ISMIS, Bari, September 27-29, 2006

Implication Strength of Classification Rules

Gilbert Ritschard
University of Geneva, Switzerland

Djamel A. Zighed
ERIC, University of Lyon 2, France

Outline

http://mephisto.unige.ch

1 Introduction

- Implicative Statistics (IS)
  - Tool for data analysis (Gras, 1979)
  - Interestingness measure for association rules mining (Suzuki and Kodratoff, 1998; Gras et al., 2001)
- IS useful for supervised classification?
  - YES, when the aim is characterizing typical profiles of outcomes

    Example 1: Physician interested in knowing typical profile of persons at risk for cancer, rather in predicting “cancer” or “not cancer”

    Example 2: Tax-collector interested in identifying groups where he has more chances to found fakers, rather in predicting “fraud” or “no fraud”
  - typical profile paradigm rather than classification paradigm

- Applying IS to decision rules
- We focus on classification rules derived from decision trees.
  - Index of implication for classification rules
    - Gras's index as a standardized residual
    - Alternative forms of residuals from modeling of contingency tables
  - Individual validation of classification rules
  - Optimal conclusion (alternative to the majority rule)

2 Trees and implication indexes

2.1 Trees and rules

- Illustrative data set and example of induced tree
- Classification rules and counter-examples (notations)

Illustrative data set (273 cases)

Civil status	Sex	Activity sector	Number of cases
married	male	primary	50
married	male	secondary	40
married	male	tertiary	6
married	female	primary	0
married	female	secondary	14
married	female	tertiary	10
single	male	primary	5
single	male	secondary	5
single	male	tertiary	12
single	female	primary	50
single	female	secondary	30
single	female	tertiary	18
divorced/widowed	male	primary	5
divorced/widowed	male	secondary	8
divorced/widowed	male	tertiary	10
divorced/widowed	female	primary	6
divorced/widowed	female	secondary	2
divorced/widowed	female	tertiary	2

Induced tree for civil status (married, single, divorced/widow)

