Apprentissage de règles caractéristiques

Teddy Turmeaux* — **Daniel Cassard****

Ansaf Salleb* ** — **Christel Vrain***

* *Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)*
Rue Léonard de Vinci, B.P 6759, 45067 Orléans Cedex 2
{turmeaux,salleb,cv}@lifo.univ-orleans.fr

** *Bureau des Recherches Géologiques et Minières (BRGM[1])*
3, avenue Claude Guillemin, B.P. 6009, Orléans cedex 2
d.cassard@brgm.fr

RÉSUMÉ. *La caractérisation est une tâche supervisée de fouille de données qui permet de résumer de manière succincte et concise un ensemble de données. Cette tâche est intéressante dans la mesure où elle ne nécessite pas de contre exemples. Nous proposons un cadre général pour la caractérisation d'un ensemble d'objets, appelé ensemble 'cible', en nous basant non seulement sur leurs propriétés propres mais aussi sur les propriétés des objets qui leur sont liés. Selon le type des objets considérés, différents liens peuvent être envisagés. Dans le cas de bases de données géographiques, ce sont les relations spatiales qui expriment des liens entre objets géoréférencés. Nous proposons des algorithmes d'extraction de règles de caractérisation et nous montrons comment nous les avons appliqués à des données géographiques réelles fournies par le BRGM.*

ABSTRACT. *Among data mining tasks, characterization does not attract much attention from researchers, in comparison to classification. It seems to us an interesting task, since it does not require negative examples, which may be a strong requirement for real applications. In this paper, we present a general framework for the characterization of a target set of objects by means of their own properties, but also the properties of objects linked to them. According to the kinds of objects, various links can be considered. In the case of geographic databases, spatial relations express links between geographic objects. We propose some algorithms for mining characterization rules, and we show how they have been applied to a real geographic application provided by BRGM.*

MOTS-CLÉS : *Fouille de données, règles caractéristiques, systèmes d'information géographiques.*

KEYWORDS: *Data Mining, Characteristic Rules, Geographic Information Systems.*

1. Le BRGM est une entreprise présente dans plus de 40 pays à travers le monde, intervenant dans le domaine des Géosciences pour la gestion durable des ressources de l'espace souterrain.

1. Introduction

La *caractérisation* est une tâche descriptive de fouille de données qui tend à résumer de manière concise un ensemble de données. Contrairement aux tâches de classification ou de discrimination, la caractérisation ne nécessite pas a priori de contre-exemples. C'est une propriété importante puisque disposer de contre-exemples n'est pas toujours possible dans les applications réelles. C'est le cas par exemple de l'application qui a motivé les développements présentés ici : à partir de données géologiques et métallogéniques stockées dans un Système d'Information Géographique (**SIG**), caractériser des gisements en fonction de leur contexte géologique.

Nous proposons dans ce papier un cadre général pour la caractérisation d'un ensemble d'objets, appelé ensemble *cible*, basé non seulement sur leurs propriétés propres mais aussi sur les propriétés des objets qui leur sont liés. Selon le type des objets considérés, différents liens peuvent être envisagés. Nous présentons ici un algorithme d'extraction de règles de caractérisation d'objets géographiques ; cependant, l'approche développée est plus générale et peut s'appliquer à d'autres types de bases de données, comme par exemple, les bases de données relationnelles.

Mentionnons que nous nous sommes d'abord intéressés à la tâche d'extraction des règles d'association [SAL 00] dans les bases de données géographiques. Cependant, elle présentait le défaut d'engendrer beaucoup trop de règles, d'autant plus qu'un SIG est formé de plusieurs couches thématiques d'objets géographiques reliés entre eux par des relations spatiales. Nous nous sommes également intéressés aux motifs fréquents [SAL 02], mais les motifs pertinents sont très souvent noyés dans des masses de connaissances qui les rendaient inaccessibles à l'expert. La recherche de règles caractéristiques a l'intérêt d'être plus dirigée, car elle se focalise sur un ensemble d'objets cibles à caractériser. Notons que tout comme la recherche de motifs fréquents, la caractérisation repose aussi sur une notion de fréquence des propriétés.

Peu de travaux se sont intéressés à cette tâche. D'une part, la caractérisation a été abordée d'un point de vue *généralisation descriptive* en Apprentissage Symbolique. Par exemple, dans [MIC 83], Michalski étudie l'apprentissage de règles caractéristiques et l'apprentissage de règles discriminantes à partir d'exemples. En Programmation Logique Inductive, la notion de plus petit généralisé, comme par exemple définie dans [PLO 70], permet de capturer toutes les propriétés communes à un ensemble d'objets, mais pose des problèmes de complexité des algorithmes. D'autre part, dans le domaine de la Fouille de Données, Han et al. [HAN 92, HAN 96] ont introduit l'induction orientée attribut pour la généralisation de données, notion étendue aux données spatiales par Lu et al. [LU 93]. Par ailleurs, Ester et al. [EST 98] ont proposé une méthode pour caractériser des objets géographiques à partir de leurs attributs non spatiaux, et aussi des propriétés de leurs voisins. Cette approche permet d'extraire seulement des propriétés de type (attribut=valeur). De plus la notion de voisinage est définie de façon statique sous forme de graphe au début du processus.

