Risques en Assurance-Finance et Modèles Statistiques
15-16 Novembre 2006, Angers

Règles d'association et détection de profils à risque

Stéphane Lallich*, Philippe Lenca**

* Laboratoire ERIC, Université Lyon 2,
** CNRS UMR 2872 TAMCIC, GET - ENST Bretagne

Plan

1. Le Data Mining
2. Représentation des données
3. Notions de base sur les règles d'association
4. L'approche support-confiance
5. Mesure de la qualité des règles
6. Les mesures statistiques
7. Identification de groupes risqués
8. Conclusion

Le Data Mining

Les nouvelles données

- Acquisition automatique
- Volumétrie en individus et variables
- Exemples : SIM,

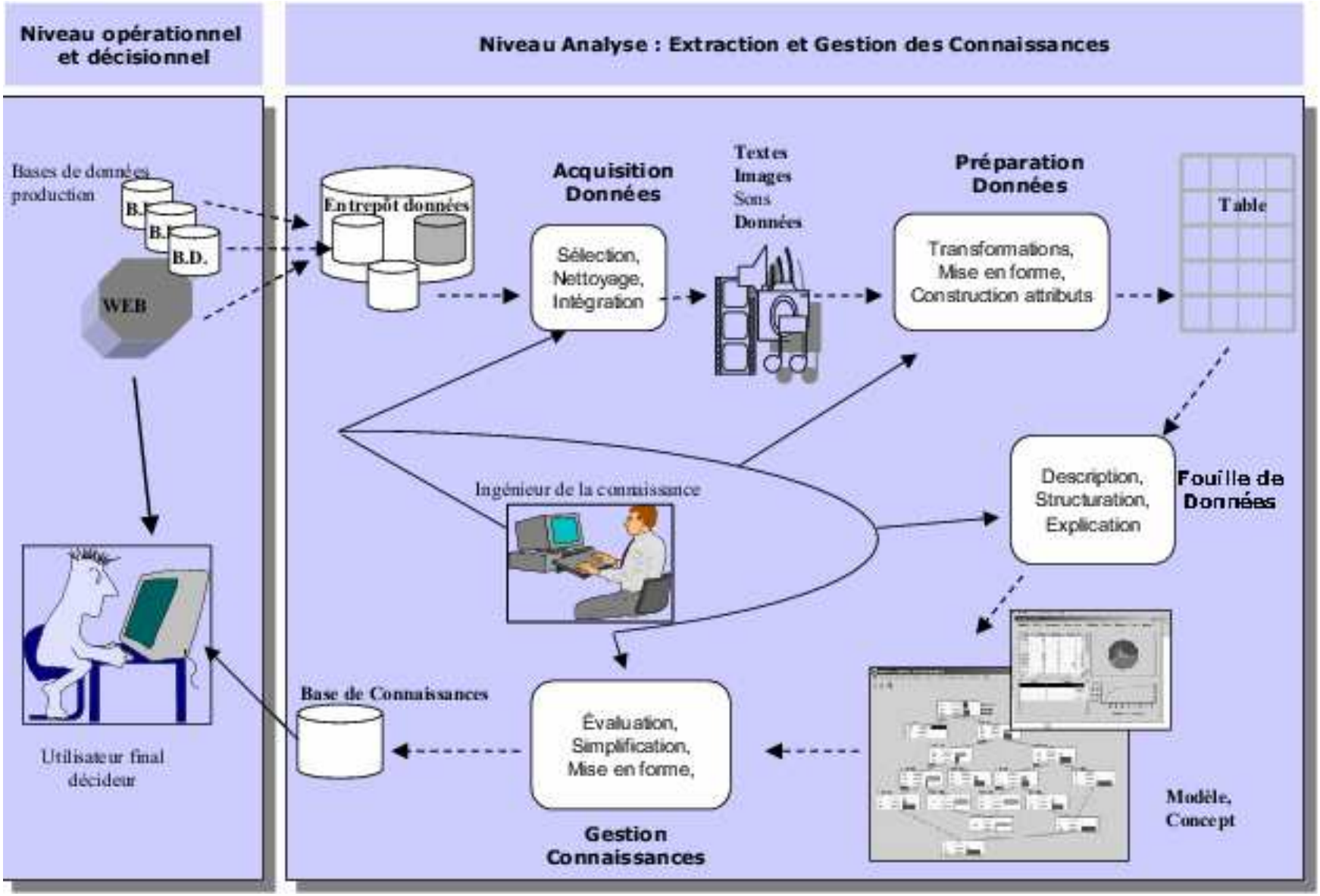
Emergence du data mining

- Data mining : combine méthodes statistiques, BD, IA
- Statistique : organise le recueil et l'analyse des données en fonction d'objectifs posés *a priori*
- Data mining : extrait l'information pertinente *a posteriori* à partir de grandes bases de données

Références

Friedman (1997), Hand (2000), Saporta (2004) …

Schéma du processus de data mining



Source : Zighed (1998)

Apprentissage

non supervisé
classification, analyse typologique
clustering

Données : n individus décrits par k descripteurs (X)

But : regrouper les individus en classes homogènes

Type de regroupement :
partitions, hiérarchies, graphes de voisinages, cartes de Kohonen,

Exemple : typologie clients

supervisé
classement
classification

Données : idem (X), plus une variable à prédire, la classe (c).

But : prédire c pour un individu dont on connaît les descripteurs

Stratégie :
- on divise les données étiquetées en LS et TS.
- on applique à LS un algorithme d'apprentissage pour trouver le meilleur classifieur
- on évalue la qualité du classifieur en généralisation (sur TS)

Représentation des données sous forme boléenne

Matrice transactionnelle

- objets = clients
- attributs = articles (items)
- x_{ij}=1 si le client i achète l'article j

