Étude des mesures d'intérêt de motifs complexes : une extension à la sélection de méta-connaissances

Présentée par : *Dhouha Grissa*

Encadrants: *Engelbert Mephu Nguifo, Sylvie Guillaume (France)*

Sadok Ben Yahia (Tunisie)

Mai 31, 2012

Plan de la Présentation

I- Problématique

Objectif de l'analyse des associations

Technique d'apprentissage non supervisé qui permet de :

- Identifier des profils ou associations entre les items ou objets dans les bases de données transactionnelles, relationnelles, ou dans les entrepôts de données.
- Autrement dit, il s'agit d'identifier les items qui apparaissent souvent ensemble lors d'un évènement.

Association rules

Une règle d'association est de la forme : $X \rightarrow Y$

- $X \cap Y = \emptyset$
- X, Y sont des conjonctions de variables binaires.

$$\textbf{\textit{Règles Valides}} \begin{cases} Support(X \rightarrow Y) \geqslant min_{sup} & (fréquence) \\ Confiance(X \rightarrow Y) \geqslant min_{conf} & (force) \end{cases}$$

Avantages : Vertus algorithmiques accélératrices
Inconvénients : Obtention de règles non pertinentes.

Les mesures d'intérêt

Obtention de règles non pertinentes

Étape supplémentaire pour analyser les règles extraites

- La proposition de plusieurs mesures d'intérêt objectives
- Une soixantaine de mesures.

Les mesures d'intérêt

Obtention de règles non pertinentes

Étape supplémentaire pour analyser les règles extraites

- La proposition de plusieurs mesures d'intérêt objectives
- Une soixantaine de mesures.

Quelle mesure choisir ?

Quelle mesure choisir ?

- Étude des "*bonnes*" propriétés de mesures
- 21 propriétés

Aider l'utilisateur dans le choix de mesures complémentaires

(*Élimination des règles inintéressantes*)

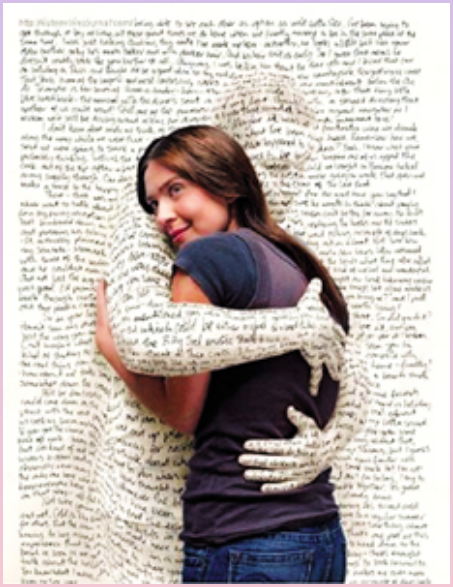
Aider l'utilisateur dans le choix de mesures complémentaires

Détection de groupes de mesures

(*Feno, 2007*), (*Heravi et Zaiane, 2010*), (*Lallich et Teytaud, 2004*), (Huynh et al. 2005), (*Y. Le Bras. 2011*)

- Catégorisation des mesures d'intérêt en utilisant la *CAH* et la méthode des *k-moyennes* ;
- Catégorisation des mesures d'intérêt utilisant *l'analyse factorielle booléenne*.

II- *État de l'art sur les mesures d'intérêt objectives*



État de l'art sur les mesures d'intérêt

Les études réalisées sur les mesures d'intérêt portent sur deux axes de recherche :

1. Étude *formelle* selon des propriétés de mesures ;
2. Étude *comparative expérimentale* du comportement des mesures.

État de l'art sur les mesures d'intérêt

Les études réalisées sur les mesures d'intérêt portent sur deux axes de recherche :

1. Étude *formelle* selon des propriétés de mesures ;
2. Étude *comparative expérimentale* du comportement des mesures.

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;
- **B. Vaillant, 2006 :** étude de 20 mesures selon 9 propriétés ;

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;
- **B. Vaillant, 2006 :** étude de 20 mesures selon 9 propriétés ;
- **Geng et Hamilton, 2007 :** étude de 38 mesures selon 11 propriétés ;

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;
- **B. Vaillant, 2006 :** étude de 20 mesures selon 9 propriétés ;
- **Geng et Hamilton, 2007 :** étude de 38 mesures selon 11 propriétés ;
- **D. Feno, 2007 :** étude de 15 mesures selon 13 propriétés ;

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;
- **B. Vaillant, 2006 :** étude de 20 mesures selon 9 propriétés ;
- **Geng et Hamilton, 2007 :** étude de 38 mesures selon 11 propriétés ;
- **D. Feno, 2007 :** étude de 15 mesures selon 13 propriétés ;
- **Heravi et Zaiane, 2010 :** étude de 53 mesures d'intérêt objectives selon 16 propriétés ;

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;
- **B. Vaillant, 2006 :** étude de 20 mesures selon 9 propriétés ;
- **Geng et Hamilton, 2007 :** étude de 38 mesures selon 11 propriétés ;
- **D. Feno, 2007 :** étude de 15 mesures selon 13 propriétés ;
- **Heravi et Zaiane, 2010 :** étude de 53 mesures d'intérêt objectives selon 16 propriétés ;
- **Lallich et Teytaud, 2004 :** étude de 15 mesures selon 13 propriétés ;

Évaluation des mesures d'intérêt selon les propriétés

- **Tan et al. 2002 :** étude de 21 mesures à travers 8 propriétés ;
- **B. Vaillant, 2006 :** étude de 20 mesures selon 9 propriétés ;
- **Geng et Hamilton, 2007 :** étude de 38 mesures selon 11 propriétés ;
- **D. Feno, 2007 :** étude de 15 mesures selon 13 propriétés ;
- **Heravi et Zaiane, 2010 :** étude de 53 mesures d'intérêt objectives selon 16 propriétés ;
- **Lallich et Teytaud, 2004 :** étude de 15 mesures selon 13 propriétés ;
- **Y. Le Bras. 2011 :** étude de 42 mesures selon 6 propriétés.