L'originalité de notre approche repose sur une définition de schémas caractéristiques qui spécifient comment prendre en compte les différents types d'objets et leurs

relations. Ils précisent de plus comment regrouper les objets les uns par rapport aux autres. Par exemple, dans la formule $\forall\ Mine\ \exists_{10km}\ Volcan :: Age(Volcan, récent)$, $\forall\ Mine\ \exists_{10km}\ Volcan$ représente un schéma caractéristique et la formule signifie que pour chaque mine, il existe un volcan récent à moins de 10 km. Le schéma $\forall\ Mine\ \exists_{10km}\ Volcan$ spécifie donc que l'on cherche des propriétés satisfaites par au moins un volcan dans un buffer de 10 km autour de chaque mine. En revanche, le schéma $\forall\ Mine\ \forall_{10km}\ Volcan$ spécifie que l'on cherche des propriétés satisfaites par tous les volcans dans un buffer de 10 km autour de chaque mine. L'algorithme que nous proposons parcourt un ensemble de schémas possibles et recherche les propriétés caractéristiques pour chacun de ces schémas.

2. Formulation du problème

Une base de données géographiques est composée d'objets géographiques organisés en couches thématiques. Dans chaque couche, les objets sont décrits par leur forme, leur position et les attributs décrivant leurs propriétés. Les informations topologiques et de distance permettent de calculer différents types de relations spatiales entre les objets comme par exemple *près_de*, *intersecte*, ...

La tâche de caractérisation qui nous intéresse peut se formuler comme suit :
• étant donné un ensemble $\mathcal{E}$ d'objets, $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cdots \cup \mathcal{E}_n$ où chaque $\mathcal{E}_i$ représente un ensemble d'objets de même type T_i,
• étant donné un ensemble $\mathcal{R}$ de relations binaires r_{ij} définies sur deux ensembles $\mathcal{E}_i$ et $\mathcal{E}_j$ quelconques,
• étant donné un ensemble $\mathcal{E}_{cible}$, tel qu'il existe un i, $\mathcal{E}_{cible} \subseteq \mathcal{E}_i$,
$\rightarrow$ trouver un ensemble de règles de caractérisation de $\mathcal{E}_{cible}$

Précisons que les relations binaires sont soit des relations prédéfinies comme *près_de*, soit paramétrées par un paramètre λ spécifiant par exemple une taille de buffer.
Par exemple, $r_{i,j}^{100km}$ représente une relation binaire et $r_{i,j}^{100km}(o, o')$ signifie que o et o' appartenant respectivement à $\mathcal{E}_i$ et $\mathcal{E}_j$ sont à une distance inférieure à 100km.
La figure 1 montre un graphe illustrant deux mines et les objets qui leur sont liés (de type faille et volcan).

Schémas caractéristiques. Un schéma est une expression de la forme :
$\forall\ X_0\ Q_1\ X_1 \ldots Q_n\ X_n$, où $n \geq 0$, X_0 est le type des objets cibles et pour chaque $i \neq 0$, $Q_i = \forall\ ou\ \exists$, X_i est un type d'objets et il existe une relation de $\mathcal{R}$ entre X_{i-1} et X_i. Notons que dans le cas où il existe plusieurs relations entre X_{i-1} et X_i, le quantificateur Q_i peut être indicé par la relation utilisée dans le schéma. La taille d'un schéma est égale au nombre de quantificateurs qui apparaissent dans la formule. On dit que deux schémas sont des *variantes* s'ils ont la même taille, s'ils portent sur les mêmes types d'objets, les mêmes relations, dans le même ordre et s'ils diffèrent par au moins un de leurs quantificateurs.

Par exemple, pour caractériser des mines, le schéma $\forall Mine\ \forall Volcan$ représente *pour chaque mine, pour chaque volcan qui lui est lié*, alors que $\forall Mine\ \exists Volcan$ représente *pour chaque mine, pour au moins un volcan qui lui est lié*.

