Notion de sémantiques bien-formées pour les règles

Marie Agier*,**, Jean-Marc Petit**
*DIAGNOGENE
83, avenue Charles de Gaulle
15000 Aurillac
**LIMOS, UMR 6158 CNRS
Univ. Clermont-Ferrand II
63177 Aubière

Résumé. La notion de règles entre attributs est très générale, allant des règles d'association en fouille de données aux dépendances fonctionnelles (DF) en bases de données. Malgré cette diversité, la syntaxe des règles est toujours la même, seule leur sémantique diffère. Pour une sémantique donnée, en fonction des propriétés induites, des techniques algorithmiques sont mises en oeuvre pour découvrir les règles à partir des données. A partir d'un ensemble de règles, il est aussi utile en pratique de *raisonner* sur ces règles, comme cela est le cas par exemple avec les axiomes d'Armstrong pour les dépendances fonctionnelles. Dans cet article, nous proposons un cadre qui permet de s'assurer qu'une sémantique donnée pour les règles est *bien-formée*, i.e. les axiomes d'Armstrong sont justes et complets pour cette sémantique. Les propositions faites dans ce papier proviennent du contexte applicatif de l'*analyse de données de biopuces*. A partir de plusieurs sémantiques pour les données d'expression de gènes, nous montrons comment ces sémantiques s'intègrent dans le cadre présenté.

1 Introduction

Les biopuces permettent aujourd'hui aux biologistes de mesurer l'expression de milliers de gènes simultanément et un des défis majeurs fixé à présent est de comprendre les *réseaux de régulation géniques*, i.e. de découvrir les interactions entre les différents gènes.

Dans le cadre de nos travaux, nous nous intéressons à définir des *règles* entre les gènes à partir de données d'expression de gènes, étant entendu que ces règles forment un modèle possible d'un réseau de régulation. Ces données sont à valeurs réelles, chaque valeur représentant le niveau d'expression d'un gène pour une expérience (ou biopuce) particulière.

La notion de *règles* entre attributs est très générale, allant des règles d'association en fouille de données aux dépendances fonctionnelles en bases de données. Malgré cette diversité, la syntaxe des règles est toujours la même, seule leur sémantique diffère. Pour une sémantique donnée, en fonction des propriétés induites, des techniques algorithmiques sont mises en oeuvre pour découvrir les règles à partir des données [Agrawal et Srikant, 1994, Lopes *et al.*, 2002, Morishita et Sese, 2000].

A partir d'un ensemble de règles, il est aussi très utile en pratique de pouvoir

raisonner sur ces règles sans accéder aux données, comme cela est le cas par exemple avec les axiomes d'Armstrong pour les dépendances fonctionnelles.

Puisqu'une seule sémantique pour les règles ne peut répondre aux diverses attentes des biologistes, nous avons proposé dans des travaux antérieurs plusieurs types de sémantiques pour les règles sur des données d'expression [Agier *et al.*, 2004]. Nous avons commencé à montrer leur intérêt d'un point de vue biologique [Agier *et al.*, 2003].

Dans ce papier, nous nous intéressons à quelques "bonnes" propriétés que nous aimerions voir satisfaites par une sémantique. Plus précisément, nous nous focalisons sur un cadre qui permet de s'assurer qu'une sémantique donnée pour les règles est *bien-formée*, i.e. les axiomes d'Armstrong sont justes et complets pour cette sémantique. De fait, ce système d'inférence s'applique aussi aux implications d'un système de fermeture [Ganter et Wille, 1999] et a donc une portée plus générale que celle induite par les dépendances fonctionnelles (voir les nombreuses applications citées dans [Ganter et Wille, 1999]).

Les intérêts pratiques d'une sémantique bien-formée se situent à la fois sur le raisonnement qui est rendu possible sur les règles mais aussi sur la définition des couvertures des règles et des possibilités algorithmiques qui en découlent lorsque l'on s'intéresse à leur découverte à partir des données. Comme exemple de raisonnement, le *problème d'implication* qui consiste à savoir si une règle est impliquée par un ensemble de règles est possible et peut alors être fait en temps linéaire [Beeri et Berstein, 1979].

Nous montrons que les sémantiques présentées dans [Agier *et al.*, 2003, Agier *et al.*, 2004] entrent dans le cadre que nous proposons dans ce papier et nous donnons une nouvelle sémantique qui n'entre pas dans ce cadre.

Etat de l'art L'hypothèse que les règles d'association sont un modèle pour la découverte de réseaux de régulation de gènes a été partiellement validée dans [Aussem et Petit, 2002, Becquet *et al.*, 2002, Creighton et Hanash, 2003, Cong *et al.*, 2004]. Les règles entre gènes obtenues à partir de données d'expression de gènes sont donc une connaissance prometteuse dans ce domaine, puisqu'elles permettent de mieux comprendre les interactions entre gènes. Il faut noter que la plupart des travaux se sont focalisés sur les règles d'association avec diverses façons de discrétiser et divers indices de qualité pour les règles. Néanmoins, nous pensons qu'au regard de la diversité des objectifs biologiques rencontrés, se restreindre à la sémantique des règles d'association est un peu restrictif et c'est donc tout naturellement que nous avons proposé d'autres sémantiques pour les règles [Agier *et al.*, 2003, Agier *et al.*, 2004].

