Mesurer l'intérêt des règles d'association

Benoît Vaillant*, Patrick Meyer***, Elie Prudhomme**,
Stéphane Lallich**, Philippe Lenca*, Sébastien Bigaret*

*GET ENST Bretagne / Département LUSSI – UMR 2872 CNRS TAMCIC
{*prenom.nom*}@enst-bretagne.fr

**Laboratoire ERIC - Université Lumière - Lyon 2
lallich@univ-lyon2.fr

***Faculté de Droit, d'Economie et de Finance, Université du Luxembourg,
patrick.meyer@uni.lu

Atelier DKQ – EGC 05, 18 Janvier 2005

Résumé/Contexte

- Action Spécifique STIC Fouille de Bases de Données (GaFoDonnées), animée par R. Cicchetti et M. Sebag (2002)
- Groupe de travail sur les Mesures de Qualité (GaFoQualité), animé par F. Guillet (2002)
- laboratoires LUSSI (ENST Bretagne) et ERIC (Université Lyon 2) engagent une collaboration sur le thème de l'intérêt des règles d'association

Synthèse des travaux réalisés sur ce thème par les deux parties.

Plan

Synthèse des travaux réalisés sur ce thème par les deux parties.

- Vingt mesures, huit propriétés.
- Contrôle du risque.
- Etude formelle des mesures selon les propriétés.
- Etude de comportement, en situation (plate-forme HERBS)
- Confrontation des typologies formelle et expérimentale.
- Assistance au choix d'une mesure de qualité.

Pour des mesures d'intérêt des règles d'association

Qu'est-ce qu'une règle d'association ?

- focalise sur les coprésences [Agrawal et al., 1993]
- règle d'association, implication, équivalence et corrélation

Règles admissibles :

- Support
- Confiance

Extraction des règles :

- But : extraire les règles admissibles en totalité
- Algorithme : Apriori et dérivés

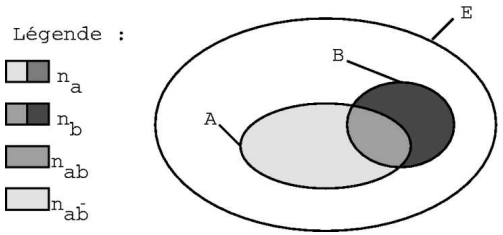
Problème : trop de règles, certaines sans intérêt

Une nécessité : mesurer l'intérêt des règles d'association

Mesures

Soit $n = |E|$, le nombre total d'enregistrements
Pour $\mathtt{A} \longrightarrow \mathtt{B}$, on note :
$n_a = |\mathtt{A}|$, le nombre d'enregistrements vérifiant $\mathtt{A}$.
$n_b = |\mathtt{B}|$, le nombre d'enregistrements vérifiant $\mathtt{B}$.
$n_{ab} = |\mathtt{A} \cap \mathtt{B}|$, le nombre d'exemples de la règle.
$n_{a\bar{b}} = |\mathtt{A} \cap \bar{\mathtt{B}}|$, le nombre de contre-exemples à la règle.
$\mathtt{A} \longrightarrow \mathtt{B}$ est évaluée à l'aide de mesures généralement **monotones décroissantes** en fonction de $n_{a\bar{b}} = n_a - n_{ab}$. $\mathtt{A} \longrightarrow \mathtt{B}$ est jugée intéressante selon la mesure μ lorsque $\mu(\mathtt{A} \longrightarrow \mathtt{B}) \geq \alpha$, α **devant être fixé par l'utilisateur**.

Pour $\mathtt{X} \subset \mathtt{E}$, on remplace n_X/n par p_X lorsque l'on considère les fréquences relatives plutôt que les fréquences absolues.