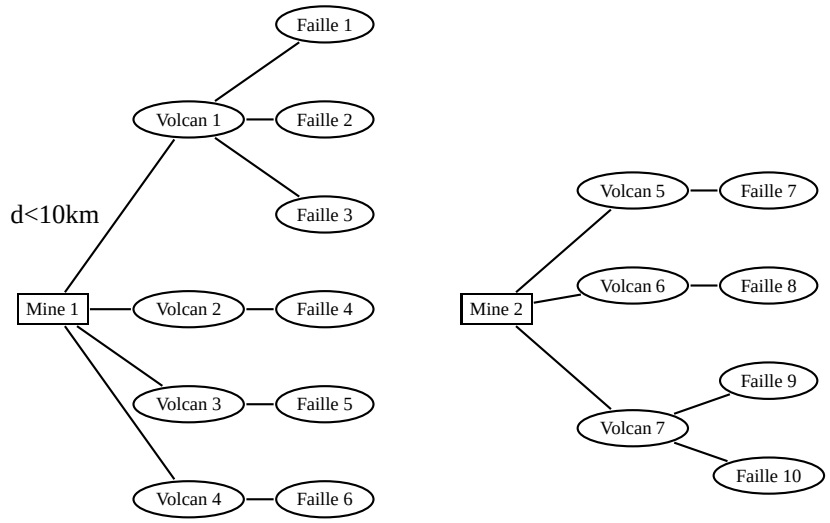

Figure 1. *Deux mines et les objets situés à moins de 10km les uns des autres*

Des objects cibles aux sacs. Dans la suite, un *sac* [1] représente simplement un ensemble d'objets, et on note par $\mathcal{B}$ l'ensemble des sacs.

Pour chaque schéma caractéristique δ, nous définissons la fonction f_δ qui associe à chaque objet de la cible, un ensemble de sacs qui représente comment regrouper les objets qui lui sont liés selon δ.

$$f_\delta : \quad \mathcal{E}_{cible} \rightarrow 2^{\mathcal{B}} \qquad o_i \mapsto \{b_j\}$$

Si nous considérons à nouveau l'exemple des mines, des volcans et des failles, nous avons :

$$\begin{array}{lcl} f_{\forall M \forall V}(Mine_1) & = & \{\{Volcan_1\}, \{Volcan_2\}, \{Volcan_3\}, \{Volcan_4\}\} \\ f_{\forall M \forall V}(Mine_2) & = & \{\{Volcan_5\}, \{Volcan_6\}, \{Volcan_7\}\} \\ f_{\forall M \exists V}(Mine_1) & = & \{\{Volcan_1, Volcan_2, Volcan_3, Volcan_4\}\} \\ f_{\forall M \forall V \exists F}(Mine_1) & = & \{\{Volcan_1, Faille_1, Faille_2, Faille_3\}, \\ & & \{Volcan_2, Faille_4\}, \{Volcan_3, Faille_5\}, \\ & & \{Volcan_4, Faille_6\}\} \end{array}$$

Pour des raisons d'efficacité, nous considérons seulement des schémas où tous les quantificateurs universels ($\forall$) précèdent les quantificateurs existentiels ($\exists$). De plus, pour de tels schémas, il est facile d'implanter la fonction f_δ en utilisant des requêtes SQL.

1. Une notion assez voisine de sacs a déjà été introduite dans [DIE 97] et [CHE 01] dans le cas de l'apprentissage à partir d'instances multiples.

Par exemple, la requête SQL suivante permet de constituer l'ensemble des sacs du schéma $\forall M \forall_{d<10km} V \exists_{d<10km} F$:

```
SELECT Mine.id, Volcan.id, Faille.* from Mine, Volcan, Faille
WHERE distance(Mine,Volcan) < 10 and distance(Volcan,Faille)<10
ORDER BY Mine.id,Volcan.id
```

Les sacs sont constitués des tuples partageant le même couple (*Mine.id,Volcan.id*). La requête donne de plus toutes les valeurs des attributs des failles concernées. Chaque sac est rattaché à un objet cible par *Mine.id*.

Propriété et couverture. Nous disons qu'un sac satisfait (ou vérifie) une propriété *p* si au moins l'un de ses objets satisfait *p* . Nous disons qu'une propriété *p couvre* un objet *o* relativement à un schéma δ si tous les sacs de $f_\delta(o)$ vérifient *p* .

Nous notons $couverture(\delta, p, \mathcal{F})$ le rapport $\frac{|\{o|o \in \mathcal{F} \text{ et } p \text{ couvre } o \text{ relativement à } \delta\}|}{|\mathcal{F}|}$. où $\mathcal{F}$ est un ensemble d'objets de $\mathcal{E}$.

Par exemple, une propriété *p* couvre l'objet $Mine_1$ pour le schéma $\forall M \forall V$ si $volcan_1, volcan_2, volcan_3, volcan_4$ satisfont cette propriété. En revanche, *p* couvre $Mine_1$ pour le schéma $\forall M \exists V$ si au moins un objet parmi $volcan_1, volcan_2, volcan_3$, $volcan_4$ satisfait cette propriété.

