A First Study in Causal Formal Concept Analysis

Zachary ASSOUMANI

Mardi 27 septembre 2022

Réalisé par :
Zachary ASSOUMANI

Encadrante Mines :
Fabienne BUFFET-THOMARAT

Encadrants LORIA :
Alexandre BAZIN
Miguel COUCEIRO
Amedeo NAPOLI

Cadre du stage

Stage de recherche

Loria
Laboratoire lorrain de recherche en informatique et ses applications

- 1er mars au 31 août
- Campus des Aiguillettes
- 29 équipes

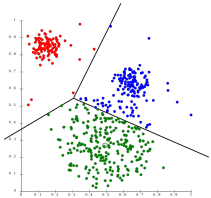
Stage de recherche

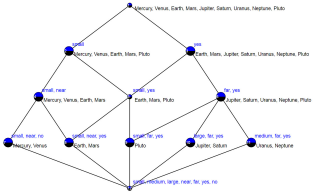
Équipe ORPAILLEUR

« Découverte et représentation de connaissances »

Base de données

⇒





Motifs et représentations

Présentation du sujet :
Une étude de la causalité en analyse formelle de concepts

Inférence causale

Que signifie « X cause Y » ?

Notions

- **Relation causale entre X et Y :** le comportement de X affecte celui de Y
- **Inférence causale :** étude des relations causales

(ex : réservoir vidé $\rightarrow$ arrêt de la voiture)

Sujet : Inférence causale

Caractéristiques

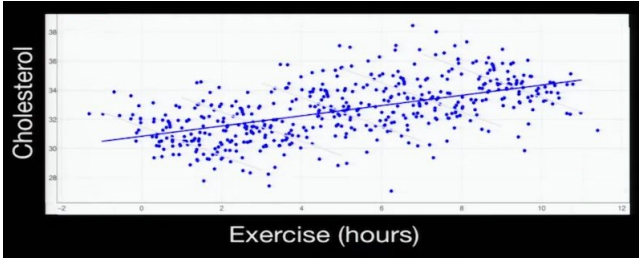
Une relation de causalité vérifie :

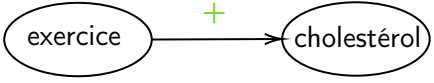
1. une **variation conjointe** entre cause et effet } *corrélation :* $P(A \cap B) \neq P(A) \cdot P(B)$
2. la cause qui **précède** l'effet
3. **l'élimination d'autres causes** plausibles.

(Shaughnessy, Zechmeister & Zechmeister in "Research Methods in Psychology") (2000)

Inférence causale

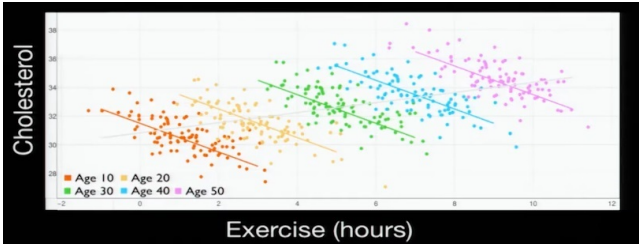
Un exemple





Inférence causale

Un exemple



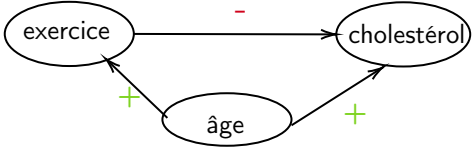


Diagramme causal

Représenter les relations causales

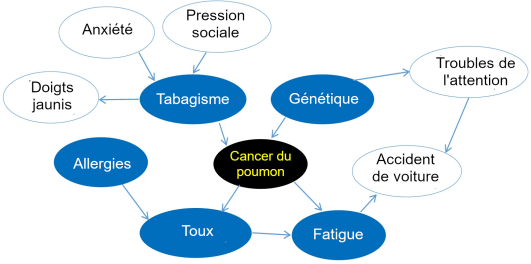


Figure – Exemple de **diagramme causal**

Source : Yu et al. in "Causality-based Feature Selection : Methods and Evaluations" (ACM Computing Surveys, 2019)