Mesures

Mesure	Abréviation	Définition
support	SUP	$\frac{n_a - n_{a\bar{b}}}{n}$
confidence	CONF	$1 - \frac{n_{a\bar{b}}}{n_a}$
linear correlation coefficient	R	$\frac{nn_{ab} - n_a n_b}{\sqrt{nn_a n_b n_{\bar{a}} \cdot n_{\bar{b}}}}$
centred confidence	CONFCEN	$\frac{nn_{ab} - n_a n_b}{nn_a}$
conviction	CONV	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$
Piatetsky-Shapiro	PS	$\frac{1}{n}(\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}})$
Loevinger	LOE	$1 - \frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$
information gain	GI	$\log(\frac{nn_{ab}}{n_a n_b})$
Sebag-Schoenauer	SEB	$\frac{n_a - n_{a\bar{b}}}{n_{a\bar{b}}}$
lift	LIFT	$\frac{nn_{ab}}{n_a n_b}$

Mesures

Mesure	Abréviation	Définition
Laplace	LAP	$\frac{n_{ab}+1}{n_a+2}$
least contradiction	MOCO	$\frac{n_{ab}-n_{a\bar{b}}}{n_b}$
odd multiplier	MC	$\frac{(n_a-n_{a\bar{b}})n_{\bar{b}}}{n_b n_{a\bar{b}}}$
example and counter example rate	TEC	$\frac{n_a-2n_{a\bar{b}}}{n_a-n_{a\bar{b}}}$
Kappa	IQC	$2\frac{nn_a-nn_{a\bar{b}}-n_a n_b}{nn_a+nn_b-2n_a n_b}$
Zhang	ZHANG	$\frac{nn_{ab}-n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$
implication index	-INDIMP	$\frac{nn_{a\bar{b}}-n_a n_{\bar{b}}}{\sqrt{nn_a n_{\bar{b}}}}$
intensity of implication	INTIMP	$P\left[poisson\left(\frac{n_a n_{\bar{b}}}{n}\right) \geq n_{a\bar{b}}\right]$
entropic intensity of implication	IIE	$\{[(1-h_1(\frac{n_{a\bar{b}}}{n})^2)\times (1-h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}\text{INTIMP}\}^{1/2}$
probabilistic discriminant index	IPD	$P\left[\mathcal{N}(0,1) > \text{INDIMP}^{CR/\mathcal{B}}\right]$

Propriétés

Distinguer $\mathtt{A} \longrightarrow \mathtt{B}$ et $A \longrightarrow \bar{B}$.
Expression des mesures en fonction de la confiance.
Relation d'équivalence *classer comme ...*

	Sémantique	Valeurs	Compétence
g_1	non symétrie	2	E_a
g_2	décroissance avec n_b	2	E_a
g_3	situation à l'indépendance	2	E_a
g_4	situation à la règle logique	2	E_a
g_5	non-linéarité autour de 0^+	3	E_r
g_6	prise en compte de n	2	E_r
g_7	facilité à fixer un seuil	2	$E_a$et E_r
g_8	intelligibilité	3	E_r

[Lallich, 2002], [Gras et al., 2004], [Lenca et al., 2004]

Contrôle du risque multiple,

[Teytaud et Lallich, 2001], [Lallich et al., 2004]

- règle significative : $p_{b/a}$ significativement supérieur à p_b
- on teste $H_0 : p_{b/a} = p_b$ contre $H_1 : p_{b/a} > p_b$
- statistique de test : $r(A, B)$, où $nr^2 = \chi^2(A, B)$
- sous H_0 : $r(A, B) \approx N(0, 1/\sqrt{n})$
- règle de décision : rejet H_0 au risque $\alpha_0 \iff$ $r(A, B) \times \sqrt{n} > u_{1-\alpha_0}$
- problème : inflation de fausses découvertes à 5%

Nécessité : contrôle du risque multiple !

Pour 10 000 règles, même si aucune n'est pertinente, on doit s'attendre à sélectionner 500 règles à tort

Contrôle du risque multiple

Réa. \ Déc.	Acc.	Rej.	Tot.
H_0 vraie	$1-\alpha$	α	1
H_1 vraie	β	$1-\beta$	1

$\Longrightarrow$

Réa. \ Déc.	Acc.	Rej.	Tot.
H_0 vraie	U	V	m_0
H_1 vraie	T	S	m_1
total	W	R	m

- nouveau critère : V_0 fausses découvertes au risque δ
  $FWER = P(V > 0) \longrightarrow UAFWER = P(V > V0)$,
- algorithme : BS_FD fondé sur le bootstrap pour évaluer la marge de fluctuation de la mesure et assurer $UAFWER = \delta$
- complexité : en $\mathcal{O}(lmn)$, où l constante due au bootstrap
- expériences : jusqu'à 50% de règles éliminées suivant la base

Perspectives

Règles différentiellement exprimées.