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;
- **Blanchard et al. 2005 :** classifient les mesures en : mesures de *déviation d'indépendance* et mesures de *déviation d'équilibre* ;

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;
- **Blanchard et al. 2005 :** classifient les mesures en : mesures de *déviation d'indépendance* et mesures de *déviation d'équilibre* ;
- **B. Vaillant, 2006 :** proposent de catégoriser les mesures selon la méthode hiérarchique "*CAH*" ;

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;
- **Blanchard et al. 2005 :** classifient les mesures en : mesures de *déviation d'indépendance* et mesures de *déviation d'équilibre* ;
- **B. Vaillant, 2006 :** proposent de catégoriser les mesures selon la méthode hiérarchique "*CAH*" ;
- **D. Feno, 2007 :** au sens de la normalisation, les mesures sont catégorisées selon ces trois classes :

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;
- **Blanchard et al. 2005 :** classifient les mesures en : mesures de *déviation d'indépendance* et mesures de *déviation d'équilibre* ;
- **B. Vaillant, 2006 :** proposent de catégoriser les mesures selon la méthode hiérarchique "*CAH*" ;
- **D. Feno, 2007 :** au sens de la normalisation, les mesures sont catégorisées selon ces trois classes :
  1. mesures M_{GK}*-normalisables* ;

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;
- **Blanchard et al. 2005 :** classifient les mesures en : mesures de *déviation d'indépendance* et mesures de *déviation d'équilibre* ;
- **B. Vaillant, 2006 :** proposent de catégoriser les mesures selon la méthode hiérarchique "*CAH*" ;
- **D. Feno, 2007 :** au sens de la normalisation, les mesures sont catégorisées selon ces trois classes :
  1. mesures M_{GK}*-normalisables* ;
  2. mesures *normalisables à normalisées différentes de* M_{GK} ;

Classification des mesures d'intérêt

Parmi ces études formelles sur les mesures d'intérêt, certaines ont permis de faire des classifications de ces mesures :

- **Lallich et Teytaud, 2004 :** catégorisent les mesures en deux classes : mesures *statistiques* et mesures *descriptives* ;
- **Blanchard et al. 2005 :** classifient les mesures en : mesures de *déviation d'indépendance* et mesures de *déviation d'équilibre* ;
- **B. Vaillant, 2006 :** proposent de catégoriser les mesures selon la méthode hiérarchique "*CAH*" ;
- **D. Feno, 2007 :** au sens de la normalisation, les mesures sont catégorisées selon ces trois classes :
  1. mesures M_{GK}*-normalisables* ;
  2. mesures *normalisables à normalisées différentes de* M_{GK} ;
  3. mesures *non normalisables*.

Classification des mesures d'intérêt

- **Heravi et Zaiane, 2010 :** appliquent la méthode de classification hiérarchique pour catégoriser les mesures ;

Classification des mesures d'intérêt

- **Heravi et Zaiane, 2010 :** appliquent la méthode de classification hiérarchique pour catégoriser les mesures ;
- **Y. Le Bras. 2011 :** catégorise les mesures selon des propriétés, e.g. déterminer celles qui vérifient ou pas la propriété d'*antimonotonie*.

État de l'art sur les mesures d'intérêt

Les études réalisées sur les mesures d'intérêt portent sur deux axes de recherche :

1. Étude *formelle* selon des propriétés de mesures ;
2. Étude *comparative expérimentale* du comportement des mesures.

Étude des mesures d'intérêt selon des jeux de données

- **Tan et al. 2002 :** ont analysé 21 mesures d'intérêt objectives suite au classement de $10,000$ tables de contingence synthétiques générées par les mesures et la détermination de la corrélation entre paire de mesures ;

Étude des mesures d'intérêt selon des jeux de données

- **Tan et al. 2002 :** ont analysé 21 mesures d'intérêt objectives suite au classement de 10,000 tables de contingence synthétiques générées par les mesures et la détermination de la corrélation entre paire de mesures ;
- **Huynh et al. 2005 :**
  - proposition d'une approche basée sur l'analyse du graphe de corrélation ;

Étude des mesures d'intérêt selon des jeux de données

- **Tan et al. 2002 :** ont analysé 21 mesures d'intérêt objectives suite au classement de 10,000 tables de contingence synthétiques générées par les mesures et la détermination de la corrélation entre paire de mesures ;
- **Huynh et al. 2005 :**
  - proposition d'une approche basée sur l'analyse du graphe de corrélation ;
  - développement d'un outil "*ARQAT*" (*Huynh et al. 2006*) pour évaluer et comparer visuellement le comportement des mesures ;

Étude des mesures d'intérêt selon des jeux de données

- **Tan et al. 2002 :** ont analysé 21 mesures d'intérêt objectives suite au classement de 10,000 tables de contingence synthétiques générées par les mesures et la détermination de la corrélation entre paire de mesures ;
- **Huynh et al. 2005 :**
  - proposition d'une approche basée sur l'analyse du graphe de corrélation ;
  - développement d'un outil "*ARQAT*" (*Huynh et al. 2006*) pour évaluer et comparer visuellement le comportement des mesures ;
  - étude de 34 mesures d'intérêt objectives en se basant sur leurs performances sur 120,000 règles d'association découvertes à partir de la base "*Mushroom*" ;

Étude des mesures d'intérêt selon des jeux de données

- **Tan et al. 2002 :** ont analysé 21 mesures d'intérêt objectives suite au classement de 10,000 tables de contingence synthétiques générées par les mesures et la détermination de la corrélation entre paire de mesures ;
- **Huynh et al. 2005 :**
  - proposition d'une approche basée sur l'analyse du graphe de corrélation ;
  - développement d'un outil "*ARQAT*" (*Huynh et al. 2006*) pour évaluer et comparer visuellement le comportement des mesures ;
  - étude de 34 mesures d'intérêt objectives en se basant sur leurs performances sur 120,000 règles d'association découvertes à partir de la base "*Mushroom*" ;
  - identification de onze groupes de mesures.

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.
- **Vaillant et al. 2005 :**

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.
- **Vaillant et al. 2005 :**
  - étude expérimentale de 20 mesures selon 5 bases réelles provenant de l'UCI repository ;

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.
- **Vaillant et al. 2005 :**
  - étude expérimentale de 20 mesures selon 5 bases réelles provenant de l'UCI repository ;
  - développement d'une plateforme "*Herbs*" (*Vaillant. 2002*) pour expérimenter les mesures sur des bases de règles ;

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.
- **Vaillant et al. 2005 :**
  - étude expérimentale de 20 mesures selon 5 bases réelles provenant de l'UCI repository ;
  - développement d'une plateforme "*Herbs*" (*Vaillant. 2002*) pour expérimenter les mesures sur des bases de règles ;
  - application d'une approche d'aide multicritères à la décision pour aider l'expert à choisir la mesure la mieux adaptée à ses critères ;

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.
- **Vaillant et al. 2005 :**
  - étude expérimentale de 20 mesures selon 5 bases réelles provenant de l'UCI repository ;
  - développement d'une plateforme "*Herbs*" (*Vaillant. 2002*) pour expérimenter les mesures sur des bases de règles ;
  - application d'une approche d'aide multicritères à la décision pour aider l'expert à choisir la mesure la mieux adaptée à ses critères ;
  - identification de cinq groupes de mesures.

