De l'usage de la statistique implicative dans les arbres de classification

Gilbert Ritschard
Dept Econométrie, Université de Genève
Journées ASI, Palerme, octobre 2005

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 ?
- 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 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)

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

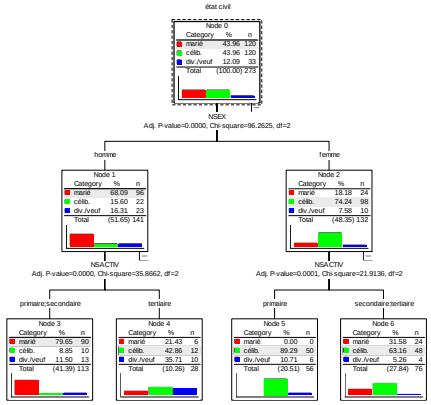


Table associée à l'arbre induit

	Homme		Femme		
Etat civil	primaire ou secondaire	tertiaire	primaire	secondaire ou tertiaire	total
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 prémisse, mais pas la conclusion (erreur de classement)

Notations :

- b conclusion de la règle j (varie avec j)
- $n_{b\cdot}$ nombre totaux de cas vérifiant b, $n_{\bar{b}\cdot} = n - n_{b\cdot}$ (varie avec j)
- n_{bj} nombre de cas avec prémisse 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 :

$$N_{\bar{b}j} \sim \text{Poisson}(n^e_{\bar{b}j})$$

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

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

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

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

classe prédite *cpred*	Homme primaire ou secondaire	Homme tertiaire	Femme primaire	Femme secondaire ou tertiaire	total
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

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

Contributions au Khi2 mesurant la distance entre observés et attendus

classe prédite *cpred*	Homme primaire ou secondaire	Homme tertiaire	Femme primaire	Femme secondaire ou tertiaire
0 (contre- exemple)	**-5.068**	**0.078**	**-4.531**	**-2.236**
1 (exemple)	5.722	-0.088	5.116	2.525

$$\chi^2 = \sum_j \underbrace{\frac{(n_{\bar{b}j} - n^e_{\bar{b}j})^2}{n^e_{\bar{b}j}}}_{\text{Imp}^2(j)} + \sum_j \frac{(n_{bj} - n^e_{bj})^2}{n^e_{bj}}$$

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

standardisé (=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_{TF}	stabilisation de la variance (Bishop et al., 1975, p 137)

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

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

Degré de signification de l'indice d'implication

p-valeur de l'indice d'implication $= p(N_{\bar{b}j} \geq n_{\bar{b}j}|H_0)$.

Prob. que le hasard génère plus de contre-exemples que le nombre observé

Calcul conditionnel à $n_{b\cdot}$ et $n_{\cdot j}$.

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

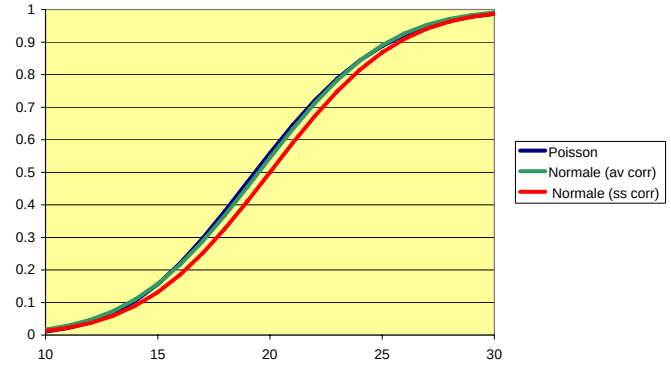
Pour l'approximation avec la normale :

correction pour la continuité

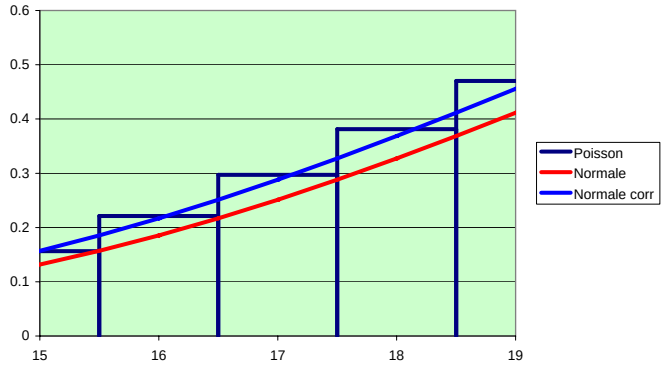
(ajouter 0,5 aux effectifs observés)