HERBS : une plate-forme d'expérimentation

Développements à poursuivre ...

- HERBS V1.0 [Vaillant, 2002], GAFOQUALITÉ
- HERBS V2.0 [Vaillant et al., 2005], démonstration à EGC 05

Deux types d'analyse :

- intra-mesure
- inter-mesures

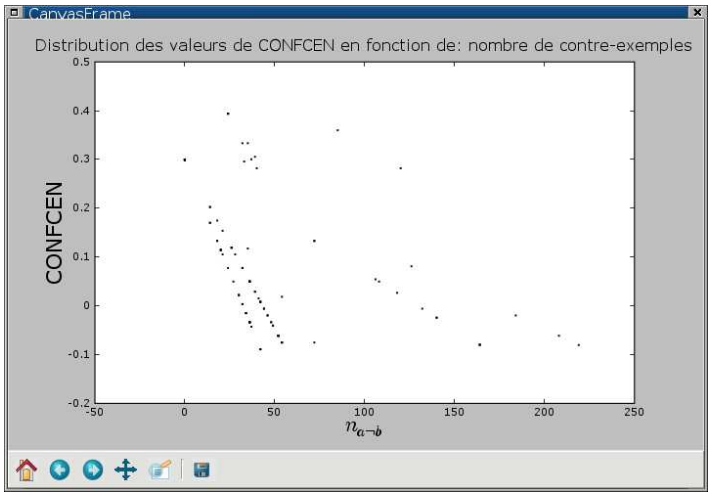
Post-analyse :

- données et règles sont des entrées
- format C4.5, APRIORI et CSV

HERBS : une plate-forme d'expérimentation

Etude d'une mesure :

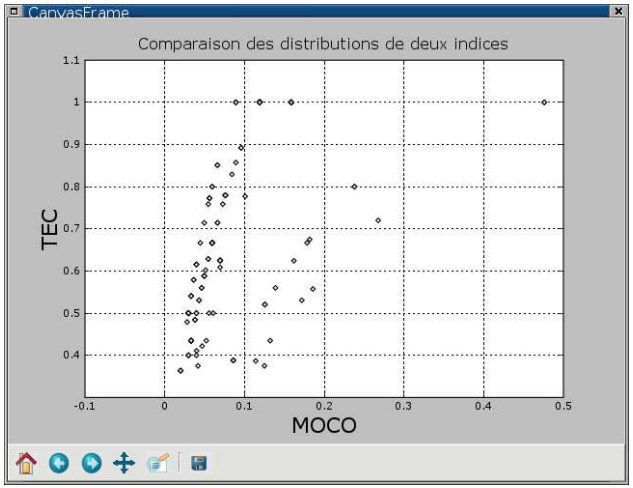
- évaluation des objets au moyen de quelques grandeurs : nombre de cas et de règles, taux de couverture, indice de recouvrement, et nombre de règles *particulières*
- sélection de l'ensemble des N meilleures règles
- tracé de la distribution des valeurs prises par la mesure



HERBS : une plate-forme d'expérimentation

Comparaison de mesures :

- extraction de l'ensemble des règles classées k fois parmi les N meilleures par p mesures.
- comparaison des préordres induits par deux mesures.
- tracé des distributions croisées des valeurs de deux mesures.