Diagramme causal

Définition

- Graphe acylique et orienté

Construire le diagramme causal

1. Mettre en place la structure du graphe
2. Trouver l'orientation des arêtes : $X \to Y$ ou $Y \to X$?

Approches :

- probabiliste : maximum de vraisemblance
- algorithmiques : $Y = f(X) + N$
- analyse formelle de concepts

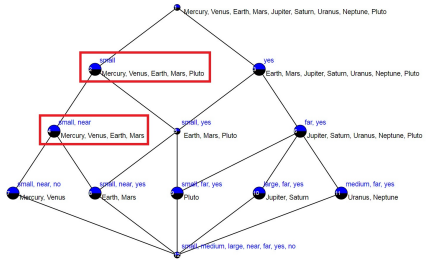
Analyse de concepts formels (FCA)

Contexte formel

Planets	Size			Distance to Sun		Moon(s)	
	small	medium	large	near	far	yes	no
Jupiter			×		×	×	
Mars	×			×		×	
Mercury	×			×			×
Neptune		×			×	×	
Pluto	×				×	×	
Saturn			×		×	×	
Earth	×			×		×	
Uranus		×			×	×	
Venus	×			×			×

⇒

Treillis de concepts



- Règles d'association : "small ⟶ near" (confiance 4/5).
- Implications : "no ⟹ near" and "near ⟹ small" (confiance 1).

Les **concepts** du treillis sont les **rectangles maximaux** du contexte.

Problématique initiale

Analyse formelle de concepts	Inférence causale
Implication	Causalité
Attributs	Variables

$\Longrightarrow$ S'inspirer de la FCA pour l'analyse causale

Source : Bazin, Couceiro, Devignes, Napoli in "Steps Towards Causal Formal Concept Analysis" (International Journal of Approximate Reasoning, 2022)

Travail bibliographique

Dans un premier temps

- Bibliographie, ouvrages clés sur la causalité ou sur FCA
- Familiarisation avec *l'état de l'art*
- Discussions avec les encadrants
- Intégration dans le projet MIRRORS

Projet Mirrors

Application pratique : projet MIRRORS : ModelIng cRop Responses tO Repeated Stresses

Collaboration entre le LORIA et l'INRAE :

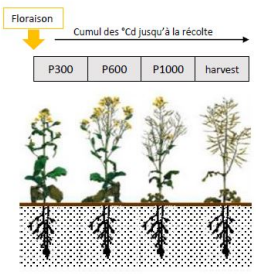
- Données agronomiques sur des plantations de colza
- 174 plantations de 2011 à 2020
- Trouver des relations de causalité entre température et rendement

Avec Lethicia MAGNO, doctorante à l'INRAE, et son encadrante Sophie BRUNEL-MUGUET.

Pré-traitement

Tri des données

Search for days with **Tmax>25°C and Tmax>30°C** in the pre-defined phenological intervals



Based on Coriouer et al. (2019)

174 rows (objects)

17 columns (attributes) = 2 qualitatives + 10 climatic + 5 plant performances

indiv	ref	local	genotype	Ptot_25	Ptot_30	P300_25	P300_30	P600_25	P600_30	P1000_25
1	LR11	LR	Aviso	6	2	0	0	1	0	2
2	LR11	LR	Aviso	6	2	0	0	1	0	2
3	LR11	LR	Aviso	6	2	0	0	1	0	2
4	LR11	LR	Montego	6	2	0	0	1	0	2
5	LR11	LR	Montego	6	2	0	0	1	0	2
6	LR11	LR	Montego	6	2	0	0	1	0	2
7	MD11	MD	Aviso	17	1	4	0	0	0	11
8	MD11	MD	Aviso	17	1	4	0	0	0	11
9	MD11	MD	Aviso	17	1	4	0	0	0	11
10	MD11	MD	Aviso	17	1	4	0	0	0	11
11	MD11	MD	Aviso	17	1	4	0	0	0	11
12	MD11	MD	Aviso	17	1	4	0	0	0	11
13	MD11	MD	Montego	17	1	4	0	0	0	9
14	MD11	MD	Montego	17	1	4	0	0	0	9
15	MD11	MD	Montego	17	1	4	0	0	0	9
16	MD11	MD	Montego	17	1	4	0	0	0	9
17	MD11	MD	Montego	17	1	4	0	0	0	9
18	MD11	MD	Montego	17	1	4	0	0	0	9
19	LR12	LR	Aviso	4	0	0	0	1	0	2
20	LR12	LR	Aviso	4	0	0	0	1	0	2
21	LR12	LR	Aviso	4	0	0	0	1	0	2
22	LR12	LR	Aviso	4	0	0	0	1	0	2
23	LR12	LR	Aviso	4	0	0	0	1	0	2
24	LR12	LR	Aviso	4	0	0	0	1	0	2
25	LR12	LR	Aviso	4	0	0	0	1	0	2

