

Extraction de Règles de Classification à partir des Données Spatiales

ABDICHE Fethi, ATMANI Baghdad

Equipe de recherche « Simulation, Intégration et Fouille de données (SIF) »
Département d'Informatique, Faculté des Sciences,
Université d'Oran Es-Senia,
BP 1524, El-M'Naouer, 31000, Oran, Algérie.
fethi_abdiche_box@yahoo.fr, atmani.baghdad@univ-oran.dz

Résumé. La fouille de données spatiales est un processus d'exploration des connaissances implicites dans des bases de données volumineuses à références spatiales. Le caractère géométrique des objets que ces bases de données représentent rend le processus d'extraction de la connaissance plus compliqué contrairement aux bases de données numériques traditionnelles. Les algorithmes de la fouille de données spatiales se caractérisent par leurs dépendances à l'analyse de voisinage entre le phénomène en question et son environnement. Ce type d'analyse est le point d'interaction entre la base de données à références spatiales et l'algorithme de la fouille de données. Notre contribution dans ce domaine est d'apporter une solution au problème d'extraction et d'optimisation des règles de classification par la mise en œuvre d'un outil de construction de graphe cellulaire.

Mots clés: Bases de données à références spatiales, Fouille de données spatiales, Requêtes spatiales, Apprentissage symbolique, Graphe d'induction, Automate cellulaire, Extraction de règles.

1 Introduction

Selon [12], [17], jusqu'à 80% des données d'une organisation ont une composante spatiale. Les données spatiales sont de plus en plus nombreuses grâce à l'évolution des outils d'acquisition de données (ex. GPS, images satellite, photo aériennes, etc.) et des méthodes de structuration (ex. raster, vecteur) et de représentation (ex. représentations 2D, 3D). De plus, des outils et des méthodes de représentation des données spatiales (ex. des outils de visualisation) ont été développés pour mettre en évidence les caractéristiques spatiales des données (position, forme, taille, orientation, etc.) et les relations qui existent entre elles (ex. intersection, adjacence, etc.) afin de faciliter leur interprétation. En outre, il y a eu des innovations importantes dans le domaine des technologies de l'information, particulièrement dans les technologies de base de données et des systèmes intelligents d'aide à la décision (SIAD). Les SIAD sont des outils issus de l'IA qui fournissent des techniques, des données et des solutions pour identifier et résoudre les problèmes liés à la prise de décisions stratégiques.

2 ABDICHE Fethi, ATMANI Baghdad

Les SIAD utilisent généralement les techniques de fouille de données qui permettent l'exploration des entrepôts de données actuelles et historiques à différents niveaux d'agrégation. Les entrepôts de données se basent généralement sur une structure multidimensionnelle, ils contiennent alors ce qu'on appelle des « cubes de données ». Les cubes de données facilitent la navigation rapide des données selon différents niveaux de granularité (ex. d'un niveau détaillé à un niveau plus général). Lorsque les cubes de données contiennent des données spatiales (nous les appelons «cubes de données spatiales»), ils permettent à la fois de profiter des avantages de la structure multidimensionnelle et de la représentation cartographique des données spatiales.

La fouille de données est l'analyse des observations de larges jeux de données dans le but d'identifier des relations non soupçonnées et de résumer la connaissance incluse au sein de ces données sous de nouvelles formes à la fois compréhensibles et utiles pour l'expert de ces données. La fouille de données fait appel principalement aux disciplines de l'intelligence artificielle, de l'apprentissage automatique, de la logique floue, des réseaux de neurones, des arbres de décisions et de la statistique et l'analyse de données. Elle s'effectue généralement sur des tables bidimensionnelles et se décompose essentiellement en trois grandes familles de méthodes : les méthodes descriptives, les méthodes de structuration et les méthodes explicatives.

La classification, qui appartient à la troisième catégorie des méthodes de fouille de donnée, est une méthode explicative qui cherche à établir un modèle décrivant un phénomène, défini à partir d'une variable endogène, à l'aide d'un ensemble de descripteurs appelés variables exogènes.