Matrice boléenne : fichier client

- objets = clients
- attributs = produits financiers
- x_{ij}=1 si le client i possède le produit j

Autre : Fichier défauts

objets = véhicules, attributs = défauts

	p1	p2	p3
c_1	1	1	1
c_2	0	1	0
c_3	1	0	1
c_4	0	0	0
c_5	0	0	1
c_6	1	0	1
c_7	0	1	0
c_8	1	1	1
c_9	1	0	1
c_{10}	1	0	0

Séquences

Exemple Clickstream : objets = internautes, attributs = pages web

Représentation simple
- matrice internautes x pages
- on perd l'enchainement des pages

Représentation plus subtile
- on cherche les séquences de pages fréquentes
- matrice internautes x séquences

Multi-niveaux : prise en compte d'une hiérachie des items

OLAP : adaptation des règles aux cubes de données

partitionnement : recherche des règles par groupe d'individus

…

Autres attributs : passage en forme disjonctive complète

- **catégoriel** : type marché, Particuliers, Entreprises ou Autres

	Marché
c_1	Part.
c_2	Autre
c_3	Part.
c_4	Part.
c_5	Entr.

	P	E	A
c_1	1		
c_2			1
c_3	1		
c_4	1		
c_5		1	

- **continu :** total virements domiciliés

	Dom
c_1	1100
c_2	0
c_3	2200
c_4	800
c_5	3800

	d0	d1	d2	d3	d4
c_1			1		
c_2	1				
c_3				1	
c_4		1			
c_5					1

⇒ **On peut toujours se ramener à une matrice booléenne !**

Notions de base sur les règles d'association

Motif (ou *Itemset*)
= ensemble d'items
= conjonction de variables logiques

Longueur d'un motif
= nb d'items figurant dans le motif

Support d'un motif
= % de transactions
qui contiennent le motif

motif	long	supp
$\alpha\beta\gamma$	3	0,2
$\alpha\beta$	2	0,2
$\alpha\gamma$	2	0,5
$\beta\gamma$	2	0,2
α	1	0,6
β	1	0,4
γ	1	0,6
vide	0	1

Exemple : A = $\alpha\gamma$ ($= \alpha \cup \gamma$), Supp(A) = 0.5

Les règles d'association, AIS 93

Données

- *n* objets (transactions) décrites par *p* attributs binaires (*items*) :
- $x_{ij}=1$ si la transaction *i* contient l'item *j*.
- *A* et *B* itemsets n'ayant pas d'item commun.