Le système d'inférence d'Armstrong pour les dépendances fonctionnelles a été étudié dans [Armstrong, 1974]. Le lecteur intéressé sur les liens entre système de fermeture et implication est référé au livre [Ganter et Wille, 1999].

La génération de règles d'association est une méthode très populaire qui a attiré un grand nombre de chercheurs ces dix dernières années [Agrawal et Srikant, 1994]. Au demeurant, de nombreux auteurs se sont intéressés à la réduction du nombre de règles d'association générées [Bastide *et al.*, 2000, Cristofor et Simovici, 2002, Li et Hamilton, 2004, Luong, 2001], sans pour autant se positionner vis-à-vis du système d'inférence d'Armstrong.

Organisation du papier La section 2 présente le cadre théorique que nous proposons à partir du cadre défini pour les dépendances fonctionnelles. La section 3 discute des intérêts pratiques du cadre présenté. Plusieurs exemples de sémantiques pour les données d'expression de gènes sont rappelées en section 4 et leur adéquation au cadre est étudié. La section 5 conclut et présente quelques perspectives.

2 Cadre théorique

Soit $U = \{A_1, A_2, ..., A_n\}$ un ensemble fini d'attributs. Chaque attribut $A \in U$ prend ces valeurs dans un ensemble infini dénombrable D. Un *tuple* ou *ligne* sur U est un élément de D^n. On appelle *relation* un ensemble de tuples et on dit que r est une relation *sur* U et que U est le *schema* de la relation r. Soient $X \subseteq U$ un ensemble d'attributs et t un tuple, on note $t[X]$ la restriction de t à X.

2.1 Syntaxe et sémantique d'une règle

Definition 1 La *syntaxe* d'une *règle* sur U est une expression de la forme $X \rightarrow Y$ qui se lit "X implique Y" avec $X, Y \subseteq U$.

La *sémantique* d'une *règle* sur U est la signification, le sens que l'on souhaite donner à cette règle. Afin d'essayer de "capturer" un grand nombre de sémantiques, nous proposons dans la définition 2, une définition relativement "générique" de la sémantique des règles. Pour écrire cette définition, nous avons fait les choix suivants :

1. Nous nous focalisons sur les *règles exactes*, même s'il est clair que les règles approximatives sont très intéressantes puisqu'elles prennent en compte le bruit dans les données. Des indices ont été proposés, comme par exemple le seuil de confiance minimum pour les règles d'association ou les mesures d'erreur pour les dépendances fonctionnelles [Kivinen et Mannila, 1995] et peuvent néanmoins s'intégrer dans notre cadre (cf section 3).
2. Nous ne prenons pas en compte les divers indices de qualité qu'il est possible de définir à posteriori sur les règles, dans la mesure où ces indices peuvent s'appliquer à de nombreuses sémantiques très différentes. Ces indices permettent de limiter le nombre de règles générées et de les trier. De nombreuses mesures de qualité ont été proposées, comme par exemple le support minimum, la dépendance, le taux informationnel [Agrawal et Srikant, 1994, Suzuki et Kodratoff, 1998, Blanchard *et al.*, 2004, Tan *et al.*, 2004]... Dans ce papier, nous avons donc choisi de ne pas définir autant de sémantiques qu'il existe de mesures de qualité tout simplement dans un but de clarté.

Etant donnée une relation r, la *satisfaction* d'une règle $X \rightarrow Y$ dans r pour une *sémantique* s, notée $r \models_s X \rightarrow Y$, peut se définir de façon générale comme suit :

Definition 2 Soient $X, Y \subseteq U$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour une sémantique s, notée $r \models_s X \rightarrow Y$, est définie par :
$r \models_s X \rightarrow Y$ ssi $\forall r' \subseteq r$ tel que $d_c(r')$, si $Pred(X, r')$ est vrai alors $Pred(Y, r')$ est vrai où :

1. $d_c(r')$ spécifie une contrainte qui doit être vérifiée par $r' \subseteq r$.
2. $Pred(X, r')$ est un prédicat spécifiant une condition de X sur r', pour $X \subseteq U$ et $r' \subseteq r$.

Une sémantique est donc définie à partir de la contrainte $d_c(r')$ et du prédicat $Pred(X, r')$.

Nous donnons ci-dessous deux exemples de définition qui entrent facilement dans ce cadre : une pour la satisfaction des dépendances fonctionnelles et l'autre pour la satisfaction des règles d'association exactes sans support minimum. En section 4, d'autres sémantiques seront proposées avec ce type de formulation.