La plupart des méthodes de classification ne prennent pas en considération la représentation des données sous format relationnelles. Une donnée spatiale est généralement composée de trois parties : une partie géométrique (la forme par exemple ligne, point, polygone), une partie descriptive (par exemple le nom, la population d'une ville) et une partie métrique calculée (par exemple la superficie, le périmètre) [16]. Pendant le processus de fouille de données spatiales, on cherche à trouver les règles qui divisent l'ensemble des objets en plusieurs classes en utilisant non seulement leurs attributs descriptifs mais aussi leurs relations spatiales.

Par ailleurs, la représentation de la connaissance est aussi une question fondamentale dans la conception d'un SIAD, car la représentation utilisée peut réduire la complexité en termes de traitement et de stockage. Nous présentons par cet article notre contribution dans ce domaine qui consiste à apporter une solution au problème d'extraction et de représentation des règles de classification par la mise en œuvre d'un outil de construction de graphe cellulaire [5], [2], [6].

2 Travaux connexes

La fouille de données spatiales est définie comme l'extraction de connaissances implicites de relations spatiales ou d'autres propriétés non explicitement stockées dans les bases de données spatiales [18]. Dans ce contexte, beaucoup de travaux existent, mais la majorité consiste à appliquer des techniques de la fouille de données pour extraire de la connaissance à partir des données non spatiales. Nous pouvons

citer à titre d'exemple les travaux de Han [13] qui proposent un algorithme pour l'extraction des relations d'hierarchie. Cette technique suppose que l'utilisateur ou l'expert doit fournir à l'algorithme une hiérarchie de concepts, qui n'est souvent pas possible. Nous pouvons citer aussi la méthode STING [19], qui est une méthode hiérarchique et qui consiste à diviser le territoire en plusieurs zones de taille plus réduite afin de minimiser la complexité de la recherche.

Nous pouvons invoquer aussi les travaux sur les règles d'associations spatiales [15] qui sont une extension des règles d'associations classiques appliquées pour l'extraction d'un motif fréquent. De même nous soulignons l'importance du clustering qui est une méthode de classification automatique permettant le regroupement d'objets par classes homogènes. Pour cela, elle cherche à maximiser la similarité intra-classe et à minimiser la similarité inter-classes. Les principales méthodes sont celles par agrégation autour de centres mobiles, comme les k-means, les nuées dynamiques, la classification automatique hiérarchique (CAH) et enfin, les méthodes par densité comme **DBSCAN** [9], **BIRCH** [22] et **OPTICS** [3]. La transposition au domaine spatial s'appuie sur une mesure de similarité d'objets localisés suivant leur distance métrique.

Nous citons aussi les travaux de Fayyad [11] qui a utilisé les arbres de décision pour la classification des objets stellaires pour les étoiles et les galaxies et, où environs 3 Terabytes d'images ont été analysées. Ces données ont été traitées à l'aide du système FOCAS, qui sélectionne les objets et fournit des attributs de base comme : la magnitude, les régions, intensité, l'orientation, Les objets ont été classifiés par des astronomes. Basé sur cette classification, dix bases de données d'apprentissages ont été conçues pour construire l'arbre de décision dont des règles sont extraites. La méthode ne s'applique pas à la base de données sous format vectoriel, souvent utilisé dans les systèmes d'informations géographique.

Par ailleurs, une méthode de classification des objets spatiaux utilisant les graphes de voisinages a été proposée par Ester [10]. L'algorithme est basé sur la méthode ID3 et les objets sont traités comme voisins s'ils satisfont les relations de voisinages, qui peuvent être une des cas de figure suivants : Les relations topologiques, les relations de métriques et les relations de directions.

Dans notre contribution nous nous sommes inspirés des travaux de Chelghoum [7], par l'algorithme **SCART (Spatial CART)** et les travaux de Zeitouni [20], [21]. La plupart de ces travaux considèrent que les bases de données spatiales et la majorité des SIG organisent les données en couches thématiques, chacune avec une description ou schéma propre. Les méthodes à bases de clustering ou de généralisation ne prennent pas en compte cette organisation et par conséquent, ne peuvent révéler des relations inter-thèmes cachées.

3 Problématique de la fouille de données spatiales

Le Clustering est une méthode en fouille de données permettant le regroupement d'objets par classes homogènes. Donc il s'agit de trouver une métrique qui peut détecter les points similaires vis-à-vis un phénomène et visualisé les concentrations.

4 ABDICHE Fethi, ATMANI Baghdad

Le point faible de cette technique est le fait qu'elle repose sur un seul critère qui est la localisation alors que la réalité est multicritère.