- **Huynh et al. 2007 :** étude de 36 mesures d'intérêt sur des bases fortement corrélées et des bases faiblement corrélées $\Longrightarrow$ la découverte de 5 groupes de mesures stables.
- **Vaillant et al. 2005 :**
  - étude expérimentale de 20 mesures selon 5 bases réelles provenant de l'UCI repository ;
  - développement d'une plateforme "*Herbs*" (*Vaillant. 2002*) pour expérimenter les mesures sur des bases de règles ;
  - application d'une approche d'aide multicritères à la décision pour aider l'expert à choisir la mesure la mieux adaptée à ses critères ;
  - identification de cinq groupes de mesures.
- **Heravi et Zaiane. 2010 :** étude de 53 mesures d'intérêt selon 20 bases de données de l'UCI pour étudier l'impact des mesures sur les classifieurs associatifs $\Longrightarrow$ il n'existe pas de mesure "unique" qui a un impact sur tous les ensembles de règles pour tous les jeux de données testés.

Limites de l'existant

- Les études sont réalisées sur certaines mesures, voir les plus utilisées ;
- Évaluation de ces mesures selon un nombre restreint de propriétés ;
- Les méthodes de classification utilisées ne sont pas variées : seule la CAH est appliquée pour catégoriser les mesures ;
- La majorité des études réalisées, analysent les mesures selon des jeux de données de petite taille.

Limites de l'existant

- Les études sont réalisées sur les mesures les plus connues et les plus utilisées ;
- Évaluation de ces mesures selon un nombre restreint de propriétés ;
- Les méthodes de classification utilisées ne sont pas variées : seule la CAH est appliquée pour catégoriser les mesures ;
- La majorité des études réalisées, analysent les mesures selon des jeux de données de petite taille.

Limites de l'existant

- Les études sont réalisées sur les mesures les plus connues et les plus utilisées ;
- Évaluation de ces mesures selon un nombre restreint de propriétés ;
- Les méthodes de classification utilisées ne sont pas variées : seule la CAH est appliquée pour catégoriser les mesures ;
- La majorité des études réalisées, analysent les mesures selon des jeux de données de petite taille.

Limites de l'existant

- Les études sont réalisées sur les mesures les plus connues et les plus utilisées ;
- Évaluation de ces mesures selon un nombre restreint de propriétés ;
- Les méthodes de classification utilisées ne sont pas variées : seule la CAH est appliquée pour catégoriser les mesures ;
- La majorité des études réalisées, analysent les mesures selon des jeux de données de petite taille.

Notre contribution

- Extension du nombre de mesures et de propriétés à étudier ;
- Formalisation des propriétés ;
- Classification des mesures :
  - application d'une méthode de "*CAH*" et d'une méthode de partitionnement "*k-moyenne*" $\Longrightarrow$ obtention de groupes de mesures disjoints ;
  - application de la méthode d'analyse factorielle des données binaires (*Belohlavek et al. 2011*) $\Longrightarrow$ obtention de groupes de mesures qui se chevauchent.

Notre contribution

- Extension du nombre de mesures et de propriétés à étudier ;
- Formalisation des propriétés ;
- Classification des mesures :
  - application de la méthode hiérarchique "*CAH*" et la méthode de partitionnement "*k-moyenne*" $\Longrightarrow$ obtention de groupes de mesures disjoints ;
  - application de la méthode d'analyse factorielle des données binaires (*Belohlavek et al. 2011*) $\Longrightarrow$ obtention de groupes de mesures qui se chevauchent.

Notre contribution

- Extension du nombre de mesures et de propriétés à étudier ;
- Formalisation des propriétés ;
- Classification des mesures :
  - application de la méthode hiérarchique "*CAH*" et la méthode de partitionnement "*k-moyenne*" $\Longrightarrow$ obtention de groupes de mesures disjoints ;
  - application de la méthode d'analyse factorielle des données binaires (*Belohlavek et al. 2011*) $\Longrightarrow$ obtention de groupes de mesures qui se chevauchent.

Notre contribution

- Extension du nombre de mesures et de propriétés à étudier ;
- Formalisation des propriétés ;
- Classification des mesures :
  - application de la méthode hiérarchique "*CAH*" et la méthode de partitionnement "*k-moyenne*" $\Longrightarrow$ obtention de groupes de mesures disjoints ;
  - application de la méthode d'analyse factorielle des données binaires (*Belohlavek et al. 2011*) $\Longrightarrow$ obtention de groupes de mesures qui se chevauchent.

III- *Évaluation des propriétés sur les mesures*

Propriétés des mesures

- 21 propriétés dégagées de la littérature
- Toutes ces propriétés ont été formalisées.
- 2 propriétés jugées subjectives

  (*basées sur la connaissance de l'utilisateur en Statistique*)

