

Choix des conclusions et validation des règles issues d'arbres de classification

Vincent Pisetta*, Gilbert Ritschard**, Djamel A. Zighed*
 *ERIC, Univ. Lyon 2, **Dépt d'économétrie, Univ. de Genève
 EGC 2007, Namur, janvier 2007

Plan

- 1 Introduction
- 2 Arbres et indices d'implication
 - Le contexte
 - Indice implication et résidu
- 3 Pertinence individuelle des règles
- 4 Choix de la conclusion dans les feuilles
- 5 Conclusion

<http://mephisto.unige.ch>

1 Introduction

- Statistique implicative (SI)
- Outil d'analyse de données (Gras, 1979)
- Intérêt pour fouille de règles d'association (Suzuki and Kodratoff, 1998; Gras et al., 2001)
- **La SI a-t-elle un intérêt pour la classification supervisée ?**
- OUI, lorsque objectif est : caractériser profil typique de la classe
 - Exemple 1** : Médecin intéressé au profil type de personnes à risque d'un cancer, plutôt que prédire "cancer" ou "pas cancer".
 - Exemple 2** : Fisc intéressé à identifier les groupes de contribuables où il a le plus de chance de trouver des fraudeurs, plutôt que prédire "fraudeur" ou "non fraudeur".
- Paradigme de **ciblage** (typicité du profil), plutôt que classification

ASI05 toc intro impl arbre res pert choix STUL conc

◀▶▲▼ 22/1/2007gr 1

ASI05 toc intro impl arbre res pert choix STUL conc

◀▶▲▼ 22/1/2007gr 2

Application aux règles de décision

- Nous discutons ici le cas des arbres d'induction (règles de classification).
 - Indice d'implication pour règles de classification
 - Présentation comme résidu standardisé
 - Variantes issues de l'analyse de tables de contingence
 - **Validation individuelle** des règles de classification
 - **Conclusion SI optimale** des règles (alternative à la règle majoritaire)

2 Arbres et indices d'implication

2.1 Le contexte

- Jeu de données et exemple d'arbre induit
- Règles de classification et contre-exemples (notations)

ASI05 toc intro impl arbre res pert choix STUL conc

◀▶▲▼ 22/1/2007gr 3

ASI05 toc intro impl arbre res pert choix STUL conc

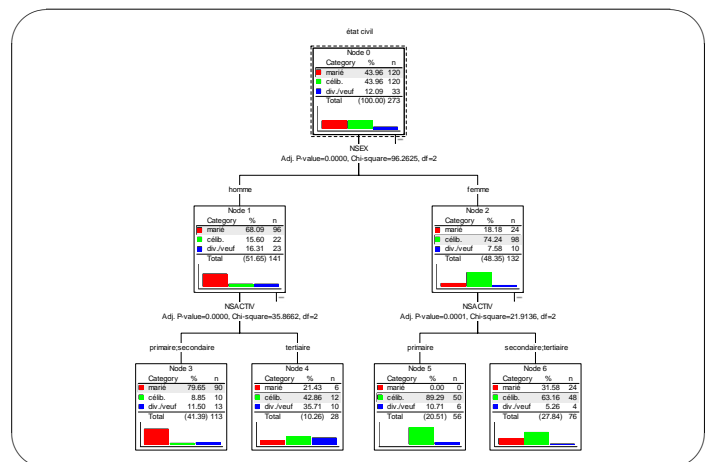
◀▶▲▼ 22/1/2007gr 4

Jeu (fictif) de 273 données

Etat civil	Sexe	Secteur activité	Nombre de cas
marié	homme	primaire	50
marié	homme	secondaire	40
marié	homme	tertiaire	6
marié	femme	primaire	0
marié	femme	secondaire	14
marié	femme	tertiaire	10
célibataire	homme	primaire	5
célibataire	homme	secondaire	5
célibataire	homme	tertiaire	12
célibataire	femme	primaire	50
célibataire	femme	secondaire	30
célibataire	femme	tertiaire	18
divorcé/veuf	homme	primaire	5
divorcé/veuf	homme	secondaire	8
divorcé/veuf	homme	tertiaire	10
divorcé/veuf	femme	primaire	6
divorcé/veuf	femme	secondaire	2
divorcé/veuf	femme	tertiaire	2

ASI05 toc intro impl arbre res pert choix STUL conc

◀▶▲▼ 22/1/2007gr 5



ASI05 toc intro impl arbre res pert choix STUL conc

◀▶▲▼ 22/1/2007gr 6

Table associée à l'arbre induit

Etat civil	Homme		Femme		total
	primaire ou secondaire	tertiaire	primaire	secondaire ou tertiaire	
Marié	90	6	0	24	120
Célibataire	10	12	50	48	120
divorcé/veuf	13	10	6	4	33
Total	113	28	56	76	273

Règles (selon classe majoritaire) :

- R1. Homme du secteur primaire ou secondaire ⇒ marié
- R2. Homme du secteur tertiaire ⇒ célibataire
- R3. Femme du secteur primaire ⇒ célibataire
- R4. Homme du secteur secondaire ou tertiaire ⇒ célibataire

Contre-exemples

Indice d'implication de Gras se définit à partir des contre-exemples.