La méthode STING comme les méthodes à base de Clustering, vise à identifier des classes d'objets fortement similaires. Elle ne fournit aucune description sur la relation des classes et leurs relations avec les caractéristiques des objets.

La méthode d'hierarchie et généralisation de concept repose sur l'intervention préalable d'un expert pour définir l'hierarchie de concept.

Ainsi nous pouvons remarquer que les méthodes précédentes n'exploitent qu'un seul thème à la fois, alors que les données géographiques sont souvent représentées par plusieurs couches thématiques qui cachent des relations topologiques entre les objets qu'elles contiennent.

Notre contribution dans ce domaine, en prenant en considération le caractère spatial des données et l'interaction avec l'environnement géographique, est la conception d'un système d'extraction et d'optimisation des règles de classification appliqué à la sécurité routière.

4 Primitives des relations spatiales

Comme il a été souligné précédemment, la mise en place de l'exemple d'apprentissage nécessite l'analyse des données spatiales, cette dernière s'appuie sur la prise en compte des relations spatiales qui sont explicites entre phénomène du territoire (accident, épidémie, séisme, ...) qui fournissent des données implicites, nécessaires pour l'algorithme de la fouille de données, nous disons que la fouille de données spatiales est une extension de la fouille de données classique [11]. Dans ce paragraphe, nous introduisons quelques primitives basées sur le concept de relations de voisinage.

4.1 Les primitives de voisinage [11]

L'influence mutuelle entre deux objets dépend des facteurs tels que la topologie, la distance ou la direction entre les objets. Par exemple, un nouvel ensemble industriel peut polluer son voisinage selon la distance et sur la direction principale du vent. La figure 1 montre une carte utilisée dans l'évaluation d'un endroit possible pour un nouvel ensemble industriel. La carte montre trois régions avec différents degrés de pollution (indiquée par les différentes couleurs) provoqués par l'usine prévue.

Dans cette section, nous présentons trois types de base de relations spatiales : *relations topologiques*, *de distance* et *de direction* qui sont des relations binaires, c'est-à-dire, relations entre les paires d'objets. Les objets spatiaux peuvent être des *points* ou des objets dans l'espace prolongés tels que des *lignes*, des *polygones*. Des objets dans l'espace prolongés peuvent être représentés par un ensemble de points sur sa surface, par exemple par les bords d'un polygone (représentation de vecteur) ou par les points contenus dans l'objet, par exemple les Pixel d'un objet.

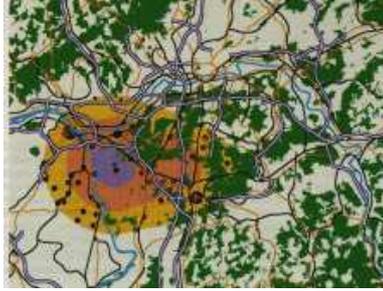


Fig. 1. Les Régions pollués au tour d'une zone industrielle

Les relations Topologiques

Les relations topologiques sont les relations qui restent invariables sous des transformations topologiques, elles sont préservées si les deux objets sont changés d'échelle, translétés ou pivotés simultanément. Les définitions formelles sont basées sur les frontières, les intérieurs et les compléments des deux objets connexes.

Ces relations selon les extensions SQL d'oracle spatial sont [4]: *Inside, Touch, Covers, Equal, Contains, Disjoint, Covered By, Overlap Boundary*. Une définition théorique est dans [8].

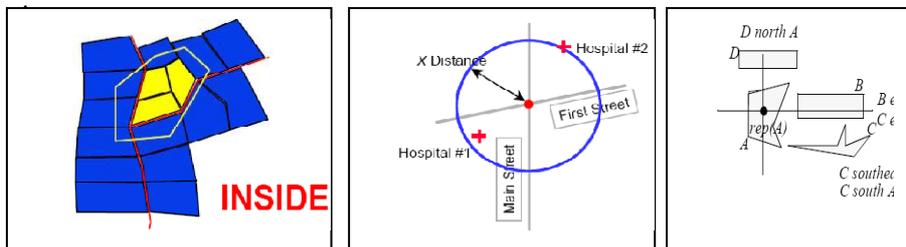
Les relations de distance

Les relations de distance sont ces relations comparant la distance de deux objets à une constante donnée utilisant un des opérateurs arithmétiques. La distance entre deux objets, ensemble de points, peut alors simplement être défini par la distance minimum entre leurs points. Nous citons selon SQL étendu d'oracle spatial [4] : *Within Distance, Nearest Neighbor*.