La différence peut encore atteindre 2.6 points de pourcentage pour $n_{\bar{b}j} = 100$.

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



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



Intensité d'implication

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

$\Rightarrow$ Intensité implication $=$ complémentaire à 1 de la p-valeur

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

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

Nous utilisons

$$\text{Intens}(j) \quad = \quad 1 - \phi\Big(\frac{n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j}}{\sqrt{n^e_{\bar{b}j}}}\Big)$$

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
déviance	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 que le nombre attendu sous H_0.

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

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).

Taux d'erreur et indice d'implication

nombre d'erreurs = nombre de contre-exemples

Taux d'erreur de la règle j :

$$\text{err}(j) \quad = \quad \frac{n_{\bar{b}j}}{n_{\cdot j}} = 1 - \text{conf}(j)$$

$\Rightarrow$ taux d'erreur a même inconvénient que confiance

Ne dit rien sur ce que la règle apporte de plus qu'une classification indépendante de toute condition !

Pour notre exemple :

	Règle 1	Règle 2	Règle 3	Règle 4	Nœud initial
taux d'erreur	0.20	0.57	0.11	0.36	0.56

Il faut comparer avec erreur au nœud initial.

En tant que résidu, l'indice d'implication inclut cette comparaison.

Indice d'implication en généralisation

En pratique, on considère le taux d'erreur en généralisation (sur échantillon test) ou en validation croisée.

On peut de même calculer les indices et intensités d'implication en généralisation.

Alternativement, on peut songer dans l'esprit BIC et MDL, à calculer sur les données d'apprentissage un

indice d'implication pénalisé pour la complexité de la règle

Indice pénalisé pour la complexité

complexité = longueur k de la règle (branche de l'arbre)

$$\text{Imp}_{pen}(j) = res_d(j) + \ln(n)k$$

Règle	res_d	k	Imp_{pen}
1	-6.826	2	4.39
2	0.788	2	12.01
3	-4.456	2	6.76
4	-4.847	2	6.37
Homme ⇒ marié	-7.119	1	-1.51
Femme ⇒ célib.	-7.271	1	-1.66

De ce point de vue, seules les règles définies au premier niveau de l'arbre seraient ici pertinentes.

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
déviance	-7.1	-6.8	0.8	-4.5	-4.8
Freeman-Tukey	-8.3	-6.3	0.1	-6.2	-2.4
ajusté	-4.3	-10.0	0.1	-7.7	-3.9

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

		Indices			Intensité		
Résidu		marié	célib.	div./v	marié	célib	div./v
Standardisé	res_s	1.6	0.1	**-1.3**	0.043	0.419	**0.891**
Déviance	res_d	3.9	0.8	**-3.4**	0.000	0.099	**0.999**
Freeman-Tukey	res_{FT}	1.5	0.1	**-1.4**	0.054	0.398	**0.895**
Ajusté	res_a	2.4	0.1	**-2.0**	0.005	0.379	**0.968**
Imp_{pen}		15.1	12.0	**7.9**			

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

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.
- L'interprétation de l'indice d'implication comme un résidu, suggère des variantes issues de la modélisation de table de contingence.
- La conclusion SI-optimale fait apparaître que la classe modale n'est pas toujours la meilleure solution .

Perspectives

- Mettre en œuvre sur des jeux de données réelles.
- Expérimenter, en particulier l'indice d'implication pénalisé,

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 règle 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.