Contre-exemple : vérifie prémisses, mais pas la conclusion (erreur de classement)

Notations :

- b conclusion de la règle j (varie avec j)
- n_b nombre totaux de cas vérifiant b , $n_{\bar{b}} = n - n_b$ (varie avec j)
- n_{bj} nombre de cas avec prémisses j qui vérifient la conclusion b
- $n_{\bar{b}j}$ nombre de contre-exemples de la règle j

H_0 Hypothèse de répartition entre b et \bar{b} indépendante de la condition

Nombre de contre-exemples sous H_0 :

$$N_{\bar{b}j} \sim \text{Poisson}(n_{\bar{b}j}^e)$$

avec $E(N_{\bar{b}j}|H_0) = \text{Var}(N_{\bar{b}j}|H_0) = n_{\bar{b}j}^e = n_{\bar{b}} n_{\cdot j} / n$. (!!! b varie avec j)

Effectifs n_{bj} et $n_{\bar{b}j}$ des contre-exemples et exemples observés

classe prédite c_{pred}	Homme		Femme		total
	primaire ou secondaire	tertiaire	primaire	secondaire ou tertiaire	
0 (contre-exemple)	23	16	6	28	73
1 (exemple)	90	12	50	48	200
Total	113	28	56	76	273

Effectifs n_{bj}^e et $n_{\bar{b}j}^e$ des contre-exemples et exemples attendus (indép.)

classe prédite c_{pred}	Homme		Femme		total
	primaire ou secondaire	tertiaire	primaire	secondaire ou tertiaire	
0 (contre-exemple)	63.33	15.69	31.38	42.59	153
1 (exemple)	49.67	12.31	24.62	33.41	120
Total	113	28	56	76	273

2.2 Indice implication et résidu

$$\text{Imp}(j) = \frac{n_{bj} - n_{\bar{b}j}^e}{\sqrt{n_{\bar{b}j}^e}}$$

Contributions au Khi2 mesurant la distance entre observés et attendus

classe prédite c_{pred}	Homme		Femme	
	primaire ou secondaire	tertiaire	primaire	secondaire ou tertiaire
0 (contre-exemple)	-5.068	0.078	-4.531	-2.236
1 (exemple)	5.722	-0.088	5.116	2.525

$$\chi^2 = \sum_j \frac{(n_{bj} - n_{\bar{b}j}^e)^2}{n_{\bar{b}j}^e} + \sum_j \frac{(n_{bj} - n_{\bar{b}j}^e)^2}{n_{\bar{b}j}^e}$$

$\text{Imp}^2(j)$

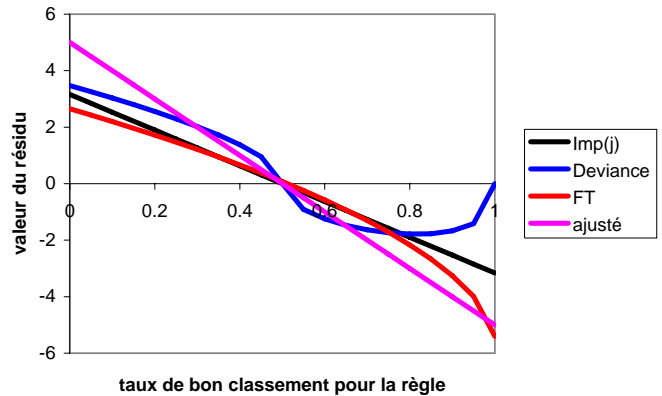
Autres résidus (utilisés pour l'ajustement de table de contingence)

- standardisé** ($=\text{Imp}(j)$) res_s contribution au Khi2 de Pearson
- déviance** res_d contribution au Khi2 du rapport de vraisemblance (Bishop et al., 1975, p 136)
- ajusté** d'Haberman res_a res_s divisé par son erreur standard (Agresti, 1990, p 224)
- Freeman-Tukey** res_{FT} stabilisation de la variance (Bishop et al., 1975, p 137)

Résidu	Règle 1	Règle 2	Règle 3	Règle 4
standardisé ($=\text{Imp}(j)$) res_s	-5.068	0.078	-4.531	-2.236
déviance res_d	-6.826	0.788	-4.456	-4.847
ajusté res_a	-9.985	0.124	-7.666	-3.970
Freeman-Tukey res_{FT}	-6.253	0.138	-6.154	-2.414

$n_{\bar{b}j}^e$ n'est qu'une estimation ⇒ variance de $\text{Imp}(j) < 1$
Et $\text{Imp}(j)$ tend à sous-estimer l'implication.
Les autres résidus sont plus proches d'une $N(0,1)$.