Les relations de direction

Pour définir la relation de direction $O_2 R O_1$, nous distinguons l'objet source O_1 de l'objet destination O_2 de la relation R de direction. Il y a plusieurs possibilités pour définir des relations de direction selon le nombre de points qu'ils considèrent dans la source et l'objet de destination.

Nous définissons la relation de direction de deux objets dans l'espace prolongés utilisant un représentant de l'objet O_1 de source et tous les points de la destination object O_2 .



a. Relation topologique

b. Relation de distance

c. Relation de direction

Fig. 2. Les relations topologiques et de distance

Dans la construction du tableau de contingence, nous nous basons sur les relations de distance comme relations de voisinage entre les points de la couche thématique des accidents et les équipements publics de la seconde couche (École, administration, supermarché, ...).

5 Contribution : La classification spatiale

La tâche de la classification consiste à assigner un objet à une classe donnée basées sur les valeurs d'attributs de cet objet. Dans la *classification spatiale* les valeurs d'attribut des objets voisins sont également considérées.

L'algorithme présenté dans [14] fonctionne comme suit : Les attributs en question sont extraits en comparant les valeurs d'attribut des objets de cible aux valeurs d'attribut de leurs voisins plus proches. La détermination des attributs appropriés est basée sur les concepts du coup le plus proche (le voisin le plus proche appartenant à la même classe) et du coup manqué le plus proche (le voisin le plus proche appartenant à des classes différentes). Dans la construction de l'arbre de décision, les voisins des objets de cible ne sont pas considérés individuellement. Un Buffer est créé autour de l'objet cible et les valeurs d'attributs de tous les objets contenus dans le Buffer sont agrégées. Par exemple, dans le cas des centres commerciaux un Buffer peut représenter le secteur où ses clients vivent ou travaillent. La taille du Buffer rapportant le gain maximum de l'information est choisi et cette taille est appliquée pour calculer les agrégats pour tous les attributs appropriés. La figure 3 montre un exemple d'arbre de décision spatial.

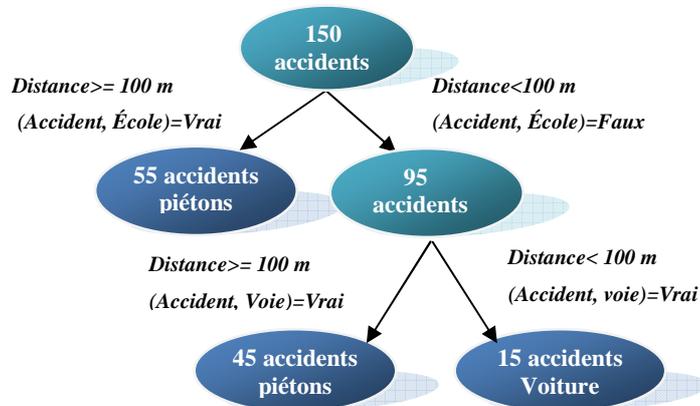


Fig. 3. Arbre de décision spatial [7]

5.1 Le model proposé

Notre système intelligent d'aide à la décision baptisé CGISS propose plusieurs services qui commencent par l'intégration des données issues de plusieurs

sources de données hétérogènes, assurer les fonctionnalités de base d'un SIG, effectuer l'analyse topologique de voisinage pour fournir l'échantillon d'apprentissage, et générer la base de règles.