Pour chaque type de schémas, nous cherchons les propriétés p satisfaites par un nombre maximum d'objets cibles, i.e., telles que $couverture(\delta, p, \mathcal{E}_{cible}) \geq \epsilon$ où ϵ est un seuil donné.

Les propriétés peuvent aussi résumer des informations concernant des objets en utilisant les *agrégats*. Un agrégat est une fonction donnant pour un ensemble d'objets une information agrégée sur ces objets. Utiliser les agrégats dans la tâche de caractérisation peut s'avérer très utile et peut produire des règles d'un pouvoir expressif très intéressant. Comme exemple d'agrégats que l'on peut utiliser dans ce cadre, citons : *sum, count, avg, min, max*. Ainsi, la propriété d'être *proche d'au moins deux volcans* peut être exprimée en fonction de l'agrégat *count* comme suit :
$Agrégat_{proche}(V, count) \geq 2$.

Règles caractéristiques. Nous définissons une règle caractéristique comme la conjonction d'un schéma caractéristique δ et d'une propriété *p* . Elle est notée $\delta :: p$. Notons qu'une propriété peut être une conjonction de propriétés élémentaires.

Règles caractéristiques intéressantes. En plus de la notion de couverture, nous utilisons un indicateur $intéressante(\delta,\ p,\ \mathcal{E}_{cible})$. Dans l'application que nous avons développée, où $\mathcal{E}_{cible}$ est strictement inclus dans $\mathcal{E}$, l'indicateur utilisé est proche de celui utilisé dans [EST 98]. Il est défini de manière à tenir compte du caractère contrastant de la propriété entre $\mathcal{E}_{cible}$ et $\mathcal{E} - \mathcal{E}_{cible}$:

$intéressante(\delta,\ p,\ \mathcal{E}_{cible})\ est\ vraie\ si\ \frac{\mathcal{RG}}{\mathcal{RL}} \ll 1$ ou $\frac{\mathcal{RG}}{\mathcal{RL}} \gg 1$
avec :
$\mathcal{RG} = \frac{|\mathcal{E}_{cible}|}{|\mathcal{E}-\mathcal{E}_{cible}|}$
$\mathcal{RL} = \frac{couverture(\delta,\ p,\ \mathcal{E}_{cible})}{couverture(\delta,\ p,\ \mathcal{E}-\mathcal{E}_{cible})}$

L'indicateur ainsi défini permet de privilégier les propriétés vérifiées plus dans l'ensemble cible qu'à l'extérieur ; par exemple, si les objets cibles sont les mines d'or, on peut s'intéresser aux propriétés vérifiées plus particulièrement par les mines d'or et nettement moins vérifiées par les mines d'argent.

L'algorithme proposé ci-dessous se base sur ces notions de couverture et d'intérêt pour rechercher les règles caractéristiques d'un ensemble d'objets cibles.

Algorithme de base
Entrées :
$-\ \mathcal{E}_{cible}$, *un ensemble d'objets cibles*
$-\Delta$, *un ensemble de schémas caractéristiques*
$-\mathcal{P}$, *un ensemble de propriétés*
$-\epsilon$, *un seuil*
Sortie :
$-\Re$, *un ensemble de règles caractéristiques*
Debut
$\Re = \emptyset$
Pour tous les $p \in \mathcal{P}$
  Pour tous les $\delta \in \Delta$
    Si $couverture(\delta, p, \mathcal{E}_{cible}) \geq \epsilon$ et $intéressante(\delta, p, \mathcal{E}_{cible})$
      ajouter la règle $\delta :: p$ à $\Re$
    FinSi
Fin

Couverture imparfaite. Nous disons qu'une propriété *p couvre imparfaitement* un objet *o* relativement à un schéma δ si au moins un des sacs de $f_\delta(o)$ vérifie *p* .

Nous notons $couverture\text{-}imparfaite(\delta, p, \mathcal{F})$ le rapport :

$$\frac{|\{o | o \in \mathcal{F}\ et\ p\ couvre\ imparfaitement\ o\ relativement\ à\ \delta\}|}{|\mathcal{F}|}$$

où $\mathcal{F}$ est un ensemble d'objets de $\mathcal{E}$.

Remarquons que la couverture imparfaite est la même pour toutes les variantes d'un même schéma : si δ_1, δ_2 sont des variantes, alors $couverture\text{-}imparfaite(\delta_1, p, \mathcal{F}) = couverture\text{-}imparfaite(\delta_2, p, \mathcal{F})$. On notera $couverture\text{-}imparfaite(\Delta, p, \mathcal{F})$ la couverture imparfaite d'un ensemble Δ de variantes d'un schéma relativement à p et $\mathcal{F}$.