La définition 3 rappelle la définition classique des dépendances fonctionnelles :

Definition 3 (*df* : sémantique des dépendances fonctionnelles) Soient $X, Y \subseteq U$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour la sémantique des dépendances fonctionnelles *df*, notée $r \models_{df} X \rightarrow Y$, est définie par :
$r \models_{df} X \rightarrow Y$ ssi $\forall t_1, t_2 \in r$, si $t_1[X] = t_2[X]$ alors $t_1[Y] = t_2[Y]$.

Pour reformuler cette définition dans ce cadre général, il suffit de prendre :

1. $d_c(r') = [r' = \{t_1, t_2\}$ avec $t_1, t_2 \in r]$.
2. $Pred(X, \{t_1, t_2\}) = [t_1[X] = t_2[X]]$, pour $X \subseteq U$.

La définition 4 rappelle la définition classique des règles d'association exactes sans support minimum, appelées règles d'association dans la suite du papier pour faire plus court :

Definition 4 (ra : sémantique des règles d'association) Soient $X, Y \subseteq U$, $D = \{0, 1\}$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour la sémantique des règles d'association ra, notée $r \models_{ra} X \rightarrow Y$, est définie par :
$r \models_{ra} X \rightarrow Y$ ssi $\forall t \in r$, si $\forall A \in X, t[A] = 1$ alors $\forall B \in Y, t[B] = 1$.

Pour reformuler cette définition dans le cadre de la définition 2, il suffit de prendre :

1. $d_c(r') = [r' = \{t\}$ avec $t \in r]$.
2. $Pred(X, \{t\}) = [\forall A \in X, t[A] = 1]$, pour $X \subseteq U$.

2.2 Rappel sur les axiomes d'Armstrong

Les axiomes d'Armstrong ont été définis [Armstrong, 1974] et ont été prouvés justes et complets pour les dépendances fonctionnelles [Ullman, 1982]. Rappelons le système d'axiomes d'Armstrong pour un ensemble de règles F défini sur U :

1. (réflexivité) si $X \subseteq Y \subseteq U$ alors $F \vdash Y \rightarrow X$
2. (augmentation) si $F \vdash X \rightarrow Y$ et $W \subseteq U$, alors $F \vdash XW \rightarrow YW$
3. (transitivité) si $F \vdash X \rightarrow Y$ et $F \vdash Y \rightarrow Z$ alors $F \vdash X \rightarrow Z$

La notation $F \vdash X \rightarrow Y$ signifie qu'une preuve de $X \rightarrow Y$ peut être obtenue en utilisant le système d'axiomes d'Armstrong de F. De plus, pour une sémantique donnée s, la notation $F \models_s X \rightarrow Y$ signifie que pour toute relation r sur U, si $r \models_s F$ alors $r \models_s X \rightarrow Y$.

Le système d'axiomes d'Armstrong est juste et complet si ces trois axiomes ne génèrent pas de règles incorrectes (la justesse) et s'ils génèrent bien toutes les règles possibles pouvant être déduites de F (la complétude). Finalement, montrer que le système d'axiomes d'Armstrong est juste et complet pour une sémantique donnée s revient à montrer que si $F \vdash X \rightarrow Y$ alors $F \models_s X \rightarrow Y$ (la justesse) et que si $F \models_s X \rightarrow Y$ alors $F \vdash X \rightarrow Y$ (la complétude).

Nous définissons ainsi la notion de sémantique bien-formée à partir des axiomes d'Armstrong :

Definition 5 Une sémantique s est *bien-formée* si le système d'axiomes d'Armstrong est juste et complet pour s.

Par exemple, on a le résultat classique suivant si on considère la sémantique des dépendances fonctionnelles [Ullman, 1982].

Théorème 1 La sémantique des dépendances fonctionnelles *df* est *bien-formée*

Pour la suite, nous avons besoin des notions suivantes : Pour un ensemble de règles F sur U et une sémantique donnée s, la *fermeture* de F notée F^+ est définie comme suit : $F^+ = \{X \rightarrow Y | F \models_s X \rightarrow Y\}$ et $X^+ = \{A \in U \mid X \rightarrow A \in F^+\}$. On dit que X est un ensemble *fermé* si $X = X^+$. L'ensemble $\{X \subseteq U | X = X^+\}$ est un système de fermeture sur U. Un ensemble de règles F_1 est une *couverture* ou *base* de l'ensemble de règles F_2 si $F_1^+ = F_2^+$. Une couverture F_1 est *minimale* si pour chaque règle $X \rightarrow Y \in F_1$, $X \rightarrow Y \notin (F_1 \setminus X \rightarrow Y)^+$. F_1 est une couverture *canonique* si F_1 contient uniquement des règles avec un attribut en partie droite et $X \rightarrow A \in F_1$ implique $\forall Y \subset X, Y \rightarrow A \notin F_1^+$. Une telle couverture est unique et minimale. Une couverture F_1 de F est *minimum* si $\nexists F_2$ tel que $F_2^+ = F^+$ et $|F_2| < |F_1|$.