  1. Compréhensibilité de la mesure ;
  2. Facilité à fixer un seuil.

19 **propriétés sont retenues**

19 propriétés

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

19 propriétés

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

Non symétrie

$$m(X \rightarrow Y) \neq m(Y \rightarrow X)$$

$$m(X \rightarrow Y) \neq m(X \rightarrow \overline{Y})$$

Oui : 1

Non : 0

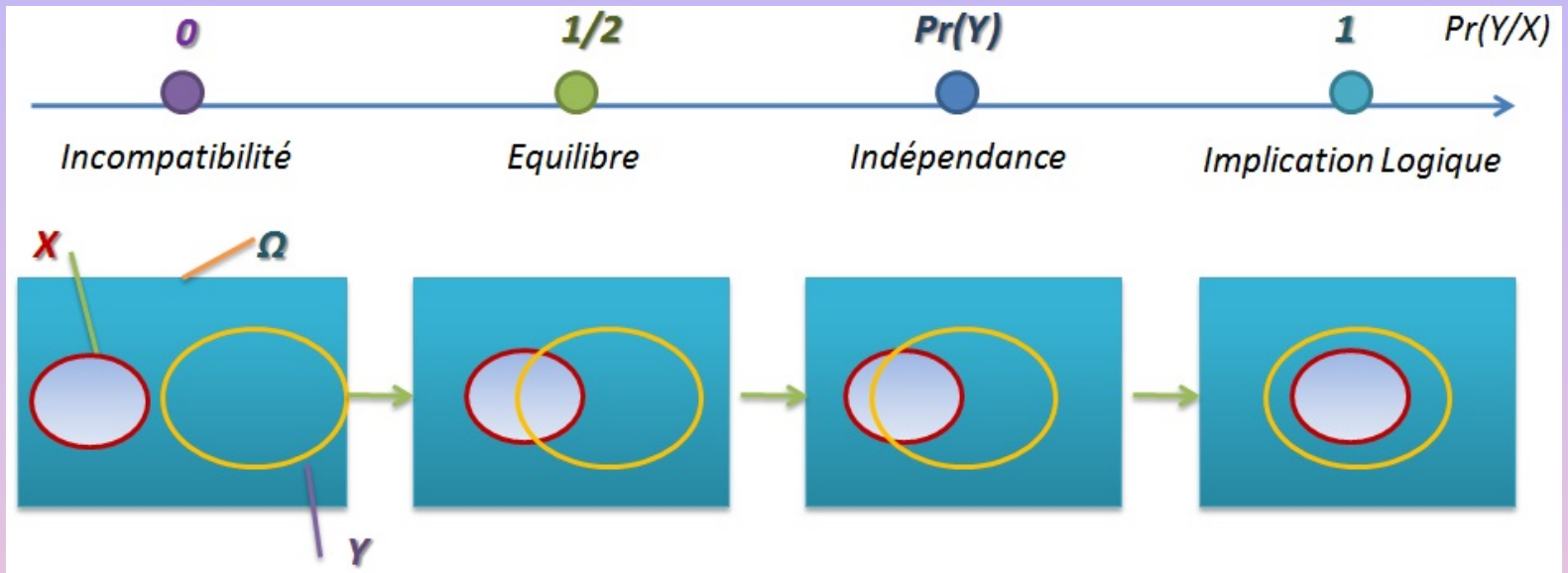
Exemple

$$Support(X \rightarrow Y) = Support(Y \rightarrow X) \Rightarrow P(XY) = P(YX)$$
$$Confiance(X \rightarrow Y) \neq Confiance(Y \rightarrow X) \Rightarrow P(Y/X) \neq P(X/Y)$$

19 propriétés

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

Valeurs fixes pour différents niveaux d'implication



$$P_{10}(m) = 0 \quad si \ \forall \ b \in \ \mathcal{R} \ \exists \ X \rightarrow Y / P(Y/X) = 1 \ et \quad m(X \rightarrow Y) \neq b$$

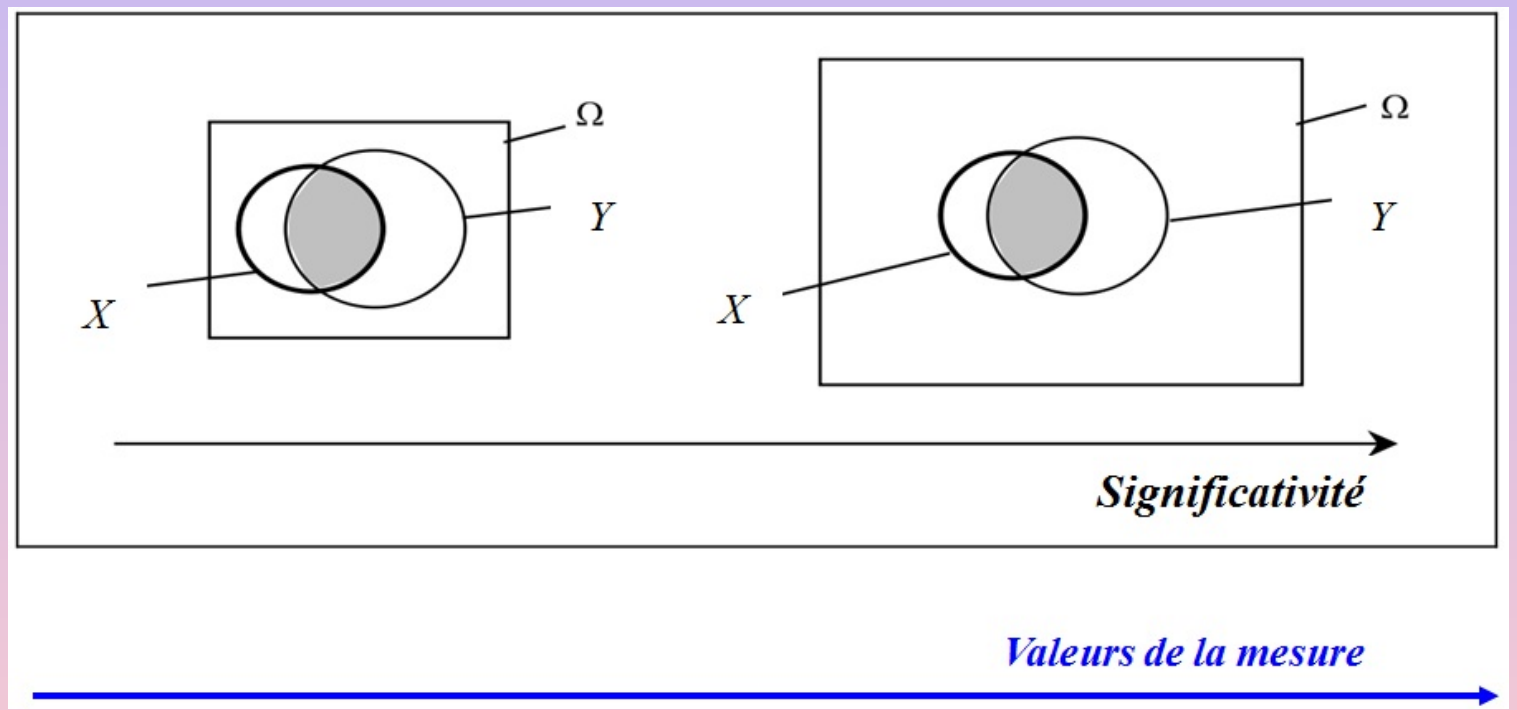
$$P_{10}(m) = 1 \quad si \ \forall \ b \in \ \mathcal{R} \ / \ \forall \ X \rightarrow Y \ P(Y/X) = 1 \ \Rightarrow \ m(X \rightarrow Y) = b$$