Objectif

Deux types d'indicateurs

- climat : $P_{t,a}$ = nombre de périodes avant t au-dessus de la température a
- plante : nombre de grains, teneur en protéines, lipides...

Extraire des règles de causalité $X \rightarrow Y$
avec X ensemble d'indicateurs climat
et Y ensemble d'indicateurs plante

exemple : $\{P_{300,25}, P_{1000,30}\} \rightarrow \{nb\_grains, lipides\}$

Ma démarche

Protocole

1. Rechercher des règles d'association ou des implications "candidates"
2. Sélectionner les règles ou les implications cohérentes
3. Vérifier la direction de causalité

Pour cela, on utilise trois approches :

- recherche d'**implications** par FCA
- recherche de **règles d'association**
- recherche de **biclusters**

FCA

Implications

$a \rightarrow b$ valide $\Leftrightarrow$ tous les objets ayant les attributs a possèdent les attributs b

Dériver des implications

$$\left.\begin{array}{l} \mathrm{a} \rightarrow b \\ \mathrm{b} \rightarrow c \end{array}\right\} \quad \mathrm{a} \rightarrow c$$

FCA et règles d'association

Planets	Size			Distance to Sun		Moon(s)	
	small	medium	large	near	far	yes	no
Jupiter			×		×	×	
Mars	×			×		×	
Mercury	×			×			×
Neptune		×			×	×	
Pluto	×				×	×	
Saturn			×		×	×	
Earth	×			×		×	
Uranus		×			×	×	
Venus	×			×			×

- Sur les 5 planètes lointaines, 2 sont de taille moyenne → **confiance = 2/5**
- Sur les 5 planètes sans lune, toutes sont proches du Soleil **confiance = 1**

Objectif

Rechercher les règles avec les plus hauts **degrés de confiance**

Règles d'association

Pré-traitement : scaling des données vers le binaire

- **Catégoriel :** $\{variété\} \rightsquigarrow \{variété\ 1, variété\ 2, variété\ 3...\}$
- **Numérique :** $\{lipides\} \rightsquigarrow \{lipides\_bas, lipides\_haut\}$

	Climat	Variété	$P_{tot,25}$
Plante 1	océanique	Aviso	4
Plante 2	méditerannéen	Montego	16

↓

	océanique	méditerranéen	Montego	Aviso	$P_{tot,25}$ bas	$P_{tot,25}$ haut
Plante 1	×			×	×	
Plante 2		×	×			×

Règles d'association

Extraction

- Seuil de confiance minimum : réglé à 50%
- Outil utilisé : CORON

(http://coron.loria.fr/)

```
\coron-0.8>sh core02_assrulexMODIF.sh rapsodynCoron_v1.rcf 50% 50% -alg:close -rule:closed
```

module | fichier d'entrée | seuil de support | seuil de confiance | algorithme de minage | type de règles extraites

FCA et biclustering

Lien avec FCA

Biclustering :
méthode de data mining **généralisant la FCA**,
en recherchant des rectangles maximaux dans des données tabulaires
où *toutes les valeurs ne sont pas nécessairement égales.*

Biclustering

Définition

Sous-matrices **cohérentes**

- à valeurs constantes (a)
- à lignes ou colonnes identiques (b)
- à cohérence additive (c)

2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2

(a)

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5

(b)

+3	+1	-5	+2	
1	4	5	0	2
4	7	8	3	5
3	6	7	2	4
5	8	9	4	6
2	5	6	1	3

(c)