La motivation d'une telle mesure est de pouvoir élaguer l'espace de recherche lorsque l'on considère des variantes d'un schéma. Par exemple, $\forall\ Mine\ \exists\ Volcan$ et $\forall\ Mine\ \forall\ Volcan$ sont deux variantes, et si la couverture imparfaite de l'un des deux n'est pas suffisante, il est alors inutile de considérer le second schéma. L'algorithme utilisant la notion de couverture imparfaite est donné ci-dessous :

Algorithme révisé 1//$élagage\ par\ couverture\ imparfaite$
Entrées :
$-\ \mathcal{E}_{cible}$, *un ensemble d'objets cibles*
$-\Delta$, *un ensemble de variantes d'un schéma caractéristique*
$-\mathcal{P}$, *un ensemble de propriétés*
$-\epsilon$, *un seuil*
Sortie :
$-\Re$, *un ensemble de règles caractéristiques*
Debut
$\Re = \emptyset$
Pour tous les $p \in \mathcal{P}$
  Si $couverture\text{-}imparfaite(\Delta, p, \mathcal{E}_{cible}) \geq \epsilon$
    Pour tous les $\delta \in \Delta$
      Si $couverture(\delta, p, \mathcal{E}_{cible}) \geq \epsilon$ et $intéressante(\delta, p, \mathcal{E}_{cible})$
        ajouter la règle $\delta :: p$ à $\Re$
  FinSi
Fin

3. Application aux bases de données géographiques

Nous illustrons le cadre général d'extraction de règles de caractérisation que nous avons proposé par une application sur un système d'information géographique réel : le SIG Andes.

Développé au *BRGM*, le *SIG Andes* est un système d'information homogène à vocation géologique et métallogénique, sur l'ensemble de la cordillère des Andes.
Ce SIG est constitué de plusieurs couches thématiques concernant les données géologiques, minéralogiques, sismiques, volcaniques, gravimétriques, ... à l'échelle de la cordillère [CAS 99, Sig], sauvegardant ainsi près de 70 mille objets géographiques. La figure 2 montre trois couches : Géologie, Mines, et Failles sur une partie des Andes. Le SIG offre un bon moyen de visualisation et de stockage des données mais il doit être valorisé pour l'aide à la prospection minière. C'est pourquoi nous avons développé un prototype, dans le but d'extraire des règles de caractérisation des mines permettant d'aider l'expert à dresser des cartes de potentiel métallogénique des Andes. Cette carte in fine pourra être utilisée pour explorer les zones non encore explorées et les plus favorables à la présence de gisements dans les Andes.

Le problème se formalise comme suit. Etant donné $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cup \mathcal{E}_4 \cup \mathcal{E}_5$, où $\mathcal{E}_1$ contient les *mines*, $\mathcal{E}_2$ représente la *géologie*, $\mathcal{E}_3$ les *volcans*, $\mathcal{E}_4$ les *failles* et enfin $\mathcal{E}_5$ les *séismes*, étant donné un ensemble $\mathcal{R}$ de relations binaires basées sur la proximité entre les objets, caractériser l'ensemble $\mathcal{E}_{cible} = \{$*mines d'or*$\} \subseteq \mathcal{E}_1$.

Comme relation de proximité, nous choisissons une relation de distance entre objets : par exemple, $r_{1,3}^{100km}$ représente une relation entre les mines et les volcans, telle que $r_{1,3}^{100km}(m, v)$ est vrai si m et v sont à une distance inférieure ou égale à 100 kilomètres.