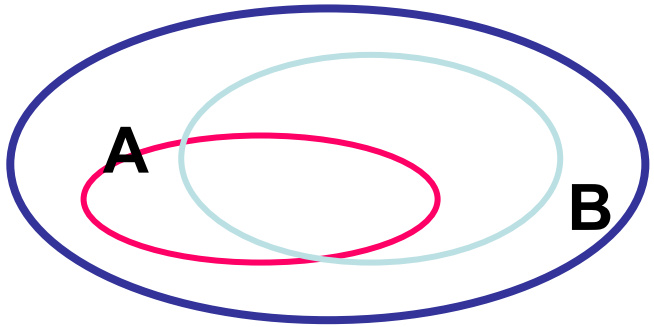
Règle d'association

- **définition :** quasi-implication du type "si les items de A figurent dans une transaction, alors ceux de B y figurent souvent"
- **notation :** A $\rightarrow$ B, **A est l'antécédent, B le conséquent**
- **généralité de la règle = support** : proportion de transactions contenant tous les items de A et B.
- **force de la règle = confiance** : proportion de transactions qui contiennent tous les items de B parmi celles contenant tous les items de A

Notations

A\B	*0*	*1*	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
total	$p_{\bar{b}}$	p_b	*1*

- A, B : itemsets
- n : nombre total d 'objets
- n_a : effectif de A (resp. B, AB)
- p_a : fréquence de A (resp. B, AB)
- support : p_{ab}
- confiance : $p_{b/a}$



Exemple : produits IARD

Hab Auto	*0*	*1*	total
0	948	22	970
1	17	13	30
total	965	35	1000

- A : Auto, B : Habitat
- n = 1000 clients banque
- $n_a = 30$, $n_b = 35$, $n_{ab} = 13$
- $p_a = 0.03$, $p_b = 0.035$
- support : $p_{ab} = 0.13$
- confiance : $p_{b/a} = {}^{13}/_{30} = 0.43$

Le "risque" d'IARD Hab est multiplié par 12 lorsque le client a un IARD Auto !

Règle d'association et corrélation, LT 04

Particularité

- une règle d'association focalise sur les coprésences,
- le 0 et le 1 ne sont pas traités de façon symétrique, au contraire de ce que fait le khi 2 ou le r^2
- ni implication logique, ni corrélation !

Exemple : même khi 2 ou r, mesures d'intérêt différentes

A \ B	*0*	*1*	total
0	65	10	75
1	5	20	25
total	70	30	100

Conf	**0,8**
Lift	**2,67**
LOE	**0,71**
BF	**9,33**

A \ B	*0*	*1*	total
0	20	5	25
1	10	65	75
total	30	70	100

Conf	**0,87**
Lift	**1,24**
LOE	**0,56**
BF	**2,79**

L'approche support-confiance

Algoritme fondateur : *Apriori, AS 94*

1. **Recherche des motifs fréquents** (Supp > min_{sup}) par balayage du treillis des itemsets.

2. Pour chaque itemset fréquent X, **on conserve les règles** du type $X \setminus Y \rightarrow Y$ (où $Y \subset X$) dont la **confiance dépasse le seuil min_{conf}.**

Antimonotonie de la condition de support

Apriori exploite l'antimonoticité de la condition de support :

- Tout sous-motif d'un motif fréquent est fréquent
- Tout sur-motif d'un motif non fréquent est non fréquent.