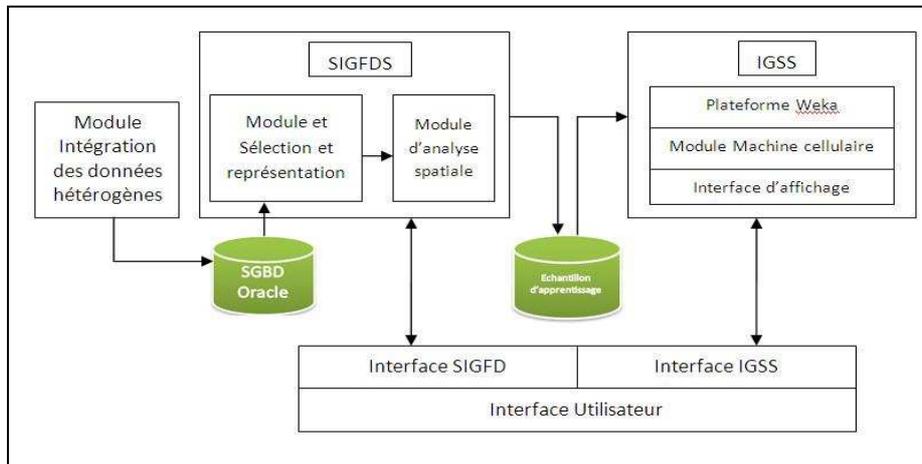


Fig .4. Architecture générale du système CGISS (Cellular Geographic Information System for Spatial datamining)

Le système CGISS proposé dans cet article est composé de trois modules :

1. Le module Intégration des données hétérogènes est chargé de stocker ces dernières dans la même base de données. En effet, les données descriptives ont été indépendantes par rapport aux formes géométriques situées dans les couches thématiques. Le résultat sera stocké dans une base de données de type Oracle [1].
2. Le Composant SIGFDS (Système d'Information Géographique pour la Fouille de Données Spatiales), est équipé de deux modules :
 - Le Module sélection et représentation est équipé des fonctionnalités de bases d'un SIG. Il permet d'extraire et d'afficher les couches thématiques à partir de la base de données stockée sous Oracle. Ainsi d'autres options avancées comme l'ajout ou élimination des couches, la modification de leurs apparitions, affichage des labels, ..., etc.
 - Le module d'analyse spatiale a pour objectif d'explicitier les relations de voisinage entre le lieu d'accident et sont environnement. SIGFDS fourni l'échantillon d'apprentissage nécessaire pour l'exécution du composant IGSS.

Le Module IGSS est le cœur du système. Il a été développé¹ afin d'intégrer le principe cellulaire [2] et enrichir l'environnement graphiques de la plateforme Weka. Il prend en entrée l'échantillon d'apprentissage issu d'une base de données géographique afin de fournir en sortie une base en binaire en appliquant le principe

¹ Au sein de notre équipe de recherche SIF.

booléen de la machine cellulaire CASI (Cellular Automata for Symbolic Induction). Le processus général d'apprentissage que le système cellulaire CASI [5] applique à une population est organisé sur quatre étapes :

1. Initialisation du graphe d'induction par automate cellulaire (principe de SIPINA).
2. Optimisation du graphe d'induction cellulaire.
3. Génération des règles de production.
4. Validation du modèle cellulaire.

5.2 Expérimentation

Soit $\Omega = \{\omega_1, \omega_2, \dots, \omega_{133}\}$ l'ensemble des observations concernées par le problème des accidents et $X = \{X_1, X_2, X_3\}$ un ensemble d'attributs appelées descripteurs pour chaque observation (accident) avec :

X_1 : La proximité de l'accident par rapport aux carrefours.

X_2 : La proximité de l'accident par rapport aux établissements scolaires.

X_3 : La proximité de l'accident par rapport aux marchés.

Nous avons choisi les trois couches : Carrefour, Etablissement et Marché afin de cibler en premier lieu une population forte et un flux important de circulation. Nous avons ignoré les cas particulier et les accidents en dehors de notre champ d'intérêt.

Pour la première couche, c'est l'ensemble des *carrefours* et des intersections représentant un flux important de circulation routière au niveau de la ville d'Oran, donc un nombre important d'accidents dans les deux catégories : voitures et piétons.

La seconde couche contient des objets spatiaux de type *établissements scolaires*. En plus de l'information géométrique contenue dans la base de données, nous distinguons aussi des informations descriptives de l'objet comme le nom de l'établissement ainsi que sont type.