3 Intérêts pratiques

Dans un contexte de découverte de connaissances, la découverte de règles satisfaites dans une relation pour une sémantique bien-formée offre plusieurs avantages :

– On peut tout d'abord *raisonner* sur les règles à partir des axiomes d'Armstrong sans accéder aux données. A partir d'un ensemble de règles F, il est possible de savoir si une règle est *impliquée* par cet ensemble de règles [Beeri et Berstein, 1979]. Ainsi, si on a une relation r qui satisfait F alors on sait que toutes les règles pouvant être déduites de F par les axiomes d'Armstrong seront satisfaites dans cette relation. Notons que seule la justesse du système d'axiomes d'Armstrong est nécessaire dans ce cas.

– On peut de plus travailler sur des "petites" *couvertures* des règles [Maier, 1980, Guigues et Duquenne, 1986] et proposer un processus de découverte spécifique à la couverture considérée, mais applicable à *toute* sémantique bien-formée. De plus, il est aussi possible de proposer des couvertures pour les règles approximatives [Gottlob et Libkin, 1990], répondant ainsi à une des limites mentionnée lors de la proposition d'une définition générique des sémantiques.

A partir des données, la première étape du processus de découverte de règles revient à calculer une *famille génératrice* du système de fermeture. Cette étape d'accès aux données peut être prépondérante si les données ne tiennent pas en mémoire et est en tout cas spécifique à la sémantique considérée. Notons qu'une famille génératrice d'un système de fermeture est une représentation équivalente de l'ensemble des règles satisfaites.

Une fois les accès aux données réalisés, deux principales techniques se dégagent pour calculer une couverture des règles :

– Celles qui passent par l'énumération du système de fermeture afin par exemple de générer une couverture minimum [Guigues et Duquenne, 1986].
– Celles qui évitent l'énumération du système de fermeture pour générer la couverture canonique [Mannila et Räihä, 1994, Demetrovics et Thi, 1995].

Le cadre théorique que nous proposons d'utiliser pour la génération des règles pour des sémantiques bien-formées, vient de l'inférence des dépendances fonctionnelles [Mannila et Räihä, 1994, Demetrovics et Thi, 1995]. Basiquement, puisque par définition les axiomes d'Armstrong s'appliquent pour n'importe quelle sémantique bien-formée, l'axiome d'augmentation implique une *propriété de monotonie* : étant donné un attribut A, $X \rightarrow A \Rightarrow \forall\ Y \supset X, Y \rightarrow A$.

Autrement dit le prédicat "X implique A" est monotone pour l'inclusion ensembliste, et donc le prédicat "X n'implique pas A" est anti-monotone. Ainsi, des caractérisations générales développées en fouille de données [Mannila et Toivonen, 1997] peuvent être utilisées pour générer les règles.

En d'autres termes, les plus grandes parties gauches n'impliquant pas A constituent la *bordure positive* du prédicat "X n'implique pas A" et les plus petites parties gauches impliquant A constituent sa *bordure négative*.

Par conséquent, cette bordure négative donne la couverture canonique des règles exactes de A (c'est-à-dire les règles ayant la plus petite partie gauche et A en partie droite) alors que la bordure positive donne la couverture de Gottlob et Libkin pour les règles non satisfaites de A [Gottlob et Libkin, 1990] (c'est-à-dire les règles ayant la plus grande partie gauche et A en partie droite).

4 Application aux données d'expression de gènes

4.1 Exemples de sémantiques

Nous reprenons dans la suite les trois sémantiques présentées dans [Agier *et al.*, 2004] que nous illustrons sur des données d'expression de gènes puis nous montrerons comment celles-ci entrent dans le cadre que nous proposons. Notons que ces sémantiques peuvent s'appliquer à n'importe quelles données numériques.

4.1.1 Sémantique s_1 : chaque ligne est prise en compte une à une

Cette première sémantique considère chaque ligne (ou tuple) de la relation une à une. Notons que la définition de cette sémantique est proche de la définition des règles d'association proposée dans la section 2.1 mais est appliquée ici directement à des attributs numériques ce qui évite une phase de discrétisation explicite.

Definition 6 Soient $X, Y \subseteq U$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour la sémantique s_1 définie avec deux seuils ϵ_1 et ϵ_2, notée $r \models_{s_1} X \rightarrow Y$, est définie par : $r \models_{s_1} X \rightarrow Y$ ssi $\forall t \in r$, si $\forall A \in X, \epsilon_1 \leq t[A] \leq \epsilon_2$ alors $\forall B \in Y, \epsilon_1 \leq t[B] \leq \epsilon_2$.

Pour reformuler cette définition dans le cadre général défini précédemment, il suffit de prendre :

1. $d_c(r') = [r' = \{t\}$ avec $t \in r]$.
2. $Pred(X, \{t\}) = [\forall A \in X, \epsilon_1 \leq t[A] \leq \epsilon_2]$, pour $X \subseteq U$.