Améliorations algorithmiques

- **parcours en largeur :** Sampling, DIC, Partition
- **parcours en profondeur :** FP-Growth, Eclat
- **extraction des itemsets fréquents maximaux :** Close, A-Close, Pascal, MaxEclat, MaxMiner

Logiciels

- **commerciaux :** SAS Enterprise Miner, SPSS Clementine, KXEN, Statistica, Intelligent Miner
- **libres :** WEKA, TANAGRA, ORANGE, HERBS

http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_TOW_Association_Rule.pdf

Intérêt de l'approche

- Antimonotonie de la condition de support : élagage efficace
- Support et confiance : intelligibilité

Limite

- La condition de support (condition même de l'efficacité) écarte les pépites ...
- On récupère beaucoup de règles "sans intérêt"

A\B	0	1	total
0	2	18	20
1	8	72	80
total	10	90	100

- Support : $p_{ab} = 0.72$
- $p_b = 0.9$
- Confiance : $p_{b/a} = p_{ab}/p_a = 0.9$

Mesure de la qualité des règles

Quelques mesures usuelles

Critères (RNTI 04)

0 ↘ avec $n_{a \text{ non } b}$!
1. décroissance avec p_b ?
2. symétrique ou non ?
3. descriptive ou statistique ?
4. situation référence basse ?
5. courbure ?
6. borne basse ?
7. borne haute ?
8. facilité à fixer un seuil ?
9. intelligibilité ?

Confiance centrée (CC)	$Conf - p_b$
Piatesky-Shapiro (PS)	$n\, p_a\, (Conf - p_b)$
Loevinger (Loev)	$\frac{Conf - p_b}{1 - p_b}$
Indice d'implication (IndImp)	$-\sqrt{np_a}\,\frac{Conf - p_b}{1 - p_b}$
Lift	$\frac{Conf}{p_b}$
Conviction (Conv)	$\frac{1 - p_b}{1 - Conf}$
Facteur de Bayes (BF)	$\frac{Conf / p_b}{(1 - Conf)/(1 - p_b)}$

Ganascia (Gan)	$2\,(Conf - 0.5)$
Moindre Contradiction (LC)	$2\frac{p_a}{p_b}(Conf - 0.5)$
Sebag (Seb)	$\frac{Conf}{1 - Conf}$
Taux d'exemples et de contre-exemples (ECR)	$2\frac{Conf - 0.5}{Conf}$

Critères d'appréciation

- 9 critères pour apprécier les mesures, LT 04, LMPVL 04
- autres critères : résistance au bruit, pouvoir discriminant
- critères non normatifs : situation basse ? Borne max ?

Intérêt

- fait réfléchir l'utilisateur et facilite son choix
- exemple indice Jaccard (P 06) : symétrique, pas de valeur de référence basse, borne haute = p_a/p_b, biais en direction des règles $p_a = p_b$
- matrice d'évaluation *mesures x critères* qui fonde une aide automatisée au choix de la mesure la plus adaptée, LMVL 07
- typologie des mesures en fonction de la matrice d'évaluation, confrontée à une typologie de comportement réel, VLL 04

Mesures statistiques

Intérêt : prendre en compte le nombre d'observations

Principe : modélisation du nombre de contre-exemples, LGR 81

- Choix de l'hypothèse nulle H_0 : A indépendant de B, $\pi_{b/a} = \pi_b$
- Loi de $N_{a\bar{b}}$ sous H_0
- Indice statistique SI par centrage et réduction sous H_0
- Indice probabiliste PI comme probabilité cumulée de SI /H_0
- intérêt : l'indice probabiliste ramène la mesure entre 0 et 1
- 4 modélisations différentes pour écrire H_0

Exemple modélisation 3

On suppose que sous H_0 :