Oui : 1 / Non : 0

19 propriétés

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

Évolution des mesures en fonction de paramètres



19 propriétés

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

Relations entre règles positives et négatives

$$m(\overline{X} \rightarrow Y) = -m(X \rightarrow Y)$$

$$m(X \rightarrow \overline{Y}) = -m(X \rightarrow Y)$$

$$m(\overline{X} \rightarrow \overline{Y}) = m(X \rightarrow Y)$$

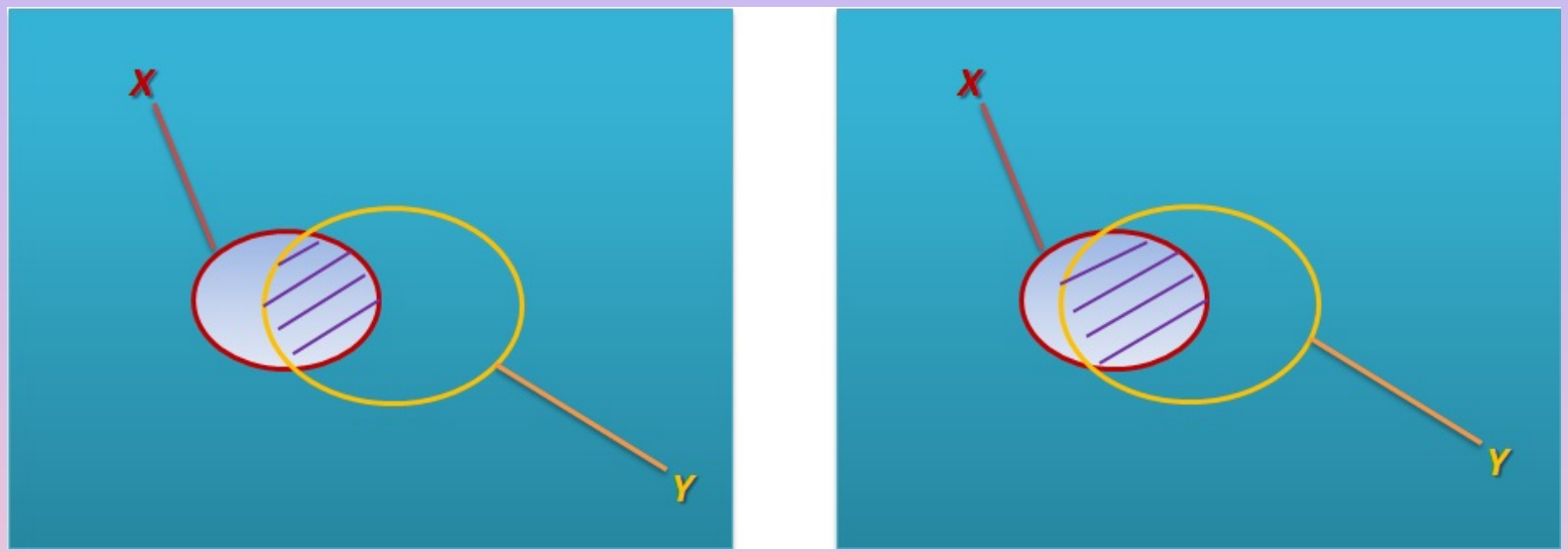
Oui : 1

Non : 0

19 propriétés

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

Discriminante en présence de données volumineuses



Mesures restituant des valeurs distinctes pour des niveaux d'implication différents

19 properties

- Non symétrie
- Valeurs fixes pour différents niveaux d'implication
- Évolution des mesures en fonction de paramètres
- Relations entre règles positives et négatives
- Discriminante en présence de données volumineuses

$\Longrightarrow$ **Évaluation des propriétés sur les mesures**

Étude de 61 mesures d'intérêt

Mesure	Formule
Cohen	$2\ \frac{p(XY)-p(X)p(Y)}{p(X)p(Y)+p(X)p(Y)}$
Confiance Causale	$1-\frac{1}{2}\left(\frac{1}{p(X)}+\frac{1}{p(Y)}\right)p(XY)$
Facteur Bayésien	$\frac{p(XY)p(Y)}{p(XY)p(Y)}$
Intensité d'Implication	$p[Poisson(np(X)p(Y)) \geq p(XY)]$
Loevinger	$1-\frac{p(XY)}{p(X)p(Y)}$
Ochiai	$\frac{p(XY)}{\sqrt{p(X)p(Y)}}$
Pearl	$p(X)\lvert\frac{p(XY)}{p(X)}-p(Y)\rvert$
Y Yule	$\frac{\sqrt{p(XY)p(\overline{XY})}-\sqrt{p(XY)p(\overline{XY})}}{\sqrt{p(XY)p(\overline{XY})}+\sqrt{p(XY)p(\overline{XY})}}$

Étude de 61 mesures d'intérêt $\times$ 19 propriétés

Construction de la matrice !

Mesure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Mesure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Mesure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Mesures non symétriques.

Mesure	P3	P4	P6	P7	P8	P9	P14	P18	P20	P21
Cohen	0	1	1	1	1	1	1	1	0	1
Conf	1	1	1	0	0	0	1	0	0	1
FB	1	1	1	1	1	1	0	0	0	1
II	1	1	1	1	1	1	2	0	1	0
Jaccard	0	1	1	0	1	0	0	0	0	1
M_{GK}	1	1	1	1	0	1	1	0	0	1
Pearl	0	0	0	0	0	1	1	1	0	1
YuleY	0	1	1	1	0	1	0	1	0	1

Mesures décroissantes en fonction de la taille du conséquent.

IV- Classification des mesures d'intérêt

Catégorisation des mesures d'intérêt

1. Catégorisation des mesures d'intérêt en utilisant la *CAH* et la méthode des *k-moyennes* (*Guillaume et al. 2011*) ;
2. Catégorisation des mesures d'intérêt en utilisant *l'analyse factorielle booléenne* (*Belohlavek et al. 2011*).

Catégorisation des mesures d'intérêt

1. Catégorisation des mesures d'intérêt en utilisant la *CAH* et la méthode des *k-moyennes* (*Guillaume et al. 2011*) ;
2. Catégorisation des mesures d'intérêt en utilisant *l'analyse factorielle booléenne* (*Belohlavek et al. 2011*).