Exemple des données d'expression de gènes Cette sémantique permet par exemple aux biologistes d'étudier les relations entre les gènes sur-exprimés (ou sous-exprimés). Pour cela, les seuils ϵ_1 et ϵ_2 doivent être choisis de façon pertinente, par exemple : $\epsilon_1 = 1.0$ et $\epsilon_2 = 2.0$, si l'on considère qu'un gène est sur-exprimé si son niveau d'expression est compris entre 1.0 et 2.0. Le choix des bons seuils est un problème difficile en soi [Pensa *et al.*, 2004].

Considérons la relation r donnée dans la Tab. 1, qui compte 6 lignes ou expériences (t_1, t_2, t_3, t_4, t_5 et t_6) et 8 attributs ou gènes (g_1,g_2,g_3,g_4,g_5,g_6,g_7 et g_8). Les valeurs représentent l'expression des gènes pour les différentes expériences.

r	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
t_1	1.9	0.4	1.4	-1.5	0.3	1.8	0.8	-1.4
t_2	1.7	1.5	1.2	-0.3	1.4	1.6	0.7	0.0
t_3	1.8	-0.7	1.3	0.8	-0.1	1.7	0.9	0.6
t_4	-1.8	0.4	1.7	1.8	0.6	-0.4	1.0	1.5
t_5	-1.7	-1.4	0.9	0.5	-1.8	-0.2	1.2	0.2
t_6	0.0	1.9	-1.9	1.7	1.6	-0.5	1.1	1.3

TAB. 1 – Exemple de données d'expression de gènes

Dans cet exemple, nous avons $r \models_{s_1} g_1 \rightarrow g_3$. La règle $g_1 \rightarrow g_3$ s'interprète de la façon suivante : Pour toute expérience, si le gène g_1 est sur-exprimé alors le gène g_3 est aussi sur-exprimé.

4.1.2 Sémantique s_2 : prise en compte des lignes deux par deux

Dans de nombreux cas, il peut être intéressant de comparer les lignes deux à deux, il s'agit d'un raisonnement bien connu en bases de données à travers la notion de dépendances fonctionnelles.

Cependant, les dépendances fonctionnelles ne peuvent s'appliquer aux données d'expression de gènes puisque la plupart des valeurs réelles de la relation diffèrent l'une de l'autre, autrement dit, pratiquement chaque gène est une clé [Aussem et Petit, 2002]. La satisfaction des DF doit donc être relâchée, par exemple de la façon suivante :

Definition 7 Soient $X, Y \subseteq U$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour la sémantique s_2 définie avec deux seuils ϵ_1 et ϵ_2, notée $r \models_{s_2} X \rightarrow Y$, est définie par : $r \models_{s_2} X \rightarrow Y$ ssi $\forall t_1, t_2 \in r$, si $\forall A \in X, \epsilon_1 \leq |t_1[A] - t_2[A]| \leq \epsilon_2$ alors $\forall B \in Y, \epsilon_1 \leq |t_1[B] - t_2[B]| \leq \epsilon_2$.

Pour reformuler cette définition dans le cadre de la définition 2, il suffit de prendre :

1. $d_c(r') = [r' = \{t_1, t_2\}$ avec $t_1, t_2 \in r]$.
2. $Pred(X, \{t_1, t_2\}) = [\forall A \in X, \epsilon_1 \leq |t_1[A] - t_2[A]| \leq \epsilon_2]$, pour $X \subseteq U$.

La satisfaction classique des dépendances fonctionnelles est réalisée quand $\epsilon_1 = \epsilon_2 = 0$.

Exemple des données d'expression de gènes Dans le contexte des données d'expression, la règle $g_1 \rightarrow g_2$ définie avec cette sémantique et avec des seuils faibles, peut s'interpréter de la façon suivante : chaque fois que g_1 a un niveau d'expression similaire pour deux expériences de r, alors g_2 a aussi un niveau d'expression similaire dans ces expériences. Si on choisit les seuils ϵ_1 et ϵ_2 de la façon suivante : $\epsilon_1 = 0.0$ et $\epsilon_2 = 0.2$, alors à partir des données décrites dans la Tab. 1, nous avons $r \models_{s_1} g_6 \rightarrow g_7$.

4.1.3 Sémantique s_3 : lignes ordonnées

La sémantique s_3 comme la précédente, prend en compte les lignes deux par deux mais ne considère pas toutes les paires possibles de lignes. En effet, on suppose ici que les expériences sont ordonnées dans le temps i.e. il existe un *ordre* sur les lignes de la relation. On s'intéresse uniquement à comparer l'expérience correspondant au temps i avec l'expérience correspondant au temps i+1. Voici la définition proposée :

Definition 8 Soient $X, Y \subseteq U$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour la sémantique s_3 définie avec deux seuils ϵ_1 et ϵ_2, notée $r \models_{s_3} X \rightarrow Y$, est définie par : $r \models_{s_3} X \rightarrow Y$ ssi $\forall t_i, t_{i+1} \in r$, si $\forall A \in X, \epsilon_1 \leq t_{i+1}[A] - t_i[A] \leq \epsilon_2$ alors $\forall B \in Y, \epsilon_1 \leq t_{i+1}[B] - t_i[B] \leq \epsilon_2$.