Biclustering

4. UN VOTES IN 1969–1970[a]

State	EASE	HUNG			CHINA			KOREA			SO AF	
	1	2	3	4	5	6	7	8	9	10	11	12
USR	1	1	1	1	3	1	2	3	1	3	2	2
BGA	1	1	1	1	3	1	1	3	1	3	2	2
YUG	1	3	3	3	3	1	1	3	1	2	3	1
SYR	1	2	2	2	3	1	1	3	1	2	3	1
UAR	1	3	3	3	3	1	1	3	2	2	3	1
KEN	1	3	3	3	3	1	1	3	2	5	3	1
TAN	1	2	2	2	3	1	1	3	2	5	3	1
SEN	1	3	3	3	1	2	2	2	2	1	3	1
DAH	1	3	3	3	1	3	1	3	3	1	3	1
USA	1	3	3	3	1	3	3	1	3	1	1	3
UNK	1	3	3	3	1	1	3	2	3	1	1	3

[J. A. Hartigan in "Direct Clustering of a Data Matrix" (Journal of the American Statistical Association, 1971)]

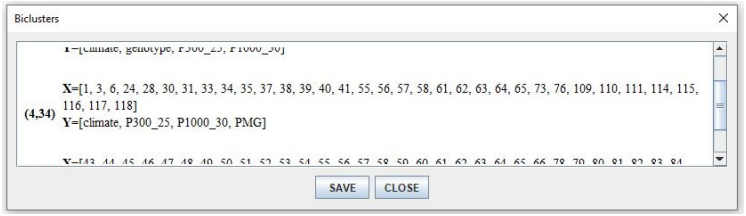
Exemple : votes au conseil de l'ONU

- Colonnes : mesures proposées
- Lignes : pays votants
- Valeur : vote (1=pour, 2=abstention, 3=contre, 5=absent)

$\Rightarrow$ Long rectangle en colonnes 11 et 12 : les votes sur ces mesures sont souvent identiques

Biclustering

Logiciel BICPAMS



Résultats : les lignes et colonnes décrivant chaque bicluster.

On obtient des ensembles d'attributs, *(ex : [climate, P300_25, P1000_30, PMG]).*

(Henriques, Ferreira, Madeira in "BicPAMS : software for biological data analysis with pattern-based biclustering" (BMC Bioinformatics, 2017))

Tri des règles

Résultats partiels

- 70 implications inférées par FCA
- 1 022 263 règles d'association (beaucoup de redondance)
- 755 règles inférées par biclustering additif
- 48 règles inférées par biclustering à colonnes constantes

De la forme suivante : $\{x_1, \ldots, x_n\} \rightarrow \{y_1, \ldots, y_m\}$

Tri des règles

Règles cohérentes

~~climat ← plante~~

climat → plante

⇒ Retirer les règles ayant un attribut plante en cause

⇒ Retirer les attributs climat de l'effet

ex : $\{climate, P_{300,25}\} \rightarrow \{lipides, nb\_graines\}$ est valide,
mais $\{climate, P_{300,25}, nb\_graines\} \rightarrow \{lipides\}$ ne l'est pas.

Tri des règles

Vérification de la direction

Algorithmes d'inférence de direction :

- Prend en entrée deux ensembles X et Y
- Infère le sens : $X \to Y$ ou $Y \to X$?

[Vreeken in "Causal Inference by Direction of Information" (SIAM International Conference on Data Mining, 2015)]

$\Rightarrow$ Ne garder que les règles dont le sens est correctement inféré.

Tri des règles

Résultats finaux

- 2 implications inférées par FCA
- 105 règles d'association
- 86 règles inférées par biclustering additif
- 1 règle inférée par biclustering à colonnes constantes

$\Rightarrow$ Lesquelles sont communes à plusieurs approches ?

Résultats

Règles communes à plusieurs approches

- **"P300_25, P1000_30 -> nb_graines"** (bicluster de taille 25)
- **"P300_25, P1000_30, harvest_25 -> poids_1000_grains"** (bicluster de taille 24)

sont inférés par biclustering additif et à colonnes constantes.