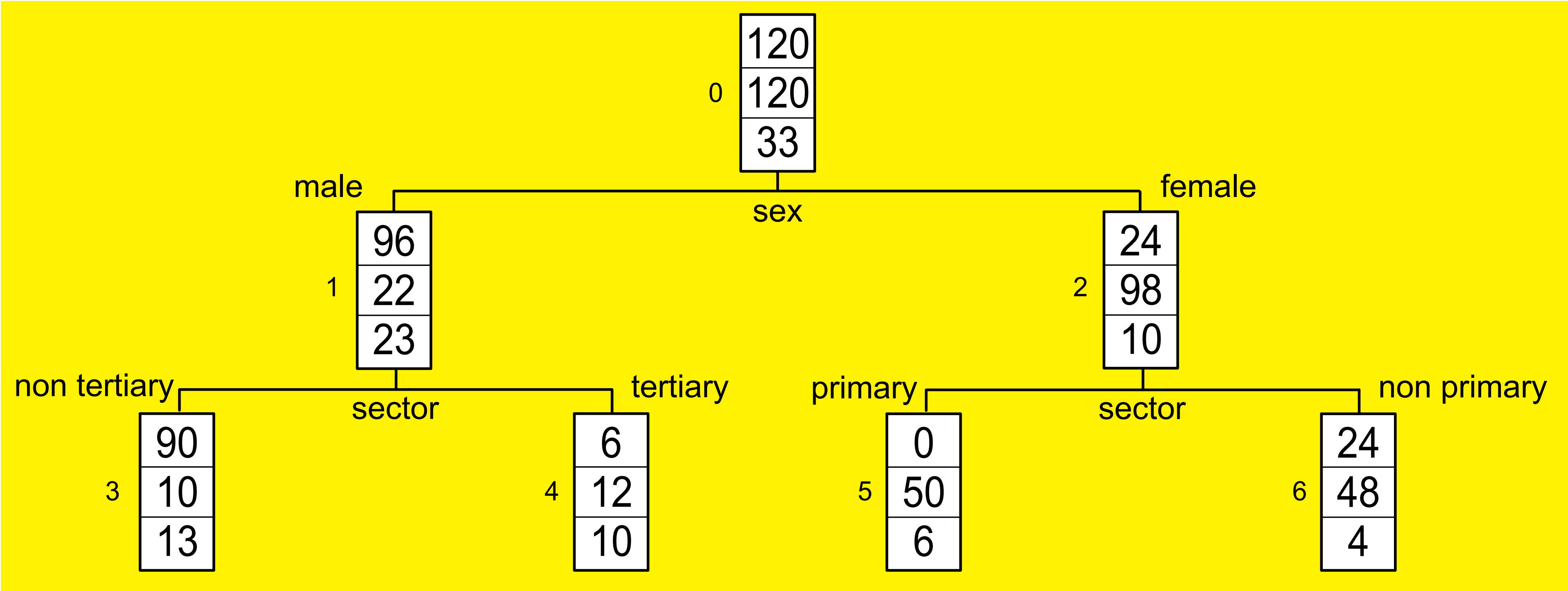


Table associated to the induced tree

Etat civil	Man: primary or secondary	Man: tertiary	Woman: primary	Woman: secondary or tertiary	total
Married	**90**	6	0	24	120
Single	10	**12**	**50**	**48**	120
Divorced/widow	13	10	6	4	33
Total	113	28	56	76	273

Rules (majority class):

R1.	Man of primary or secondary sector	$\Rightarrow$	married
R2.	Man of tertiary sector	$\Rightarrow$	single
R3.	Woman of primary sector	$\Rightarrow$	single
R4.	Woman of secondary or tertiary sector	$\Rightarrow$	single

Counter-examples

Gras's Implication Index defined from counter-examples.

Counter-example: verifies premise, but not conclusion (classification error)

Notations:

b	conclusion of rule j (changes with j)
$n_{b\cdot}$	total number of cases verifying b, $n_{\bar{b}\cdot} = n - n_{b\cdot}$ (changes with j)
n_{bj}	number of cases with premise j which verify conclusion b
$n_{\bar{b}j}$	number of counter-examples for rule j
H_0	Hypothesis that distribution among b and $\bar{b}$ is independent of the condition (same as marginal distribution)

Number of counter-examples under H_0:

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

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

Observed counts $n_{\bar{b}j}$ and n_{bj} of counter-examples et examples

predicted class *cpred*	Man: primary or secondary	Man: tertiary	Woman: primary	Woman: secondary or tertiary	total
0 (counter-example)	23	16	6	28	73
1 (example)	**90**	**12**	**50**	**48**	200
Total	113	28	56	76	273

Expected counts $n^e_{\bar{b}j}$ and n^e_{bj} of counter-examples et examples (Indep)

predicted class *cpred*	Man: primary or secondary	Man: tertiary	Woman: primary	Woman: secondary or tertiary	total
0 (counter-example)	63.33	15.69	31.38	42.59	153
1 (example)	49.67	12.31	24.62	33.41	120
Total	113	28	56	76	273

2.2 Implication Index and residuals

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

Contribution to Chi-square measuring distance between observed and expected

predicted class *cpred*	Man: primary or secondary	Man: tertiary	Woman: primary	Woman: secondary or tertiary
0 (counter-example)	**-5.068**	**0.078**	**-4.531**	**-2.236**
1 (example)	5.722	-0.088	5.116	2.525

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

Alternative residuals (used in statistical modeling of contingency tables)

standardized (=Imp(j))	res_s	contribution to Pearson Chi-square
deviance	res_d	contribution to Likelihood-ratio Chi-square (Bishop et al., 1975, p 136)
adjusted (Haberman)	res_a	res_s divided by its standard error (Agresti, 1990, p 224)
Freeman-Tukey	res_{TF}	variance stabilization (Bishop et al., 1975, p 137)

Residual		Rule 1	Rule 2	Rule 3	Rule 4
standardized (=Imp(j))	res_s	-5.068	0.078	-4.531	-2.236
deviance	res_d	-6.826	0.788	-4.456	-4.847
Freeman-Tukey	res_{FT}	-6.253	0.138	-6.154	-2.414
adjusted	res_a	-9.985	0.124	-7.666	-3.970

n^e_{bj} is mere an estimation $\Rightarrow$ variance of Imp$(j) < 1$

and Imp(j) tends to under-estimate the implication strength.

Other residuals are closer to $N(0, 1)$.

Degree of significance of the implication index

p-value of implication index $= p(N_{\bar{b}j} \leq n_{\bar{b}j} | H_0)$.