Pour reformuler cette définition dans le cadre proposé, il faut prendre :

1. $d_c(r') = [r' = \{t_i, t_{i+1}\}$ avec $t_i, t_{i+1} \in r]$.
2. $Pred(X, \{t_i, t_{i+1}\}) = [\forall A \in X, \epsilon_1 \leq t_{i+1}[A] - t_i[A] \leq \epsilon_2]$, pour $X \subseteq U$.

Nous avons supprimé la valeur absolue puisque ici l'ordre des expériences a de l'importance.

Exemple des données d'expression de gènes Nous devons supposer pour cette sémantique que les expériences décrites dans la Tab. 1, sont ordonnées. Par exemple, l'expérience t_1 peut représenter l'état d'une cellule à un instant initial i_0, puis suite à l'injection d'une drogue, on analyse à nouveau la cellule 6 heures plus tard donnant l'expérience t_2 et ainsi de suite jusqu'à l'expérience t_6 qui représente la cellule 30 heures plus tard. Ce processus permet de visualiser l'impact d'une drogue sur l'expression des gènes de la cellule dans le temps. Supposons que les biologistes s'intéressent aux gènes qui croissent dans le temps, les seuils peuvent par exemple être définis de la façon suivante : $\epsilon_1 = 1.0$ et $\epsilon_2 = 4.0$, c'est-à-dire que l'expression d'un gène croît entre l'instant i et l'instant i+1, si son niveau d'expression à l'instant i+1 est supérieur de plus de 1 point à son niveau d'expression à l'instant i. Ainsi, nous avons $r \models_{s_3} g_2 \rightarrow g_4$. La règle $g_2 \rightarrow g_4$ s'interprète de la façon suivante : Entre deux instants quelconques i et i+1, si le niveau d'expression du gène g_2 croît alors le niveau d'expression du gène g_4 croît.

4.2 Adéquation au cadre théorique

Les trois sémantiques présentées entrent dans le cadre théorique défini précédemment :

Théorème 2 Les sémantiques s_1, s_2 et s_3 sont *bien-formées*.

Nous montrons le résultat uniquement pour s_1, la preuve est basiquement la même pour les deux autres sémantiques. Nous devons montrer que le système d'axiomes d'Armstrong est juste et complet pour s_1.

Lemme 1 Le système d'axiomes d'Armstrong est juste pour s_1.

Preuve 1 Soit F un ensemble de règles sur U. Nous devons montrer que si $F \vdash X \rightarrow Y$ alors $F \models_{s_1} X \rightarrow Y$.
Soit r une relation sur U.

1. (réflexivité) évident.
2. (augmentation) Soit $t \in r$ tel que $\forall A \in X \cup W, \epsilon_1 \leq t[A] \leq \epsilon_2$. Nous devons montrer que $\forall A \in Y \cup W, \epsilon_1 \leq t[A] \leq \epsilon_2$, ce qui implique que $r \models_{s_1} XW \rightarrow YW$. Par hypothèse $F \vdash X \rightarrow Y$, donc nous avons $\forall A \in Y, \epsilon_1 \leq t[A] \leq \epsilon_2$. Le résultat suit.
3. (transitivité) Soit $t \in r$ tel que $\forall A \in X, \epsilon_1 \leq t[A] \leq \epsilon_2$. Nous devons montrer que $\forall A \in Z, \epsilon_1 \leq t[A] \leq \epsilon_2$, ce qui implique que $r \models_{s_1} X \rightarrow Z$. Par hypothèse, $F \vdash X \rightarrow Y$ et $F \vdash Y \rightarrow Z$, donc $\forall A \in Y, \epsilon_1 \leq t[A] \leq \epsilon_2$ et $\forall A \in Z, \epsilon_1 \leq t[A] \leq \epsilon_2$ respectivement. Le résultat suit.

Lemme 2 Le système d'axiomes d'Armstrong est complet pour s_1.

Preuve 2 Nous devons montrer que si $F \models_{s_1} X \rightarrow Y$ alors $F \vdash X \rightarrow Y$ ou de façon équivalente que, si $F \nvdash X \rightarrow Y$ alors $F \not\models_{s_1} X \rightarrow Y$.

Par conséquent, en supposant que $F \nvdash X \rightarrow Y$, il suffit de donner une relation r sur U telle que $r \models_{s_1} F$ mais $r \not\models_{s_1} X \rightarrow Y$.

X^+			$U - X^+$		
a	...	a	b	...	b

TAB. 2 – Contre-exemple

Soit r la relation décrite dans la Tab. 2, avec $a \in [\epsilon_1, \epsilon_2]$ et $b \notin [\epsilon_1, \epsilon_2]$.