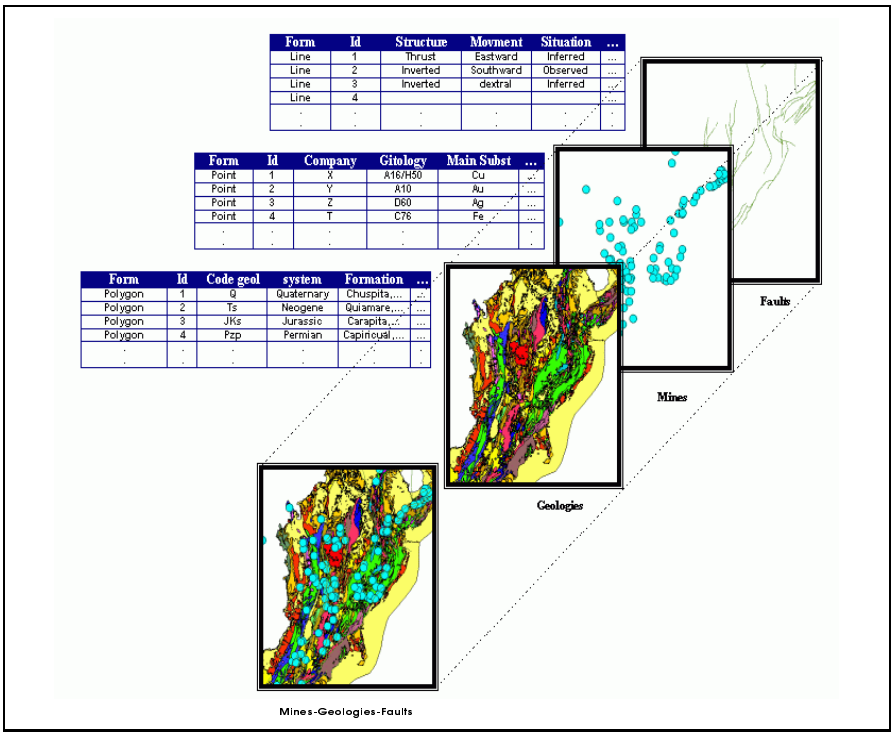

Figure 2. *Exemple de couches du SIG Andes*

Afin de gérer cette notion de distance de façon dynamique durant le processus, l'algorithme construit des buffers croissants autour des objets cibles progressivement tout en cherchant les propriétés satisfaites par les objets entrant dans le buffer. Cette notion de buffer croissant est illustrée en figure 3, où des buffers sont construits autour des mines d'or.

On remarque que les buffers possèdent des propriétés de monotonie intéressantes : si l'on note r^{λ} la relation binaire signifiant *être à moins de* λ *km*, alors si $\lambda_1 \leq \lambda_2$ on a $r^{\lambda_1} \subseteq r^{\lambda_2}$. Ceci permet d'exploiter l'ordre de généralité sur les schémas défini comme suit :

Relation de généralité. On dit qu'un schéma caractéristique δ_1 est plus général qu'un schéma caractéristique δ_2 (noté $\delta_1 \succeq \delta_2$) si pour tout objet $o \in \mathcal{E}_{cible}$, quelque soit le sac $b_i \in f_{\delta_1}(o)$ considéré, il existe un sac $b_j \in f_{\delta_2}(o)$ tel que $b_j \subseteq b_i$. Dans ce cas, si tous les sacs de $f_{\delta_2}(o)$ satisfont une propriété p, alors tous les sacs de $f_{\delta_1}(o)$ satisfont cette propriété. Autrement dit pour tout p tel que $couverture(\delta_2, p, \mathcal{E}_{cible}) \geq \epsilon$ (où ϵ est un seuil donné), alors $couverture(\delta_1, p, \mathcal{E}_{cible}) \geq \epsilon$.

Ceci permet de diminuer le nombre des règles à considérer ; par exemple, dans le cas des règles caractérisques à un seul paramètre, on a $\delta_\lambda \succeq \delta_{\lambda'}$ ssi ($\lambda \geq \lambda'$ et λ *porte* sur un $\exists$) ou ($\lambda \leq \lambda'$ et λ porte sur un $\forall$). Ainsi :

$\forall M \forall_{3Km} F \succeq \forall M \forall_{5Km} F \succeq \forall M \forall_{10Km} F$
$\forall M \exists_{10Km} F \succeq \forall M \exists_{5Km} F \succeq \forall M \exists_{3Km} F$

Intuitivement, cela signifie que si une propriété est vraie pour toutes les failles situées à moins de 10 km d'une mine, cette propriété sera *a fortiori* vraie pour toutes les failles à moins de 5 km et de 3 km de cette mine. Inversement, si il existe une faille à moins de 3 km d'une mine présentant une certaine propriété, alors *a fortiori* il existera une faille présentant cette propriété à moins de 5 km et de 10 km.

Dans le cas des schémas à plusieurs paramètres, on peut induire un ordre partiel (sur l'ensemble des schémas) en considérant la relation $\delta_{\lambda_1,\ldots,\lambda_n} \succeq \delta_{\lambda'_1,\ldots,\lambda'_n}$ si $\forall i$, ($\lambda_i \geq \lambda'_i$ et λ_i, λ'_i porte sur un $\exists$) ou $\forall i$, ($\lambda_i \leq \lambda'_i$ et λ_i, λ'_i porte sur un $\forall$). L'élagage de l'espace de recherche basé sur la notion de généralité sur les schémas est utilisé dans l'algorithme suivant :