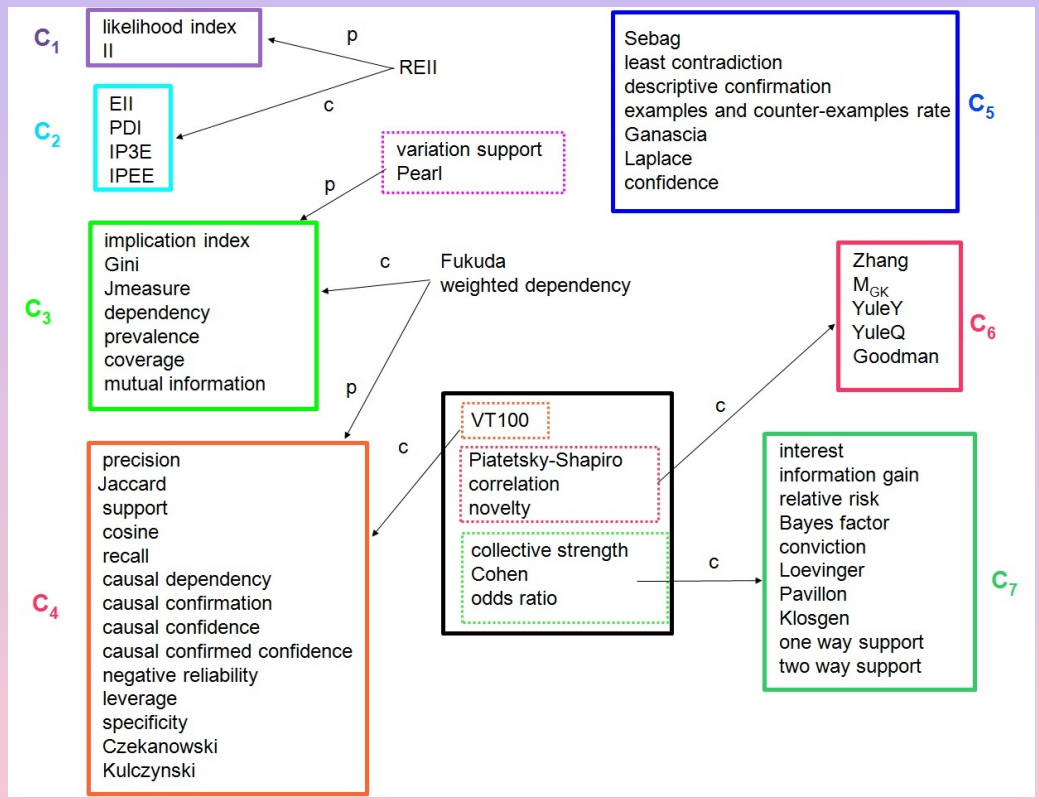
Catégorisation des mesures d'intérêt en utilisant les méthodes CAH et k-moyennes.

Simplification de la matrice : matrice de 52 *mesures* $\times$ 19 *propriétés* ;

1. Une méthode de la classification ascendante hiérarchique
   - distance euclidienne entre paires de mesures
   - distance de Ward pour la phase d'agrégation
2. Une méthode des k-moyennes
   - distance euclidienne

Catégorisation des mesures en utilisant les méthodes CAH et k-moyennes.

- *Consensus pour 7 classes*
- *Divergence pour 12 mesures*



Catégorisation des mesures d'intérêt

1. Catégorisation des mesures d'intérêt en utilisant la *CAH* et la méthode des *k-moyennes* (*Guillaume et al. 2011*) ;
2. Catégorisation des mesures d'intérêt en utilisant *l'analyse factorielle booléenne* (*Belohlavek et al. 2011*).

Catégorisation des MIs en utilisant l'analyse factorielle booléenne.

Analyse Factorielle Booléenne (AFB) = décomposition de la matrice de données binaires objet-attribut I en un produit booléen de la matrice A objet-facteur et de la matrice B facteur-attribut :

$$I_{ij} = (A \circ B)_{ij} = \max_{l=1}^{k} \min(A_{il}, B_{lj})$$

$A_{il} = 1$ . . . facteur l s'applique à l'objet i
$B_{lj} = 1$ . . . attribut j est l'un des manifestations du facteur l

$(A \circ B)_{ij}$ . . . "l'objet i possède un attribut j ssi il existe un facteur l tel que l s'applique à i et j est l'un des manifestations de l"

PROBLÈME : trouver le plus petit nombre k de facteurs que possible !

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} = \overbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}^{k} \circ \left.\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}\right\} k$$

Les matrices A et B sont construites à partir de l'ensemble $\mathcal{F}$ de concepts formels des données d'entrée I, appelés **concepts factoriels**.

Méthode

- Nous avons étendu la matrice originale d'évaluation mesure-propriété 61×21 par l'ajout pour chaque propriété de sa négation, et nous avons obtenu une matrice de mesure-propriété 61×42.
- Nous avons calculé la décomposition de la matrice en utilisant un algorithme gourmand d'approximation (à partir de l'article mentionné) et nous avons obtenu 38 facteurs, dénotés $F_1, \ldots . F_{38}$.
- Nous prenons les facteurs découverts comme étant des clusters et nous cherchons une interprétation à ces groupes.

I : 61 *mesures x* 42 *propriétés* matrice binaire d'entrée (avec des propriétés négatives)

=

	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14.1	P15	P16	P17
correlation	0	1	1	1	1	1	1	0	0	1	1	0	0	1	1
Cohen	0	1	1	1	1	1	1	0	0	1	1	0	0	0	0
confidence	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0
causal confidence	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
Pavillon	1	1	0	1	1	1	1	0	0	1	1	0	0	0	1
Ganascia	1	1	1	1	0	0	0	1	1	0	0	0	0	0	1
causal confirmation	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
descriptive confirmation	1	1	0	1	0	0	0	0	1	0	0	0	0	0	1
conviction	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0
cosine	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
coverage	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A_F : la matrice binaire
61 *mesures x* 38 *facteurs*

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24
correlation	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
Cohen	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0
confidence	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
causal confidence	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
Pavillon	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
Ganascia	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
causal confirmation	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
descriptive confirmation	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
conviction	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0
cosine	0	1	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0
coverage	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1

○

B_F : la matrice binaire
38 *facteurs x* 42 *propriétés*

	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14.1	P15	P16	P17	P18	P19	P20	P21	P14.2	P14.3	P3	P4	P5	P6
F1	0	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
F3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
F4	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F5	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
F6	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
F7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
F8	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0
F9	0	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	1	0	0	0	0
F10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Interprétation des résultats

Nous avons calculé la décomposition de la matrice *I* et nous avons obtenu 38 facteurs :

- Les premiers 21 facteurs couvrent 94% de la matrice d'entrée mesure-propriété.
- Les premiers neuf couvrent 72%.
- Les premiers cinq couvrent 52.4%.
- Les premiers dix couvrent toutes les mesures.

