : $n^{\circ}_{fils_gauche}=2*n^{\circ}_{père}$ et $n^{\circ}_{fils_droit}=2*n^{\circ}_{père} + 1$.

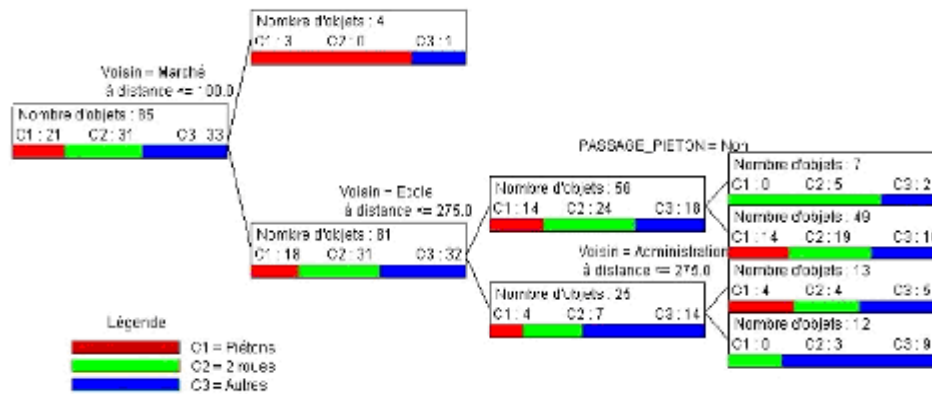


Figure 1. Exemple d'arbre de décision spatial résultat.

Cette méthode mériterait d'autres validations sur un plan opérationnel avec les experts du métier et sur d'autres jeux de test. En particulier, il faudra tester ou simuler les performances du coût d'exécution sur des volumes de données plus importants. D'ores et déjà, des techniques d'optimisation comme l'utilisation de références directes dans l'index de jointure (ROWID dans Oracle) ont été mises en œuvre.

5. Conclusion et perspectives

Contrairement aux méthodes d'arbres de décisions classiques, qui prennent en entrée une seule table dont les tuples sont considérés comme des objets à classer, la méthode que nous proposons dépasse cette limite et étend ces dernières à des données spatiales multi-relationnelles. L'algorithme proposé est une méthode particulière de fouille de données multi-relationnelles car l'index de jointure spatial n'est autre qu'une table de liens multiples entre deux tables en relationnel. Du point de vue FDS, la démarche générale de représentation des liens sous forme tabulaire est très prometteuse. Cette structure a priori entre objets pourra être considérée dans d'autres méthodes comme la classification automatique (non supervisée). Quant à cette méthode de classification supervisée, d'un côté, elle nous permet de classer les objets spatiaux selon à la fois leurs attributs et les attributs de leurs voisins. D'un autre côté, elle effectue le choix automatique de la "bonne" relation de voisinage. C'est là une originalité de la méthode, car le classement ne prend pas seulement en considération les propriétés du voisinage, mais aussi la relation spatiale qui relie les objets à classer avec leur voisinage. En outre, l'organisation en couches thématiques est tout à fait intégrée.

6 Extraction et gestion des connaissances

En perspective, nous orientons nos recherches vers deux axes importants :

- L'étude et l'amélioration des performances : des versions orientées disque des algorithmes d'arbres de décision telles que SLIQ [MEH 96] pourront être adaptées pour éviter les dégradations des performances sur des données volumineuses.
- L'extension de cette démarche à d'autres méthodes de fouille de données spatiales, dont la classification automatique et les associations spatiales.
- Le test sur des données de type image ou spatio-temporelles encore plus importantes en volume de données.

Références

- [BRE 84] Breiman L., Friedman J.H., Olshen R.A., and Stone C. J., *Classification and Regression Trees*, Ed; Wadsworth & Brooks, Monterey, California, 1984.
- [EST 97] Ester M., Kriegel H.P., Sander J., "Spatial Data Mining: A Database Approach", In proceedings of 5th Symposium on Spatial Databases, Berlin, Germany, 1997.
- [HUG 00] Huguenin-Richard F., "Approche géographique des accidents de la circulation : proposition de modes opératoire de diagnostic, application au territoire de la métropole lilloise", Thèse de doctorat, Université de Franche-Comté, Décembre 2000.
- [KNO 99] Knobbe. A.J., Siebes A., Wallen V., Daniel M.G., "Multi-relational Decision Tree Induction", In Proceedings of PKDD' 99, Prague, Czech Republic, Septembre 1999.
- [KOP 98] Koperski K., Han J., Stefanovic N., "An Efficient Two-Step Method for Classification of Spatial Data", In proceedings of International Symposium on Spatial Data Handling (SDH'98), p. 45-54, Vancouver, Canada, July 1998.
- [MEH 96] Mehta M., Agrawal R., Rissanen J. "SLIQ: A Fast Scalable Classifier for Data Mining", In Proc. of Int. Conf. On Extending Database Technology (EDBT'96), Avignon, France, March 25-29, pp 18-32, 1996.
- [QUI 86] Quinlan J.R., "Induction of Decision Trees", *Machine Learning* (1), pp 82 - 106, 1986.
- [ZEI 99] Zeitouni K., *Data mining spatial, Revue internationale de géomatique n° 4/99*, Numéro spécial, Edition Hermès Sciences.
- [ZEI 00] Zeitouni K., Yeh L., Aufaure M-A., "Join indices as a tool for spatial data mining", Int. Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Notes in Artificial Intelligence n° 2007, Springer, pp 102-114, Lyon, France, September 12-16, 2000.
- [ZIG 00] Zighed A., Ricco R., *Graphes d'induction - Apprentissage et Data Mining*, Edition Hermès Sciences, 2000.