Algorithme révisé 2 $//élagage\ par\ ordre\ de\ généralité\ sur\ les\ schémas$
Entrées :
$-\mathcal{E}_{cible}$, *un ensemble d'objets cibles*
$-\Delta$, *un ensemble ordonné de schémas caractéristiques*
$-\mathcal{P}$, *un ensemble de propriétés*
$-\epsilon$, *un seuil*
Sortie :
$-\Re$, *un ensemble de règles caractéristiques*
Debut
$\Re = \emptyset$
Pour toutes les propriétés $p \in \mathcal{P}$
  Pour tous les $\delta \in \Delta$ du plus général au plus spécifique
    Si $couverture(\delta, p, \mathcal{E}_{cible}) < \epsilon$
      sortir de la boucle interne
    Sinon
      Si $intéressante(\delta,\ p,\ \mathcal{E}_{cible})$
        ajouter la règle $\delta :: p$ à $\Re$
      FinSi
    FinSi
Fin

Résultats. La figure 4 représente le nombre de mines proches d'une faille dans un buffer de taille A autour de la mine et telle que la faille est proche d'un volcan dans un buffer de taille B autour de cette faille.

Le processus d'extraction des règles de caractérisation sur les Andes Centrales a conduit à l'extraction de la règle caractéristique suivante couvrant près de 60% des mines d'or et rejetant la majeure partie des autres mines.

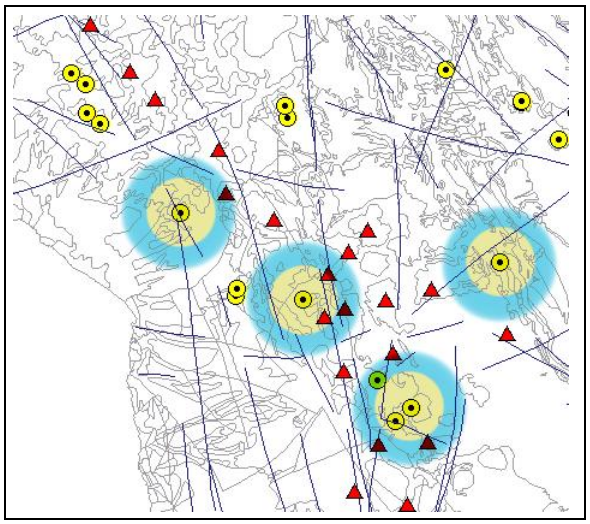

Figure 3. *Buffers autour des mines d'or. Les couches représentées ici sont la géologie, les mines, les failles et les volcans.*

$$
\begin{array}{ll}
\forall M \exists_{10km} G :: & Mine(M) \wedge Géologie(G) \wedge \\
& Substance(M, or) \wedge \\
& Profondeur\_Benioff(M) \in [75..150] \wedge \\
& Distance\_Benioff(M) \in [170..275] \wedge \\
& Pente(M) \in [8°..16°] \wedge \\
& Age(G, tertiaire) \wedge \\
& Lithologie(M, volcanique) \wedge \\
& Gitologie(M, épithermale) \wedge \\
& Morphologie(M, veines)
\end{array}
$$

Cette règle caractéristique exprime le fait que les mines d'or sont des gisements épithermaux qui ont souvent une morphologie en veines (ensemble de fissures de la croûte terrestre) et se sont formés grâce à un phénomène volcanique, lié à l'échappement en surface de magma provenant des couches profondes de la terre. L'âge des roches encaissant les gisements d'or est quant à lui obtenu grâce à la géologie porteuse des gisements et est situé autour de l'ère tertiaire. Cette règle met également en évidence des propriétés très intéressantes au vu des experts concernant la tectonique des plaques. En effet, il apparaît que la subduction, i.e. le passage de la lithosphère océanique sous la lithosphère continentale [2] également appelé plan de bénioff joue un rôle très important dans la formation des gisements d'or.

2. qui est un phénomène géodynamique naturel qui conduit à la génèse de fluides minéralisateurs

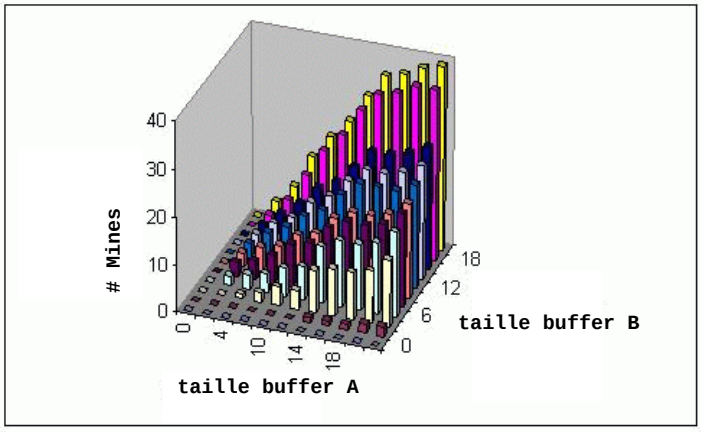

Figure 4. *Couverture de la règle* $\forall M \; \exists_A F \; Agrégat_B(V, count) \geq 1$

Cette règle a été validée par les experts comme une règle caractérisant les gisements d'or dans les Andes Centrales par rapport à une approche "expert" [BIL] effectuée précédemment par l'équipe de géologues du BRGM.