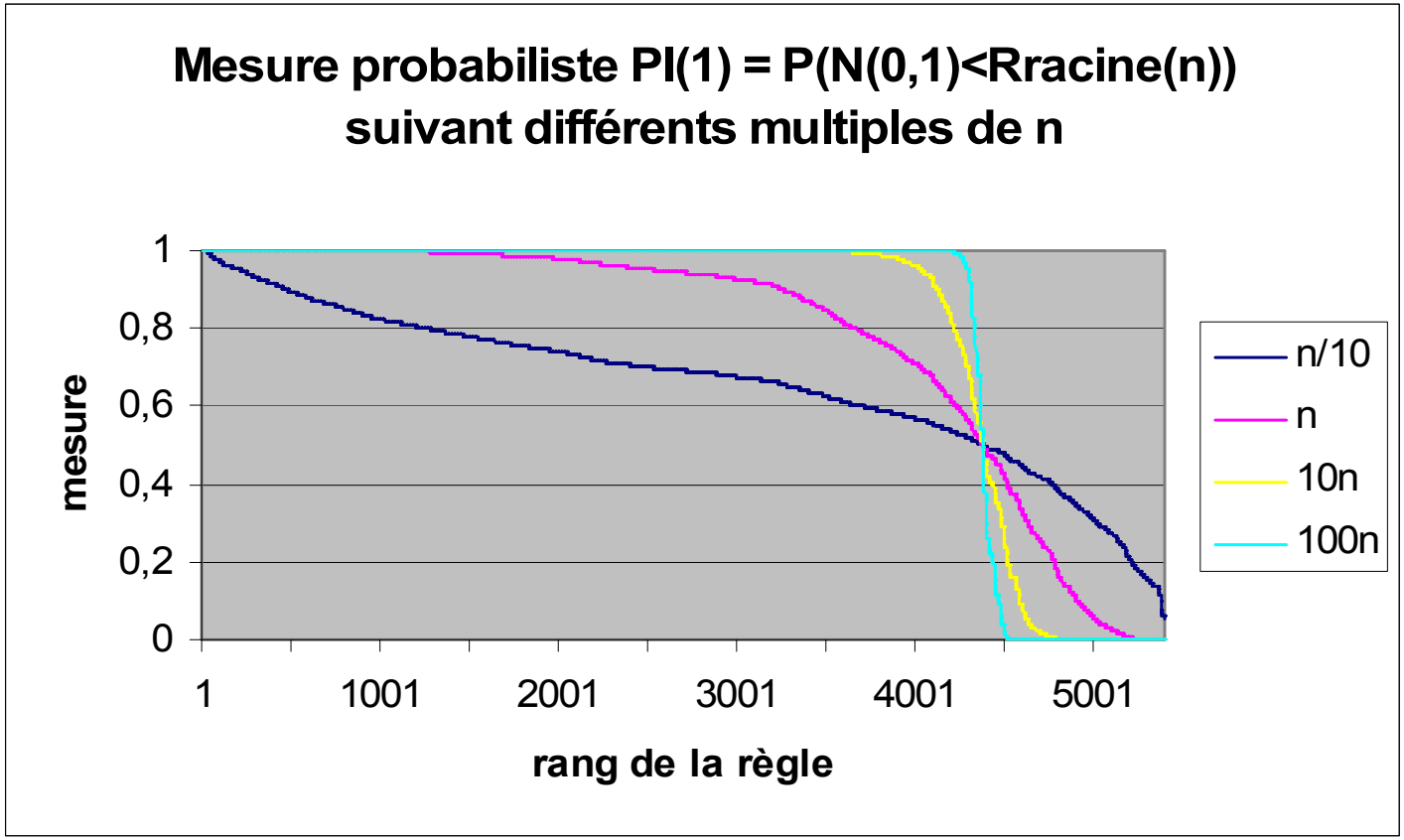
- $N \equiv Poi(n)$
- $/N = n,\ N_a \equiv B(n, p_a)$
- $/N = n,\ N_a = n_a, N_{ab} \equiv B(n_a, p_b)$

Alors :

$$SI = -INDIMP = -\frac{N_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}}}}$$

PI = INTIMP = $P(N(0,1) > INDIMP)$

Inconvénient : les mesures probabilistes sont collées à 1 !



Solution

- contextualiser SI par anamorphose gaussienne, IPD, LA 04
- pondérer PI par un indice entropique, IIE, GKCG 01

Mesures statistiques généralisées, LLV 06

Principe. La confiance est comparée à θ ou le lift à λ :

$H_0 : \pi_{b/a} = \theta$ ou $H_0 : \pi_{b/a} = \lambda\pi_b$

Construction : on suit la démarche précédente, à l'aide des fonctions génératrices

Résultat. A chacune des 4 modélisations sont associées :