Premièrement, nous devons montrer que $r \models_{s_1} F$. Pour cela, nous allons supposer le contraire i.e. $r \not\models_{s_1} F$ et ainsi, $\exists V \rightarrow W \in F$ telle que $r \not\models_{s_1} V \rightarrow W$. On en déduit par la construction de r que $V \subseteq X^+$ et que $\exists A \in W$ tel que $A \in U - X^+$. Puisque $V \in X^+$, nous avons $F \vdash X \rightarrow V$ et puisque $F \vdash V \rightarrow W$, nous avons $F \vdash V \rightarrow A$. A partir de là, $F \vdash X \rightarrow A$ par l'axiome de transitivité et donc $A \in X^+$. Cela mène à une contradiction puisque $A \in W$, donc $r \models_{s_1} F$.

Deuxièmement, nous devons montrer que $r \not\models_{s_1} X \rightarrow Y$. Supposons là encore le contraire, i.e. $r \models_{s_1} X \rightarrow Y$. On en déduit par la construction de r que $Y \subseteq X^+$ et donc que $F \vdash X \rightarrow Y$. On arrive à une contradiction puisque par hypothèse $F \nvdash X \rightarrow Y$, et donc $r \not\models_{s_1} X \rightarrow Y$. □

Comme on peut s'y attendre, toute sémantique ne va pas tomber dans ce cadre. Considérons à présent l'exemple d'une sémantique, notée s_d, qui n'est pas "bien-formée" :

Definition 9 Soient $X, Y \subseteq U$ et r une relation sur U. La *satisfaction* de $X \rightarrow Y$ dans r pour la sémantique s_d définie avec deux seuils ϵ_1 et ϵ_2, notée $r \models_{s_d} X \rightarrow Y$, est définie par : $r \models_{s_d} X \rightarrow Y$ ssi $\forall t_1, t_2 \in r$, si $\epsilon_1 \leq d(t_1[X], t_2[X]) \leq \epsilon_2$ alors $\epsilon_1 \leq d(t_1[Y], t_2[Y]) \leq \epsilon_2$, où d est la distance euclidienne entre deux vecteurs.

Pour reformuler cette définition dans le cadre de la définition 2, il suffit de prendre :

1. $d_c(r') = [r' = \{t_1, t_2\}$ avec $t_1, t_2 \in r]$.
2. $Pred(X, \{t_1, t_2\}) = [\epsilon_1 \leq d(t_1[X], t_2[X]) \leq \epsilon_2]$, pour $X \subseteq U$.

Nous avons le résultat suivant :

Théorème 3 La sémantique s_d n'est pas *bien-formée*.

Preuve 3 Nous allons montrer que l'axiome de réflexivité n'est pas juste pour cette sémantique. Considérons la relation r avec 2 lignes (t_1 et t_2) et 2 attributs (A et B) décrite dans la Tab. 3 et montrons que $r \not\models_{s_d} AB \rightarrow A$.

r	A	B
t_1	2	10
t_2	5	6

TAB. 3 – Contre-exemple pour s_d

Nous avons $d(t_1[AB], t_2[AB]) = 5$ et $d(t_1[A], t_2[A]) = 3$. Avec les seuils $\epsilon_1 = 4$ et $\epsilon_2 = 10$, nous avons donc bien $r \not\models_{s_d} AB \rightarrow A$. Ainsi, le résultat est prouvé puisque l'axiome de réflexivité n'est pas juste.

5 Conclusion

Dans ce papier, nous avons proposé un cadre qui permet de caractériser des sémantiques pour les règles, ces sémantiques peuvent être spécifiées par un expert en fonction du domaine d'application considéré. Le cadre est basé sur le système d'inférence d'Armstrong à partir duquel il devient possible de raisonner sur les règles.

Nous avons montré comment ce cadre a été mis en œuvre pour des sémantiques définies sur des données d'expression de gènes. Nous avons exhibé des sémantiques qui adhérent au cadre et d'autres qui n'y adhèrent pas.

Les sémantiques proposées pour les données d'expression de gènes ont été implémentées comme une extension d'un logiciel libre dédié à l'analyse de données de biopuces MeV de TIGR[Agier *et al.*, 2004, Saeed et al, 2003].

Comme perspective de ce travail, nous pensons intéressant de dresser une typologie des différentes sémantiques existantes pour les règles. Par exemple, celles qui sont bien-formées, celles qui ne le sont pas mais pour lesquelles il existe de bonnes propriétés (e.g. l'anti-monotonie). En outre, on peut toujours s'intéresser à la définition d'autres systèmes d'inférence justes et complets pour telle ou telle sémantique jugée intéressante, quand de tels systèmes existent.

Références

[Agier *et al.*, 2003] M. Agier, V. Chabaud, J-M. Petit, V. Sylvain, C. D'Incan, V. Vidal, et Y-J. Bignon. Towards meaningful rules between genes from gene expression data. In *poster, MGED'03, Aix en Provence*, 2003.