La troisième couche est celle des *marchés* ainsi que les grands magasins. L'objectif d'utilisation de cette dernière comme variable descriptive dans le processus de l'extraction des règles est de viser une concentration de la population. La variable à prédire Y prend ses valeurs dans l'ensemble des classes $C = \{\text{piéton}, \text{voiture}\}$, avec: 1 : accident de type voiture, 2 : accident de piéton.

$$Y : \Omega \rightarrow C = \{\text{piéton}, \text{voiture}\}$$

$$\omega_i \rightarrow Y(\omega_i) = c_j$$

5.3 Construction d'un graphe d'induction

La méthode appliquée pour la construction du graphe d'induction est SIPINA. Son principe consiste à générer une succession de partitions par fusion et/ou éclatement des nœuds du graphe selon les trois variables exogènes X_1, X_2, X_3 (voir table 1) :

Table 1. Attributs prédictifs, valeurs et significations

Variable	Signification	Valeurs
X_1 ($l_1 = 2$)	Carrefour	Distance < 10 PRET , Distance \geq 10 LOIN
X_2 ($l_2 = 2$)	Etablissement	Distance < 10 PRET , Distance \geq 10 LOIN
X_3 ($l_3 = 2$)	Marché	Distance < 10 PRET , Distance \geq 10 LOIN

1. X_1 : Carrefour : Voisinage du lieu de l'accident par rapport aux carrefours et des grands croisements, qui prend les valeurs : 2 : PRET, 1 : LOIN.
2. X_2 : Etablissement : Voisinage du lieu de l'accident par rapport aux établissements scolaires, prend les valeurs : 2 : PRET, 1 : LOIN.
3. X_3 : Marché : Voisinage du lieu de l'accident par rapport aux Marchés et super marché, prend les valeurs : 2 : NON, 1 : OUI.

La valeur de proximité est un paramètre subjectif, nous considérons que deux objets spatiaux sont proches si la distance euclidienne est inférieure à 10 mètres. À la fin de cette phase d'analyse de propriété de voisinage, nous fournissons au module chargé de la génération des règles un échantillon d'apprentissage. Nous utilisons l'outil IGSS développé au sein de notre équipe de recherche, pour la partie arbre de décision et génération des règles.

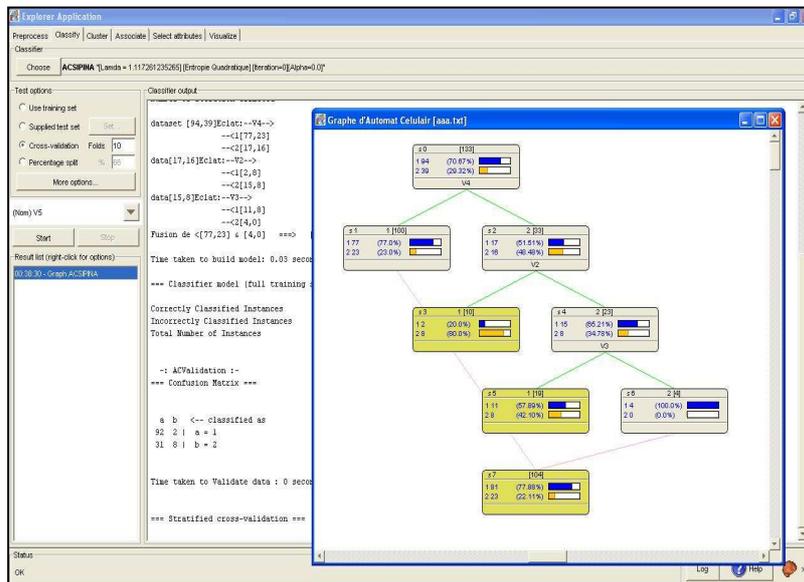


Fig. 5. Construction de l'arbre de décision par IGSS

5.4 Génération des règles conjonctives

À partir de notre échantillon d'apprentissage issu de l'analyse de voisinage, nous appliquons le principe de la machine cellulaire au sein du module IGSS. Ce dernier aide à l'extraction des règles conjonctives, illustrées par la figure 6, qui seront insérées dans la base de règles de la machine CASI [5] (Cellular Automata for Symbolic Induction).