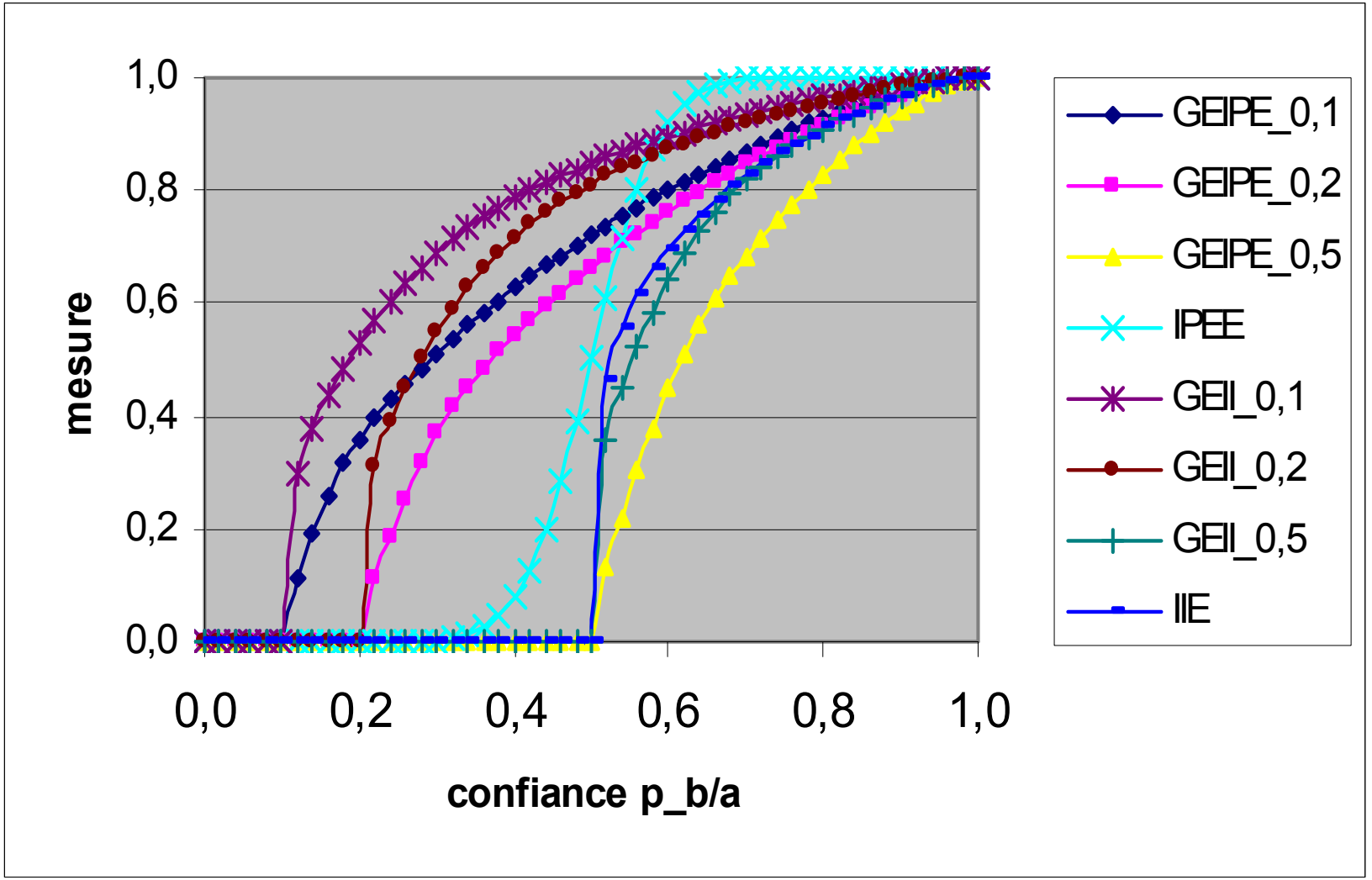
- une mesure statistique,
- une mesure probabiliste
- deux variantes discriminantes, contextualisées ou pondérées

Intérêt : fournit des mesures statistiques particulièrement utiles pour le ciblage

Comportement des mesures statistiques généralisées

$n = 1000$, $p_a = 0.05$, $p_b = 0.10$;