Typologie formelle

Classification formelle des mesures :

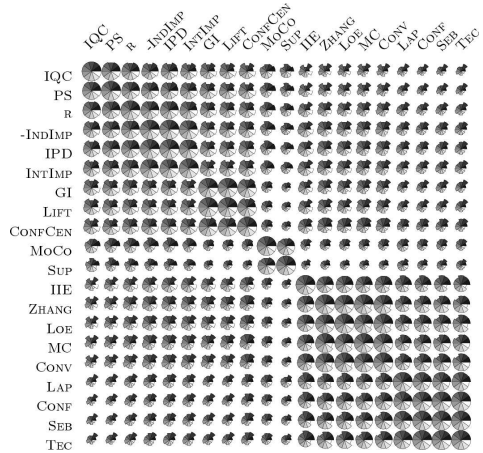
- construction d'une matrice de distance entre les mesures (matrice de décision sous forme disjonctive complète – carré des distances euclidiennes)
- classification ascendante hiérarchique avec le critère de Ward

Quatre classes principales :

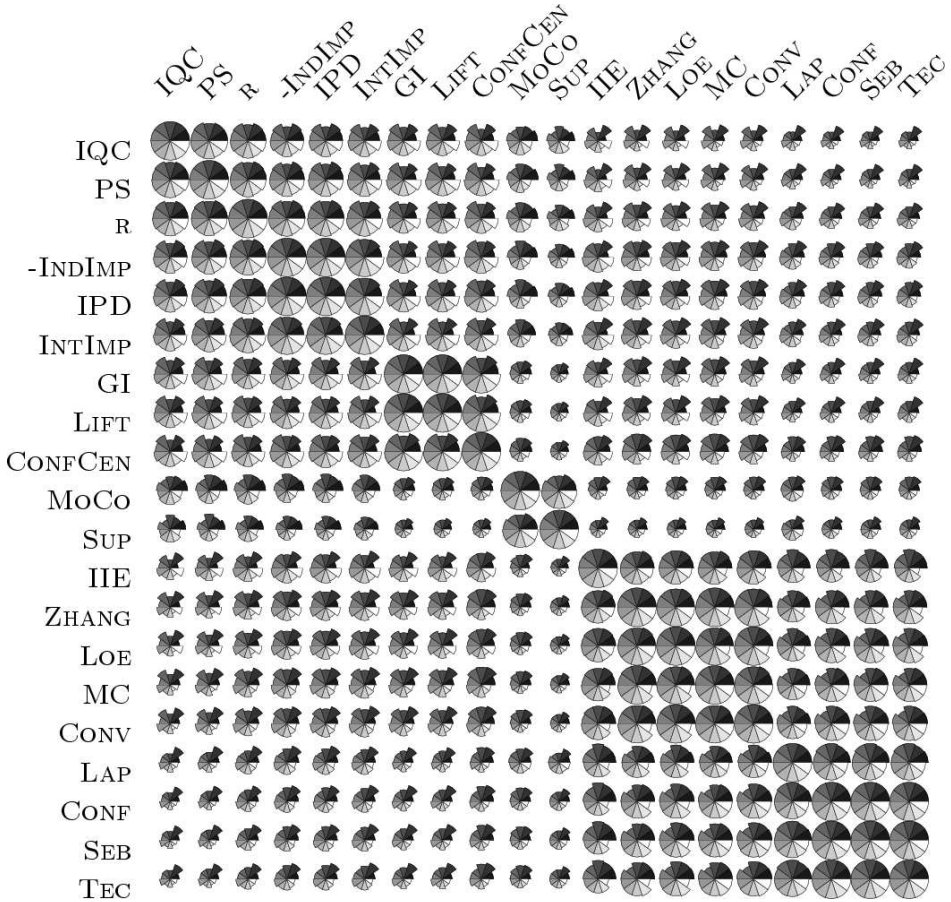
- {PS, IQC, GI, ConfCen, Lift, r, -IndImp, IPD}
- {IntImp, IIE, Loe, Zhang, MC, Conv}
- {Conf, Seb, Tec}
- {Lap, MoCo, Sup}

Typologie expérimentale

- coefficients d'accords entre préordres
- 10 bases (UCI Repository of machine learning databases)
- réarrangement de l'ordre des lignes et des colonnes afin de mieux mettre en évidence les structures de blocs (AMADO, [Chauchat et Risson, 1998])
- classification expérimentale des mesures de qualité



Typologie expérimentale



Confrontation

Du point de vue expérimental, seule la classe 3 présente de forts désaccords avec les résultats formels :

For. \ Exp.	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	PS, IQC, GI ConfCen, Lift, r -IndImp, IPD			
Classe 2	IntImp	IIE	Loe, Zhang, MC, Conv	
Classe 3			Conf, Seb, Tec	
Classe 4			Lap	MoCo, Sup