Couverture cumulative de la matrice d'entrée

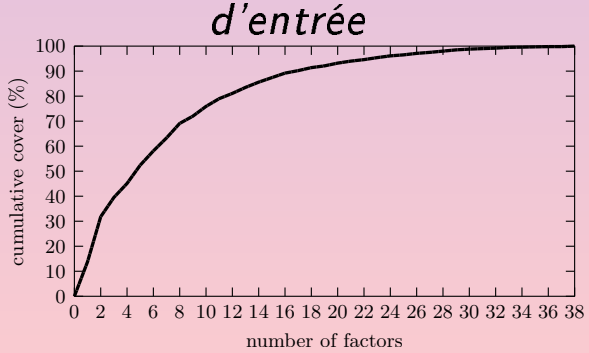
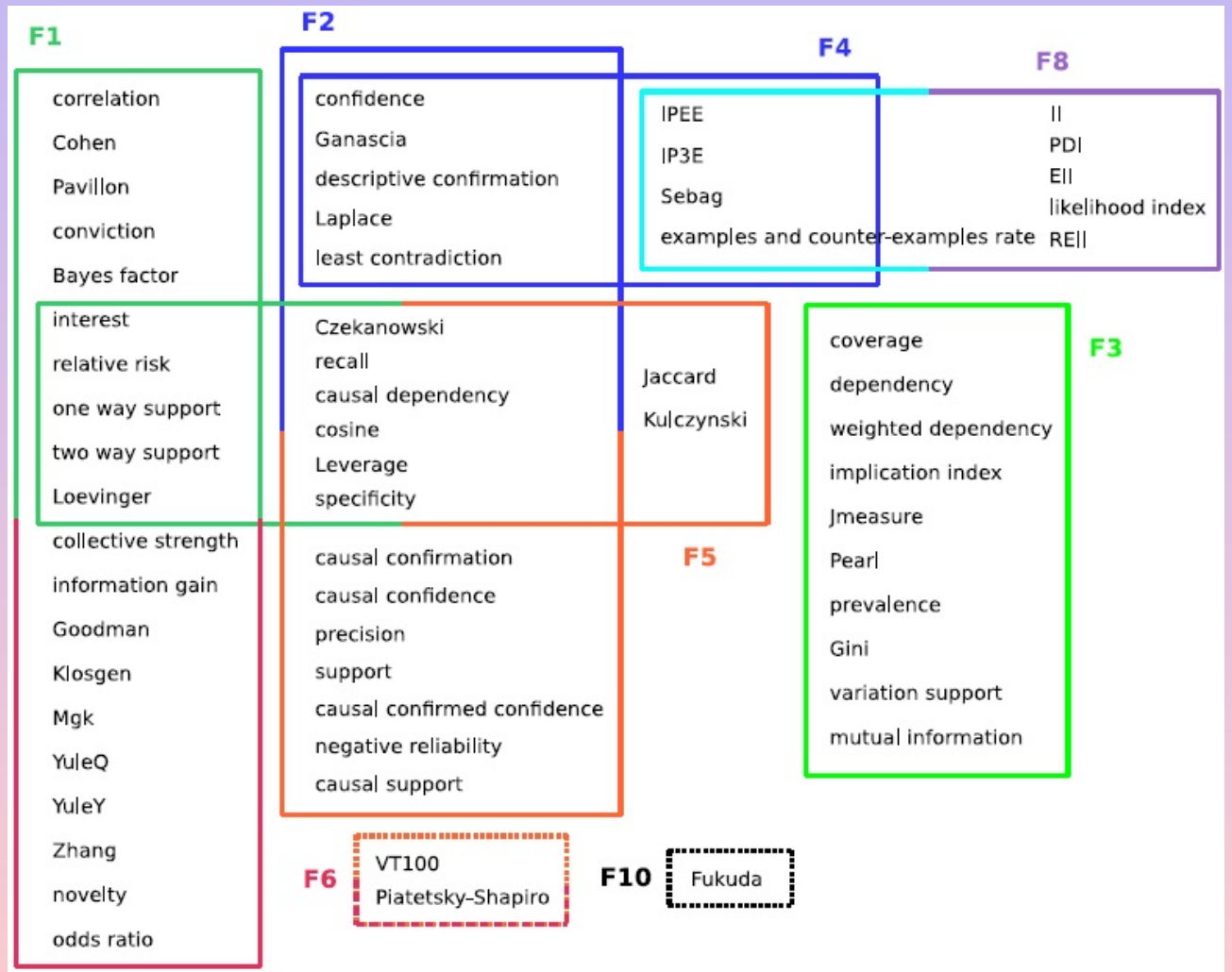


Diagramme de Venn des Facteurs Booléens



L'interprétation des premiers 4 facteurs, qui couvrent près de la moitié de la matrice (45.1%), montre :

- Une forte *similarité* avec les 7 autres classes de mesures.
- Des groupes de mesures significatifs, clairement interprétables qui se chevauchent.

V- Conclusion et Perspectives

Conclusion

- **Objectif :** aider l'utilisateur dans le choix de mesures complémentaires ;

Conclusion

- **Objectif :** aider l'utilisateur dans le choix de mesures complémentaires ;
- État de l'art sur les propriétés (*formalisation*) et sur les mesures ;

Conclusion

- **Objectif :** aider l'utilisateur dans le choix de mesures complémentaires ;
- État de l'art sur les propriétés (*formalisation*) et sur les mesures ;
- Évaluation des propriétés sur les mesures ;

Conclusion

- **Objectif :** aider l'utilisateur dans le choix de mesures complémentaires ;
- État de l'art sur les propriétés (*formalisation*) et sur les mesures ;
- Évaluation des propriétés sur les mesures ;
- Classification pour regrouper les mesures aux propriétés et aux comportements similaires :

Conclusion

- **Objectif :** aider l'utilisateur dans le choix de mesures complémentaires ;
- État de l'art sur les propriétés (*formalisation*) et sur les mesures ;
- Évaluation des propriétés sur les mesures ;
- Classification pour regrouper les mesures aux propriétés et aux comportements similaires :
  1) Catégorisation des mesures d'intérêt en utilisant la *CAH* et la méthode des *k-moyennes* (*Guillaume et al. 2011*) ;

Conclusion

- **Objectif :** aider l'utilisateur dans le choix de mesures complémentaires ;
- État de l'art sur les propriétés (*formalisation*) et sur les mesures ;
- Évaluation des propriétés sur les mesures ;
- Classification pour regrouper les mesures aux propriétés et aux comportements similaires :
  1) Catégorisation des mesures d'intérêt en utilisant la *CAH* et la méthode des *k-moyennes* (*Guillaume et al. 2011*) ;
  2) Catégorisation des mesures d'intérêt en utilisant *l'analyse factorielle booléenne* (*Belohlavek et al. 2011*).

Perspectives ?

- Chercher des liens entre les mesures en se basant sur les formules mathématiques ;

Perspectives ?