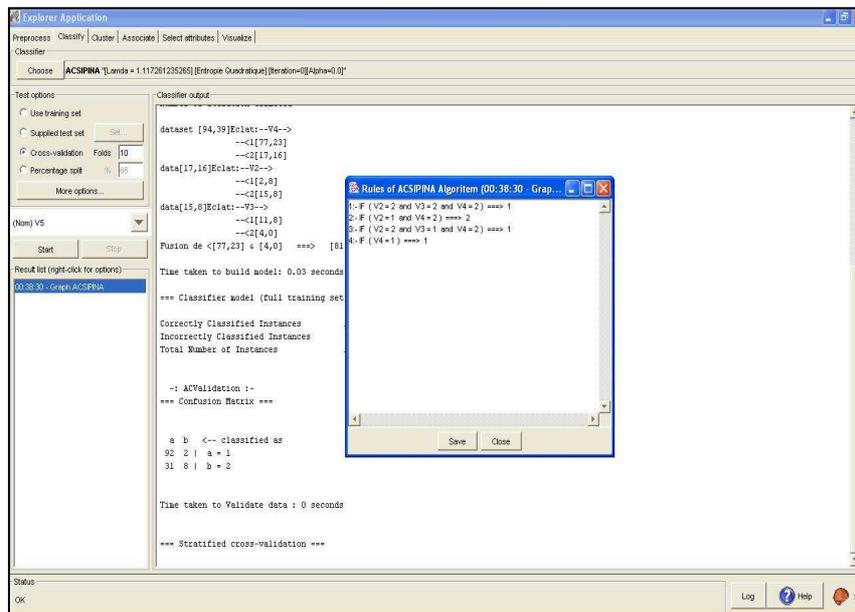


Fig. 6. Génération des règles conjonctives par l'IGSS

L'interprétation des règles est comme suit :

R_1 : IF ($X_1 = 2$ et $X_2 = 2$ et $X_3 = 2$) => 1

Si l'accident est à proximité d'un carrefour, un établissement scolaire ainsi qu'un marché alors il est de type accident de Piéton.

Cette règle est consistante car un établissement scolaire ainsi qu'un marché est un endroit de forte population. La présence d'un carrefour augmente la possibilité d'un accident vue le flux élevé de circulation de voitures.

R_2 : IF ($X_1 = 1$ et $X_3 = 2$) => 2

Si l'accident est à proximité d'un marché mais loin d'un Carrefour alors il est de type accident de voiture.

Les accidents qui sont près d'un marché sont de type accident de voiture, cela est justifiable par le fait que la présence des véhicule de transport des acheteurs et des livreurs dans un endroit restreint augmente la possibilité d'accident entre les véhicules.

R₃: IF ($X_1 = 2$ et $X_2 = 1$ et $X_3 = 2$) => 1

Si l'accident est à proximité d'un carrefour et un marché mais loin d'un établissement scolaire alors il est de type piéton.

Cette règle est consistante aussi car au moins deux facteurs d'un accident de type piéton sont présents : la forte population autour d'un marché est le carrefour d'où le flux élevé de circulation des voitures.

R₄: IF ($X_2 = 1$) => 1

Si l'accident est loin d'un carrefour alors il est de type piéton.

Cette règle ne représente pas un grand intérêt du fait qu'elle possède une seule prémisses donc elle n'est pas porteuse d'information consistante.

En exploitant les propriétés de voisinages entre les objets spatiaux de trois couches thématiques, nous avons réussi à extraire trois règles utiles et porteuses d'information qui n'ont pas été implicites au préalable.

6 Conclusion

Dans cet article nous avons abordé la fouille de données spatiale comme un cas particulier de la fouille de données classique par l'enrichissement avec des requêtes de voisinage. Étant donné la diversité de solutions logicielles existantes implémentant une multitude d'algorithmes d'apprentissage, nous avons choisi dans un premier temps d'intégrer les fonctionnalités d'analyse des propriétés de voisinages. Nous avons réussi donc à ramener plusieurs couches thématiques en une table relationnelle afin d'appliquer n'importe quel algorithme de classification.

Dans le contexte de l'analyse des accidents de la route, le graphe d'induction cellulaire engendré est un modèle booléen qui nous permettra de voir de plus près les relations entre l'accident et son environnement ainsi que les éléments responsables de ce dernier, donc le graphe d'induction engendré facilitera l'identification des problèmes de sécurité sur le réseau routier en vue de proposer des mesures de sécurité pour y remédier.