- **"P300_25, P600_25 -> recolte_grains"** (de support 53, et de confiance 67.09%)
- **"Ptot_20-25, P1000_>20, P1000_>30 -> nb_grains"** (de support 51, et de confiance 64.56%)

sont inférés par biclustering additif et par règle d'association.

Bilan collaboratif

Du côté de Lethicia

- apport des données de terrain
- explications du contexte
- choix des indicateurs

De mon côté

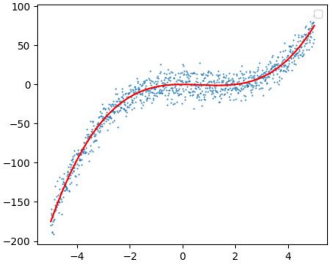
- protocole de sélection des règles
- pré-traitement et post-traitement
- mise en place de l'architecture

Un générateur de données synthétiques pour la causalité

Générateur de causalité : motivation

Deux types de données d'expérimentation

- Données de terrain (base de données médicales, sociologiques...)
- Données synthétiques (créées artificiellement)



Modèle ANM (Additive Noise Model) :

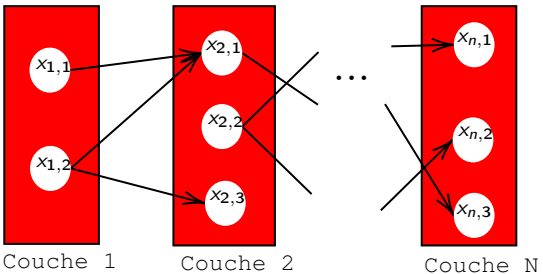
$$Y = f(X) + N$$

(Hoyer et al. in "Nonlinear causal discovery with additive noise models" (2008))

Générateur de causalité : présentation

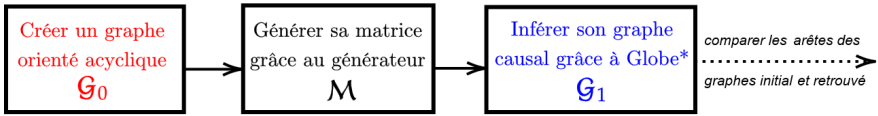
A partir du graphe...

...on génère une matrice numérique.

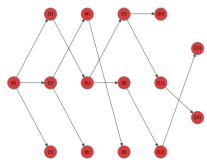


- Couche 1 : $x_{1,1}$ et $x_{1,2}$ initialisés
- Couche 2 : $x_{2,1} = f(x_{1,1}, x_{1,2}) + n_{2,1}$, $x_{2,3} = f(x_{1,2}) + n_{2,3}$
- ...
- Les attributs de la couche n sont fonction de la couche $n-1$.

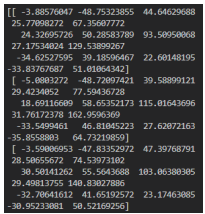
Protocole d'évaluation



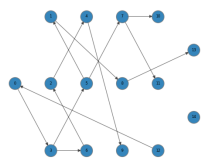
*(*Mian et al. in "Discovering Fully Oriented Causal Networks" (2021))*



$\Rightarrow$



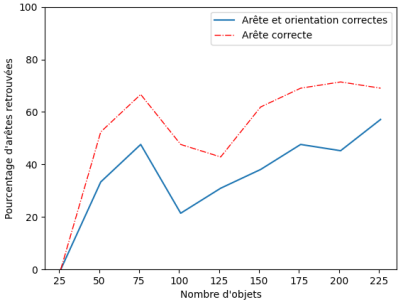
$\Rightarrow$



$\Longrightarrow$ Parmi les arêtes du graphe initial, quelle proportion est retrouvée dans le graphe final ?

Générateur de causalité : évaluation

Résultats



$\Rightarrow$ Plus il y a d'objets observés, plus la matrice générée est adéquate.

Possibles améliorations :

- $Y = f(X) + N$ avec différents polynômes
- Combiner plusieurs valeurs
- Probabilité d'apparition d'une arête

Conclusions

Missions accomplies

- Collaboration dans un projet avec l'INRAE
  - Partage de résultats
- Travail sur un générateur de données

Mise en perspective

- Travail interdisciplinaire
- Sujet évolutif
- Échanges avec des chercheurs & doctorants

Merci de votre attention.