[Agier *et al.*, 2004] M. Agier, J-M. Petit, V. Chabaud, C. Pradeyrol, Y-J. Bignon, et V. Vidal. Vers différents types de règles pour les données d'expression de gènes-application à des données de tumeurs mammaires. In *Actes du Congrès* INFORSID *'04, Biarritz*, 2004.

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the* 20^{th} *VLDB, Chile*, pages 487–499, 1994.

[Armstrong, 1974] William Ward Armstrong. Dependency structures of data base relationships. In *Proc. of the IFIP Congress 1974*, pages 580–583, 1974.

[Aussem et Petit, 2002] A. Aussem et J-M. Petit. ϵ-functional dependency inference : application to dna microarray expression data. In *BDA'02, Evry, France*, Octobre 2002.

[Bastide *et al.*, 2000] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, et L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *Computational Logic - CL 2000, London, UK, 2000*, volume 1861 of *LNCS*, pages 972–986, 2000.

[Becquet *et al.*, 2002] C. Becquet, S. Blachon, B. Jeudy, J-F. Boulicaut, et O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis : a case study on human sage data. *Genome Biology*, 3(12), 2002.

[Beeri et Berstein, 1979] C. Beeri et P.A. Berstein. Computational problems related to the design of normal form relation schemes. *ACM TODS*, 4(1) :30–59, 1979.

[Blanchard *et al.*, 2004] J. Blanchard, F. Guillet, R. Gras, et H. Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. *EGC'04, Clermont-Fd, France*, 1 :287–98, 2004.

[Cong *et al.*, 2004] G. Cong, X. Xu, F. Pan, A.K.H.Tung, et J. Yang. Farmer : Finding interesting rule groups in microarray datasets. In *Proceedings of SIGMOD'04, Paris*, 2004.

[Creighton et Hanash, 2003] C. Creighton et S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19 :79–86, 2003.

[Cristofor et Simovici, 2002] L. Cristofor et D. A. Simovici. Generating an informative cover for association rules. In *Proceedings of IEEE ICDM'02, Japan*, pages 597–600, 2002.

[Demetrovics et Thi, 1995] J. Demetrovics et V.D. Thi. Some remarks on generating Armstrong and inferring functional dependencies relation. *Acta Cybernetica*, 12(2) :167–180, 1995.

[Ganter et Wille, 1999] B. Ganter et R. Wille. *Formal Concept Analysis*. Springer-Verlag, 1999.

[Gottlob et Libkin, 1990] G. Gottlob et L. Libkin. Investigations on Armstrong relations, dependency inference, and excluded functional dependencies. *Acta Cybernetica*, 9(4) :385–402, 1990.

[Guigues et Duquenne, 1986] J-L. Guigues et V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Math. Sci. Hum.*, 24(95) :5–18, 1986.

[Kivinen et Mannila, 1995] Jyrki Kivinen et H. Mannila. Approximate inference of functional dependencies from relations. *TCS*, 149(1) :129–149, 1995.

[Li et Hamilton, 2004] G. Li et H. Hamilton. Basic association rules. In *Proceedings of the Fourth SIAM International Conference on Data Mining, Florida, USA*. SIAM, 2004.

[Lopes *et al.*, 2002] S. Lopes, J-M. Petit, et L. Lakhal. Functional and approximate dependencies mining : Databases and FCA point of view. *JETAI*, 14(2/3) :93–114, 2002.

[Luong, 2001] Viet Phan Luong. The representative basis for association rules. In *IEEE ICDM'01*, pages 639–640, 2001.

[Maier, 1980] D. Maier. Minimum covers in the relational database model. *JACM*, 27(4) :664–674, 1980.

[Mannila et Räihä, 1994] Heikki Mannila et Kari-Jouko Räihä. Algorithms for inferring functional dependencies from relations. *DKE*, 12(1) :83–99, Feb. 1994.

[Mannila et Toivonen, 1997] H. Mannila et Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery. *DMKD*, 1(3) :241–258, 1997.

[Morishita et Sese, 2000] Shinichi Morishita et Jun Sese. Traversing itemset lattice with statistical metric pruning. In *ACM PODS*, pages 226–236, 2000.

[Pensa *et al.*, 2004] R. Pensa, C. Leschi, J. Besson, et J. Boulicaut. Assessment of discretization techniques for relevant pattern discovery from gene expression data. In *BIOKDD'04*, pages 24–30, 2004.

[Saeed et al, 2003] AI. Saeed et al. TM4 : a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2) :374–78, 2003.

[Suzuki et Kodratoff, 1998] E. Suzuki et Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. In *PKDD'98*, pages 10–18, 1998.

[Tan *et al.*, 2004] P-N. Tan, V. Kumar, et J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4) :293–313, 2004.

[Ullman, 1982] J.D. Ullman. *Principles of Database Systems*. Computer Science Press, 1982.