Valeurs indices selon taux de biens classés, $n_b/n = .5$



Degré de signification de l'indice d'implication

p -valeur de l'indice d'implication = $p(N_{ij} \leq n_{ij} | H_0)$.

Prob. que le hasard génère **moins** de contre-exemples qu'observés.

Calcul conditionnel à n_b et n_j .

- avec Poisson lorsque n petit
- approximation normale pour n grand (≥ 5)

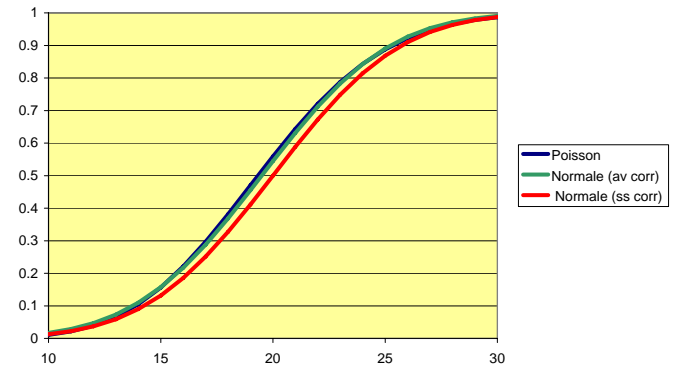
Pour l'approximation avec la normale :

correction pour la continuité

(ajouter 0,5 aux effectifs observés)

La différence peut encore atteindre 2.6 points de pourcentage pour $n_{ij} = 100$.

Distributions (cumulées) de Poisson, normale et normale avec correction



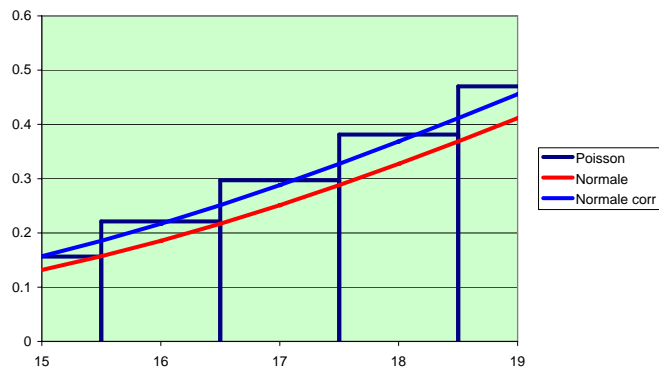
ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 13

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 14

Détail des distributions (cumulées) de Poisson, normale et normale avec correction



ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 15

Intensité d'implication

Intensité d'autant plus forte que la p -valeur est petite

⇒ **Intensité implication** = complémentaire à 1 de la p -valeur

Prob. obtenir sous H_0 **plus** de contre-exemples qu'observés

Gras et al. (2004) la définissent en termes de l'approximation normale, sans correction pour la continuité

Nous utilisons

$$\text{Intens}(j) = 1 - \phi\left(\frac{n_{ij} + 0.5 - n_{ij}^e}{\sqrt{n_{ij}^e}}\right)$$

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 16

Variantes d'intensité d'implication (avec correction continuité)

Résidu		Règle 1	Règle 2	Règle 3	Règle 4
standardisé	res_s	1.000	0.419	1.000	0.985
déviante	res_d	1.000	0.099	1.000	1.000
Freeman-Tukey	res_{FT}	1.000	0.350	1.000	0.988
ajusté	res_a	1.000	0.373	1.000	1.000

Intensité < 0.5 signifie qu'on a plus de contre-exemples qu'attendus sous H_0 .

Règle 2 sans intérêt puisqu'elle fait moins bien que le hasard.

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 17

3 Pertinence individuelle des règles

En classification, et avec les arbres en particulier, il est d'usage d'évaluer la performance du classifieur globalement avec par exemple le taux d'erreur du classifieur en généralisation.

L'**intensité d'implication** et ses variantes sont des mesures utiles pour juger de la **pertinence individuelle des règles**.

Dans notre exemple

- R1, R3 et R4 sont clairement pertinents
- R2 ne l'est pas

Que faire des règles non pertinentes? (L'ensemble des règles devant définir une partition).

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 18

Que faire des règles non pertinentes ?