Assistance au choix des mesures

Sélectionner les *bonnes* règles c'est aussi sélectionner les *bonnes* mesures [Lenca et al., 2002], [Lenca et al., 2003] :

- Aide Multicritère à la Décision
- six éléments définissant le contexte : l'ensemble de données, l'ensemble de règles, l'ensemble de mesures, l'ensemble de propriétés des mesures, l'ensemble de préférences de l'utilisateur, l'ensemble de critères de décision [Lenca et al., 2004]
- diverses méthodes : sous-ensemble de bonnes mesures, classement des mesures

Assistance au choix des mesures

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Sup	0	0	0	0	1	0	1	2
Conf	1	0	0	1	1	0	1	2
r	0	1	1	0	1	0	1	1
ConfCen	1	1	1	0	1	0	1	2
PS	0	1	1	0	1	1	1	1
Loe	1	1	1	1	1	0	1	1
Zhang	1	1	1	1	2	0	0	0
- IndImp	1	1	1	0	1	1	1	0
Lift	0	1	1	0	1	0	1	1
MoCo	1	1	0	0	1	0	1	1
Seb	1	0	0	1	0	0	1	1
MC	1	1	1	1	0	0	1	2
Conv	1	1	1	1	0	0	1	1
Tec	1	0	0	1	2	0	1	1
IQC	0	1	1	0	1	0	1	0
GI	0	1	1	0	2	0	1	0
IntImp	1	1	1	1	2	1	1	0
IIE	1	1	1	1	2	1	0	0
IPD	1	1	1	0	1	1	1	0
Lap	1	0	0	0	1	0	1	0

Matrice de décision

g_1	non symétrie
g_2	décroissance avec n_b
g_3	situation à l'indépendance
g_4	situation à la règle logique
g_5	non-linéarité autour de 0^+
g_6	prise en compte de n
g_7	facilité à fixer un seuil
g_8	intelligibilité

Assistance au choix des mesures

- deux scénarios utilisateur (tolérance –**Sc1**– ou non –**Sc2**– de l'apparition de contre-exemples dans les règles) ; [Gras et al., 2004]
- rangements totaux pour **Sc1** et **Sc2** – poids égaux PROMETHEE [Brans et Mareschal, 1994]

Rang:	1	2	3	4	5	6	7
Sc1:	IntImp	Loe	MC	ConfCen	Conv	-IndImp,IPD	
Sc2:	MC	Conv	Loe	ConfCen	IntImp	-IndImp, IPD	
Rang:	8	9	10	11	12	13	14
Sc1:	IIE,Zhang		PS	Tec	Conf	GI	r, Lift
Sc2:	PS	Seb	Conf	r, Lift		MoCo	IIE
Rang:	15	16	17	18	19	20	
Sc1:		MoCo	Seb	IQC	Sup	Lap	
Sc2:	Zhang	IQC	Tec	Sup	GI	Lap	

Conclusion

- Double approche à la fois formelle et expérimentale.
- Diverses extensions à développer :
  - capacité discriminante d'une mesure
  - résistance au bruit
  - autres mesures . . .
- Une collaboration fructueuse !

Références

Agrawal, R., Imielinski, T., et Swami, A. (1993).
Mining association rules between sets of items in large databases.
In Buneman, P. et Jajodia, S., editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C.

Brans, J. et Mareschal, B. (1994).
The PROMETHEE-GAIA decision support system for multicriteria investigations.
Investigation Operativa, 4(2):102–117.

Chauchat, J.-H. et Risson, A. (1998).
Visualization of Categorical Data, chapter 3, pages 37–45.
Blasius J. & Greenacre M. ed.
New York : Academic Press.

Gras, R., Couturier, R., Bernadet, M., Blanchard, J., Briand, H., Guillet, F., Kuntz, P., Lehn, R., et Peter, P. (2004).
Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication.
Revue des Nouvelles Technologies de l'Information.
A paraître.

Lallich, S. (2002).
Mesure et validation en extraction des connaissances à partir des données.
Habilitation à Diriger des Recherches – Université Lyon 2.