4. Conclusion et perspectives

Nous avons proposé un cadre pour l'extraction des règles de caractérisation d'un ensemble d'objets cibles dans des données géographiques. Notre approche repose sur une définition de schémas caractéristiques, spécifiant comment prendre en compte les objets les uns par rapport aux autres. Les algorithmes que nous proposons parcourent un ensemble de schémas et recherchent les propriétés caractéristiques pour chacun de ces schémas. Ils reposent tous sur des notions de couverture et d'intérêt. La notion de couverture imparfaite est introduite comme heuristique générale permettant de filtrer l'espace de recherche. Une relation de généralité sur les schémas qui permet aussi d'élaguer la recherche est introduite et appliquée dans le cadre des bases de données géographiques. Nous envisageons d'autres algorithmes utilisant aussi les caractéristiques des propriétés recherchées (monotonie/anti-monotonie, par exemple).

Notons enfin que le cadre que nous avons proposé est généralisable à d'autres types de bases de données représentant des objets et leurs relations, comme par exemple les bases de données relationnelles.

Remerciements
Nous adressons nos chaleureux remerciements au personnel de l'équipe de REM (Ressources Minérales) du BRGM et en particulier à Lips Andor, Bruno Tourlière et Mario Billa pour leur collaboration dans la préparation des données et l'interprétation des résultats obtenus.

5. Bibliographie

[BIL] BILLA M., CASSARD D., BOUCHOT V., TOURLIÈRE B., STEIN G., GUILLOU-FROTTIER L., « Predicting epithermal-porphyry gold systems in the central Andes with the continental-scale metallogenic GIS Andes », Soumis à la revue Ore Geology Review.

[CAS 99] CASSARD D., « GIS ANDES : A Metallogenic GIS of the Andes Cordillera », *4th Int. Symp. on Andean Geodynamics*, IRD Paris, October 1999, p. 147–150.

[CHE 01] CHEVALEYRE Y., ZUCKER J.-D., « A Framework for Learning Rules from Multiple Instance Data », *12th European Conference on Machine Learning*, vol. 2167 de *LNCS*, Springer, 2001, p. 49–60.

[DIE 97] DIETTERICH T. G., LATHROP R. H., LOZANO-PEREZ T., « Solving the Multiple Instance Problem with Axis-Parallel Rectangles », *Artificial Intelligence*, vol. 89, n° 1-2, 1997, p. 31-71.

[EST 98] ESTER M., FROMMELT A., KRIEGEL H.-P., SANDER J., « Algorithms for characterization and trend detection in spatial databases », *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, New York City, NY, 1998, p. 44-50.

[HAN 92] HAN J., CAI Y., CERCONE N., « Knowledge Discovery in Databases : An Attribute-Oriented Approach », YUAN L.-Y., Ed., *Proceedings of the 18th International Conference on Very Large Databases*, San Francisco, U.S.A., 1992, Morgan Kaufmann Publishers, p. 547–559.

[HAN 96] HAN J., FU Y., « Exploration of the Power of Attribute-Oriented Induction in Data Mining », FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P., UTHURUSAMY R., Eds., *Advances in Knowledge Discovery and Data Mining*, AIII Press/MIT Press, 1996.

[LU 93] LU W., HAN J., OOI B. C., « Discovery of General Knowledge in Large Spatial Databases », *1993 Far East Workshop on GIS (IEGIS 93)*, Singapore, 1993, p. 275–289.

[MIC 83] MICHALSKI R. S., « A Theory and Methodology of Inductive Learning », *Machine Learning : An Artificial Intelligence Approach*, vol. 2(4), 1983, p. 83–134.

[PLO 70] PLOTKIN G., « A note on inductive generalization », *Machine Intelligence*, vol. 5, p. 153-163, Edinburgh University Press, 1970.

[SAL 00] SALLEB A., VRAIN C., « An Application of Association Rules Discovery to Geographic Information Systems », ZIGHED D. A., KOMOROVSKI J., ZYTKOW J., Eds., *4th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'2000*, vol. 1910 de *Lecture Notes in Artificial Intelligence*, Lyon, France, 2000, Springer, p. 613–618.

[SAL 02] SALLEB A., MAAZOUZI Z., VRAIN C., « Mining Maximal Frequent Itemsets by a Boolean Approach », VAN HARMELEN F., Ed., *ECAI*, Lyon, France, 2002, Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press Amsterdam, p. 385-389.

[Sig] « http ://www.brgm.fr/sigand ».