Références

1. Abdiche, F., Atmani, B.: Vers un système d'unification des systèmes d'informations géographiques, International Conference on Web and Information Technologies ICWIT'08, pp 140--146 (2008)
2. Abdelouhab, FZ, Atmani, B.: Intégration automatique des données semi-structurées dans un entrepôt cellulaire, Troisième atelier sur les systèmes décisionnels, Mohammadia – Maroc, pp 109--120, (2008)
3. Ankerst M., Breunig M.M., Kriegel H-P., Sander J.: OPTICS: Ordering Points To Identify the Clustering Structure. SIGMOD Conference, pp 49--60 (1999)

4. Armitage, T.: Getting started with oracle spatial, Oracle Corporation, United States, (2006)
5. Atmani, B., Beldjilali, B., Knowledge Discovery in Database : Induction Graph and Cellular Automaton, Computing and Informatics Journal, Vol.26, N°2 pp 171--197, (2007)
6. Benamina, M., Atmani, B., WCSS: un système cellulaire d'extraction et de gestion des connaissances, Troisième atelier sur les systèmes décisionnels, 10 et 11 octobre 2008, Mohammadia – Maroc, pp 223--234 (2008)
7. Chelghoum N., Zeitouni K., Boulmakoul A.: A Decision Tree for Multi-layered Spatial Data, In 10th International Symposium on Spatial Data Handling (SDH'02), pp 1--10, Ottawa, Canada (2002)
8. Egenhofer, M. J.: Reasoning about Binary Topological Relations, Proceeding of the 2nd International Symposium. On Large Spatial Databases, Zurich, Switzerland, pp 143--160 (1991)
9. Ester, M., Frommelt, A., Hans-Peter, K., Jörg S.: Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support. Data Mining and Knowledge Discovery, Kluwer Academic Publishers (1999)
10. Ester, M., Kriegel, H. P., Sander, J.: Spatial data mining: A database approach, in Proceedings of the Fifth SSD Symposium (1997)
11. Fayyad, U.M., Djorgovsky, S.G., Weir, N.: Automating the Analysis and Cataloging of Sky Survey. AAAI/MIT Press, Menlo Park, CA (1996)
12. Franklin, C.: An introduction to geographic information systems: linking maps to databases. Database, vol. 15, no. 2, pp.13--21 (1992)
13. Han J., Koperski K., Stefanovic N.: GeoMiner : A System Prototype for Spatial Data Mining, ACM-SIGMOD Int'l Conf. on Management of Data (SIGMOD'97), Tucson, Arizona (1997)
14. Koperski K., Han J., Stefanovic N.: An Efficient Two-Step Method for Classification of Spatial Data, Proceeding. Symposium on Spatial Data Handling (SDH '98), Vancouver, Canada (1998)
15. Kopersky, K.: A progressive refinement approach to spatial data mining, a thesis submitted in spatial fulfillment of the requirement for the degree of Doctor of Philosophy, Simon Fraser University (1999)
16. Rigaux, P., Scholl, M., Voisard, A.: Spatial databases with application to GIS. Morgan Kaufmann Publishers Inc, San Francisco, United States (2002)
17. Sboui, T., Salehi, M., Bédard, Y., Rivest, S. : Catégorisation des problèmes d'intégration des modèles des cubes de données spatiales. Atelier Fouille de données complexes, 8èmes Journées Francophones, Extraction et Gestion des Connaissances, Sophia Antipolis 29 janvier (2008)
18. Shekhar, S., Zhang P., Yan H., Ranga V.: Trends in Spatial Data Mining : Next Generation Challenges and Future Directions, Hillol Kargupta and Anupam Joshi(editors), AAAI/MIT Press (2003)
19. Wang, W., Yang, J., Muntz, R.: STING: A statistical information grid approach to spatial data mining, in Proceedings of the International VLDB Conference (1997)
20. Zeitouni, K. : Analyse et extraction de connaissances des bases de données spatiotemporelles. Habilitation à Diriger des Recherches Spécialité Informatique, Université de Versailles Saint-Quentin-en-Yvelines (2006)
21. Zeitouni, K., Chelghoum, N. : Application du Data Mining Spatial pour la prédiction du risque d'accidents de la circulation en milieu urbain : Extension du projet TOPASE par la prise en compte des interactions entre le réseau viaire et l'environnement urbain, Université de Versailles Saint-Quentin-en-Yvelines (2006)
22. Zhang T., Ramakrishnan R., Livny M.: BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD Conference, pp 103--114 (1996)