θ **= 0.1**, indépendance ; θ **= 0.2,** ciblage, θ **= 0.5,** prediction



Identification de groupes risqués

Groupe risqué : groupe pour lequel la probabilité de réalisation d'un itemset B est accrue

Exemple : détention IARD auto et habitat

Situation 1 : non supervisé, en aveugle

- **pas d'itemset prédéterminé**
- **choix d'une mesure d'intérêt adaptée** (μ)
- **règle extraite :** elle satisfait les seuils d'extraction et la mesure d'intérêt dépasse un seuil préfixé
- **règle A $\rightarrow$ B intéressante :** les individus pour lesquels A est vrai ont un risque plus grand que B soit vrai
- **top rules :** liste de toutes les règles intéressantes
- **Fixation du seuil pour μ :** contrôle des tests multiples (LT 04)

Situation 2 : supervisée, caractéristique prédéterminée C

Règle d'association de classe, A $\rightarrow$ C : On se limite aux règles dont le conséquent est C.

Mesure d'intérêt : toutes les règles d'association de classe ont la même valeur de p_c, donc beaucoup de mesures classent de la même façon.

Associative classification, ZD 04 : on peut prévoir si un nouvel individu est risqué en faisant voter les différentes règles retenues

Conclusion

Bilan

- Une utilisation originale des règles d'association pour détecter des groupes risqués grâce à des mesures adaptées ;
- Possibilité d'utiliser ces règles pour prévoir le risque d'un client sur une caractéristique donnée.

Perspectives

- Mise en œuvre en cours sur une extraction du fichier client d'une banque ;
- Intérêt pour des collaborations sur données réelles.

Bibliographie

- AS 94, Agrawal R. and Srikant R. (1994), Fast algorithms for mining association rules. Proceedings of the 20th VLDB, 487-499.
- AIS 93, Agrawal R., Imielinski T., Swami A.N. (1993), Mining association rules between sets of items in large databases. ACM SIGMOD, 207-216
- GKCG 01, Gras R., Kuntz P., Couturier R., Guillet F. (2001), Une version entropique de l'intensité d'implication pour les corpus volumineux. EGC 2001, 1(1-2):69-80
- LMVPL 04, Lenca P., Meyer P., Vaillant B., Picouet P., Lallich S. (2004), Evaluation et analyse multicritères des mesures de qualité des règles d'association, n° spécial *Mesures de qualité pour la fouille des données*, Revue RNTI, E-1, 219-246.
- LLV 06, Lallich S., Lenca P., Vaillant B. (2006), A probabilistic framework towards the parametrisation of association rule interestingness measures. To appear in *MCAP, Methodology and Computing in Applied Probability*.

LT 04, Lallich S., Prudhomme E., Teytaud O. (2004), Contrôle du risque multiple en sélection de règles d'association significatives. RNTI-E-2, 2:305-316

- LMVL 07, Lenca P., Meyer P., Vaillant B., Lallich S. (2007), On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. To appear in *EJOR, European Journal of Operational Research*.

- LA 04, Lerman I.C., Azé J. (2004) Indice probabiliste discriminant de vraisemblance du lien pour des données volumineuses. Revue RNTI-E-1, 69-94
- LGR 81, Lerman I.C., Gras R., Rostam H. (1981), Elaboration d'un indice d'implication pour les données binaires, i et ii. Math. et Sc. Hum., (74, 75):5-35, 5-47
- P 06, Plasse M. (2006), Utilisation conjointe des méthodes de recherche de règles d'association et de classification, thèse de doctorat CNAM, Paris
- R 03, Rakotomalala R. (2003), Tanagra. http://eric.univ-lyon2.fr/_ricco/tanagra
- RNTI 04, n° spécial "Mesures de qualité pour la fouille des données", Briand H., Sebag M., Gras R., Guillet F. (ed.) *Revue des Nouvelles Technologies de l'Information, RNTI-E-1*, pp. 219-246.
- VLL 04, Vaillant B., Lenca P., Lallich S. (2004), A clustering of interestingness measures, 7th International Conference on Discovery Science, DS'2004, Padoue, LNAI 3245, Springer-Verlag, pp. 290-297.