- Chercher des liens entre les mesures en se basant sur les formules mathématiques ;
- Analyser le comportement des mesures sur des jeux de données dans l'optique de valider les classes obtenues par l'étude formelle ;

Perspectives ?

- Chercher des liens entre les mesures en se basant sur les formules mathématiques ;
- Analyser le comportement des mesures sur des jeux de données dans l'optique de valider les classes obtenues par l'étude formelle ;
- Proposer une solution de mise en oeuvre de ces mesures dans un contexte applicatif ;

Perspectives ?

- Chercher des liens entre les mesures en se basant sur les formules mathématiques ;
- Analyser le comportement des mesures sur des jeux de données dans l'optique de valider les classes obtenues par l'étude formelle ;
- Proposer une solution de mise en oeuvre de ces mesures dans un contexte applicatif ;
- Étudier les propriétés spécifiques à un cadre applicatif bien déterminé : faut-il ajouter/diminuer des propriétés ?

Perspectives

- Problème de sélection des règles intéressantes : pourrions-nous proposer des critères pour comparer les groupes obtenus par l'étude formelle et dégager un groupe gagnant ;

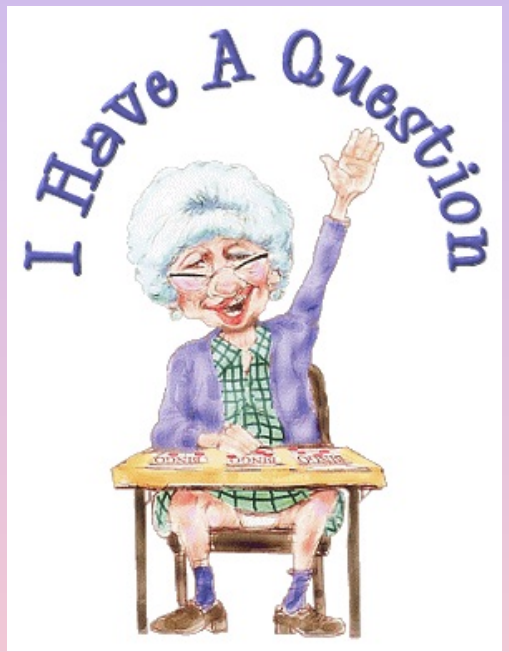
Perspectives

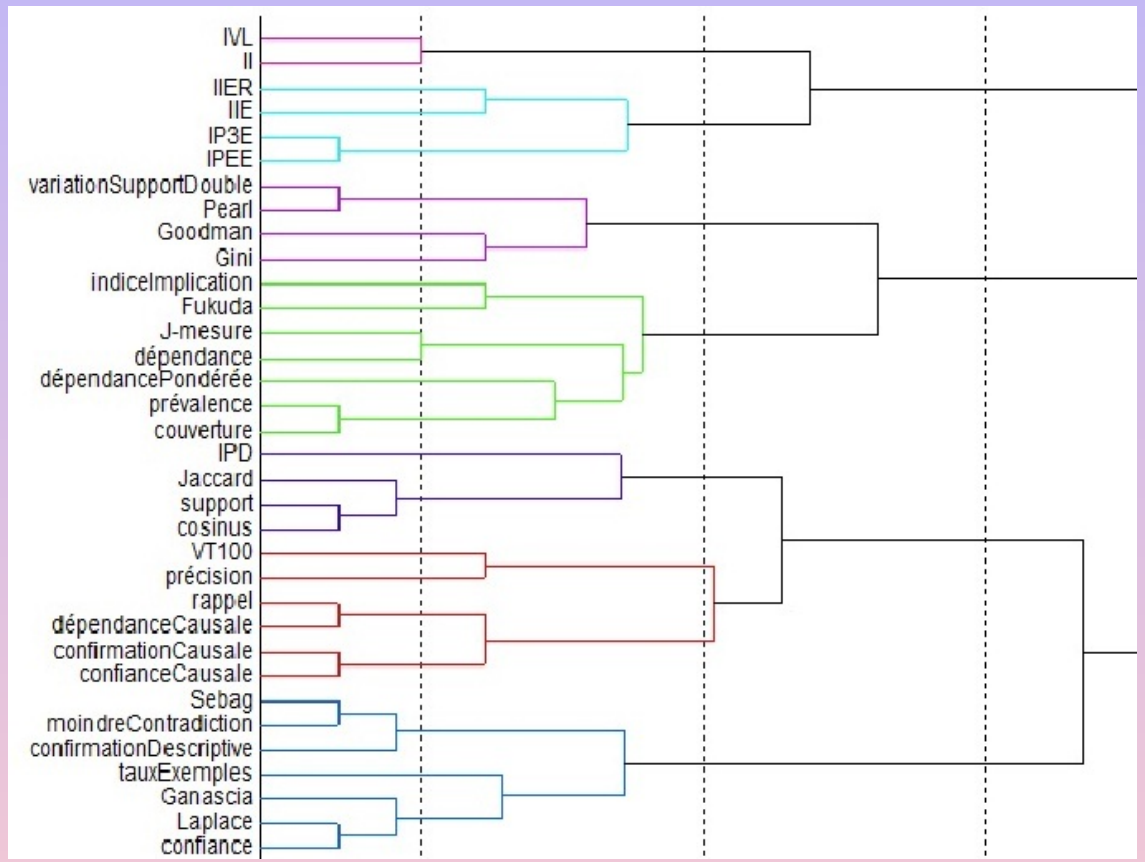
- Problème de sélection des règles intéressantes : pourrions-nous proposer des critères pour comparer les groupes obtenus par l'étude formelle et dégager un groupe gagnant ;
- Pourrions-nous étudier le degré d'appartenance/stabilité d'une mesure à une classe ?

Perspectives

- Problème de sélection des règles intéressantes : pourrions-nous proposer des critères pour comparer les groupes obtenus par l'étude formelle et dégager un groupe gagnant ;
- Pourrions-nous étudier le degré d'appartenance/stabilité d'une mesure à une classe ?
- Premiers résultats de l'étude empirique : est-ce que les mesures ayant un comportement unique sont intéressantes ?

Merci pour votre attention !




IVL
II
IIER
IIE
IP3E
IPEE
variationSupportDouble
Pearl
Goodman
Gini
indiceImplication
Fukuda
J-mesure
dépendance
dépendancePondérée
prévalence
couverture
IPD
Jaccard
support
cosinus
VT100
précision
rappel
dépendanceCausale
confirmationCausale
confianceCausale
Sebag
moindreContradiction
confirmationDescriptive
tauxExemples
Ganascia
Laplace
confiance

Return