1. Fusionner les cas concernés avec une autre règle.
2. Changer la conclusion de la règle.

Fusion

Dans exemple, fusion de R2 non pertinente avec la règle sœur R1

Résidu	Règle 1+2	Règle 1	Règle 2	Règle 3	Règle 4
standardisé	-3.8	-5.1	0.1	-4.5	-2.2
déviante	-7.1	-6.8	0.8	-4.5	-4.8
ajusté	-4.3	-10.0	0.1	-7.7	-3.9
Freeman-Tukey	-8.3	-6.3	0.1	-6.2	-2.4

4 Choix de la conclusion dans les feuilles

Conclusion SI-optimale :

classe pour laquelle on maximise l'intensité d'implication.

(Zighed and Rakotomalala, 2000, pp 282-287)

Exemple : choix de la conclusion pour la règle R2

Résidu		Indices			Intensité		
		marié	célib.	div./v	marié	célib	div./v
Standardisé	res_s	1.6	0.1	-1.3	0.043	0.419	0.891
Déviante	res_d	3.9	0.8	-3.4	0.000	0.099	0.999
Ajusté	res_a	2.4	0.1	-2.0	0.005	0.379	0.968
Freeman-Tukey	res_{FT}	1.5	0.1	-1.4	0.054	0.398	0.895

La conclusion "divorcé/veuf" est préférable à "célibataire" (classe majoritaire). R2 devient ainsi pertinente.

ASI05 toc intro impl arbre res pert choix STUL conc

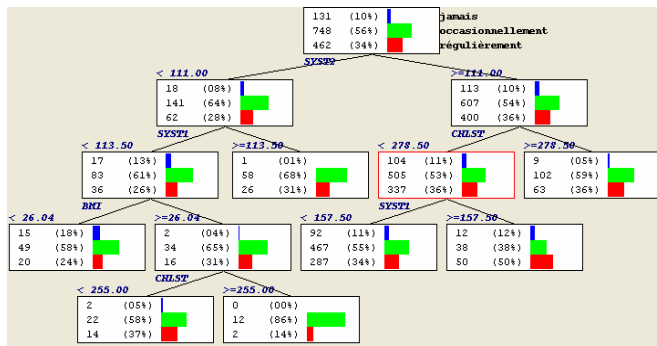
◀ ▶ ▲ ▼ 22/1/2007gr 19

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 20

Exemple : Données STULONG, Hôpital Univ. Prague

Consommation alcool : jamais, occasionnellement, régulièrement selon pression systolique 1 et 2, BMI, taux cholestérol, nbr cigarettes



ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 21

Matrices de confusion

Règle Majoritaire, taux d'erreur = 43%

état réel	prédiction		
	jamais	occasionnellement	régulièrement
jamais	0	119	12
occasionnellement	0	710	38
régulièrement	0	412	50

Maximisation Implication, taux d'erreur = 72%

état réel	prédiction		
	jamais	occasionnellement	régulièrement
jamais	107	12	12
occasionnellement	516	194	38
régulièrement	307	105	50

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 22

Conclusion selon le résidu utilisé comme critère

	Imp(j)	Déviante	Freeman-Tukey	Ajusté	Majorité
R1	2	2	2	2	2
R2	1	1	1	1	2
R3	2	3	2	3	2
R4	2	2	2	2	2
R5	2	2	2	2	2
R6	1	1	1	1	2
R7	3	3	3	3	3

1 = jamais, 2 = occasionnellement, 3 = régulièrement

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 23

5 Conclusion

– Indices et intensités d'implication complètent utilement les mesures traditionnelles de qualité d'arbres d'induction en donnant des indications précieuses sur la pertinence individuelle de chaque règle.

– Utiles dans contexte de ciblage : identifier profil type des classes.

– La conclusion SI-optimale fait apparaître que la classe modale n'est pas toujours la meilleure solution.

– L'interprétation de l'indice d'implication comme un résidu, suggère des variantes issues de la modélisation de table de contingence.

Perspectives

- Utilisation de critères SI dans la procédure de développement de l'arbre.
- Indice d'implication pénalisé pour complexité de la règle.

ASI05 toc intro impl arbre res pert choix STUL conc

◀ ▶ ▲ ▼ 22/1/2007gr 24

MERCI

Références

- Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge MA: MIT Press.
- Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques*. Thèse d'état, Université de Rennes 1, France.
- Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI E-1*, 3–30.
- Gras, R., P. Kuntz, et H. Briand (2001). Les fondements de l'analyse statistique implicative et leur prolongement pour la fouille de données. *Mathématique et Sciences Humaines* 39(154-155), 9–29.
- Suzuki, E. et Y. Kodratoff (1998). Discovery of surprising exception rules based on intensity of implication. In J. M. Zytkow et M. Quafafou (Eds.), *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, September 23-26, Proceedings*, pp. 10–18. Berlin : Springer.
- Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'induction : apprentissage et data mining*. Paris : Hermes Science Publications.