Prob. to get, by chances under H_0, less counter-examples than observed

Assuming fixed $n_{b\cdot}$ and $n_{\cdot j}$, can be computed

- with Poisson when n is small
- normal approximation when n is large (≥ 5)

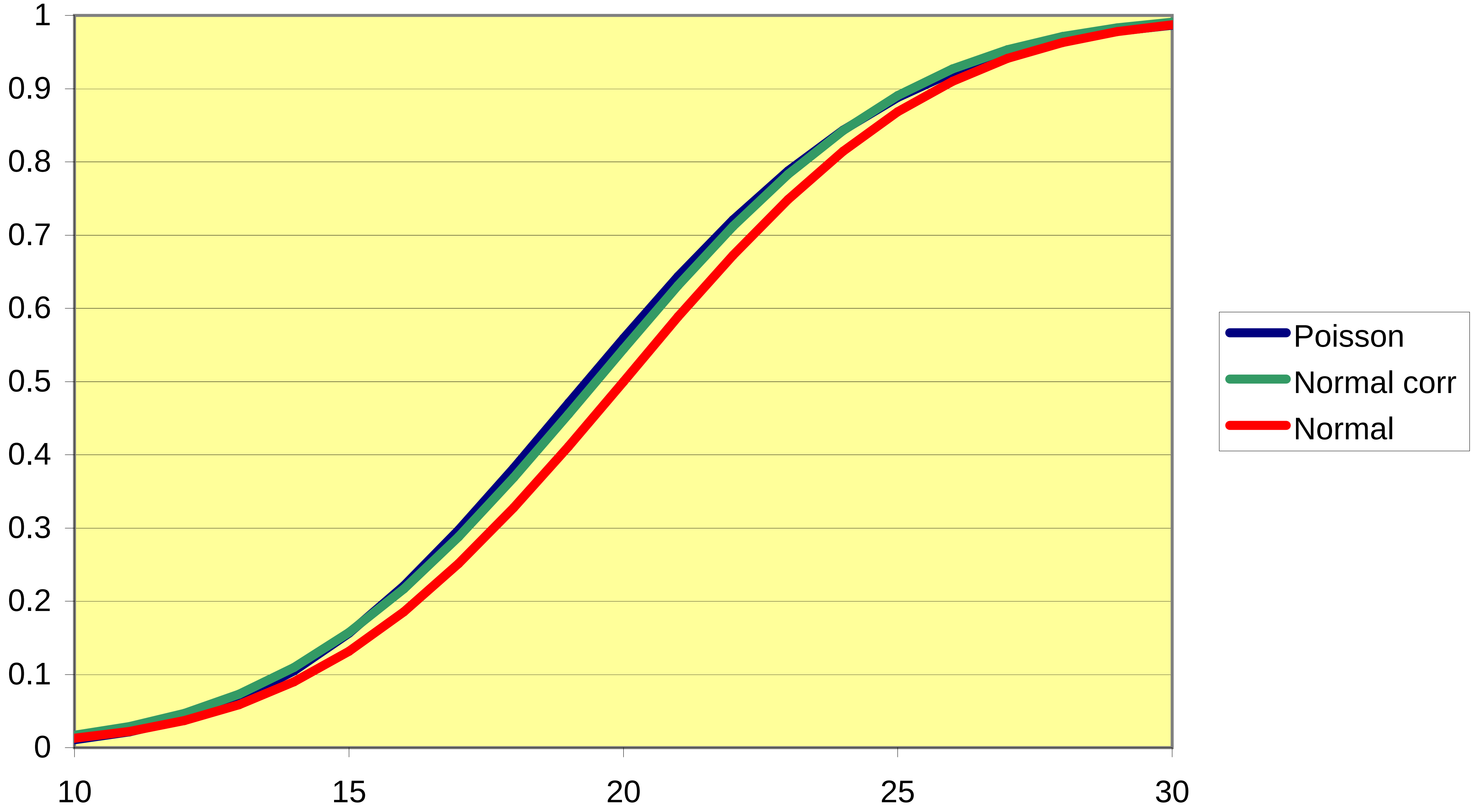
For normal approximation:

continuity correction

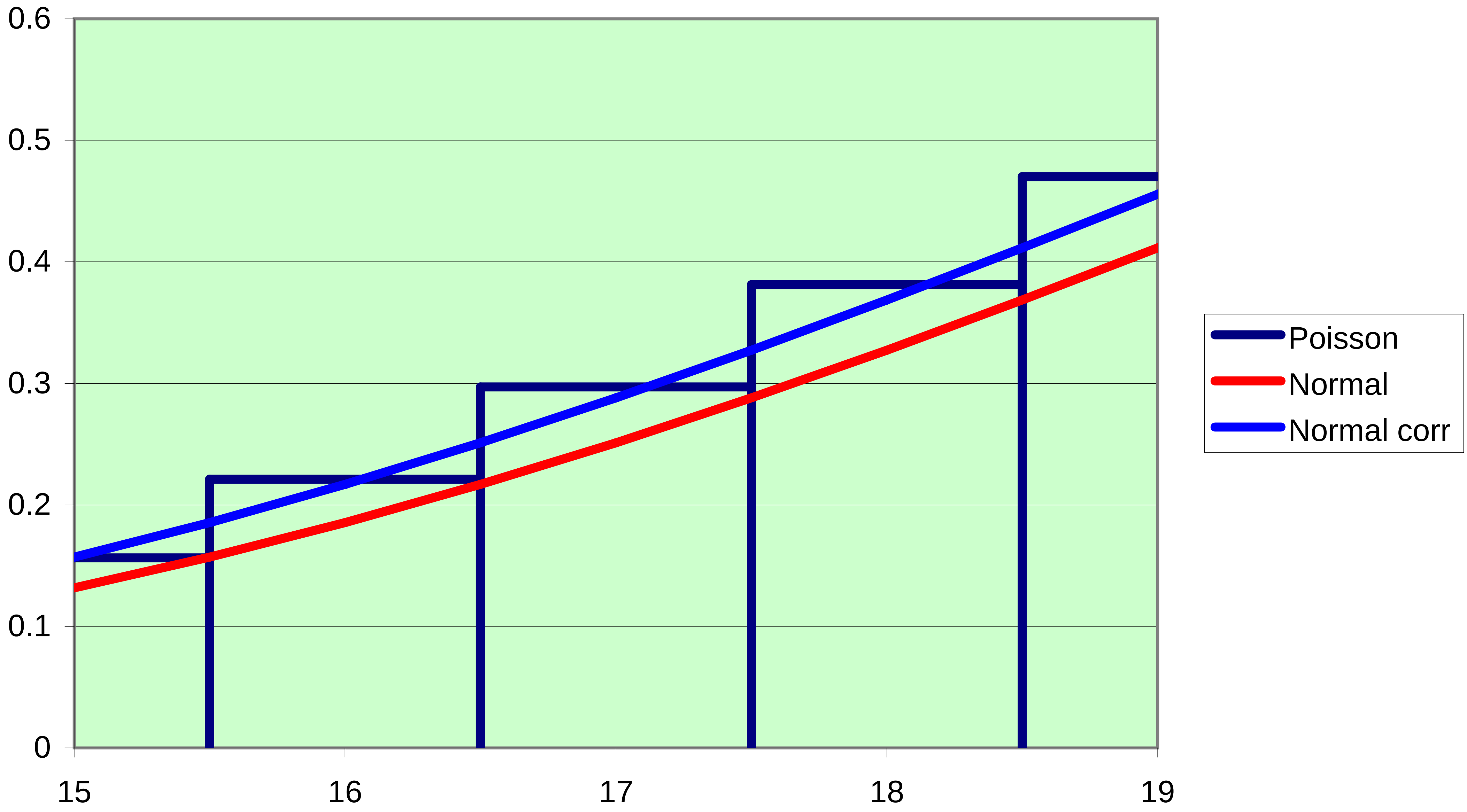
(add 0.5 to observed counts)

Difference may be as large as 2.6 points of percentage when $n_{\bar{b}j} = 100$.

Poisson, normal and normal with correction distributions



Details of Poisson, normal and normal with correction distributions



Implication intensity

The smaller the p-value, the greater the intensity

$\Rightarrow$ Intensity of implication $=$ complement to 1 of p-value

Prob. to get, by chances under H_0, more counter-examples than observed

Gras et al. (2004) define it in terms the normal approximation,
without continuity correction

We use

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

Variants of implication intensities (with continuity correction)

Residual		Rule 1	Rule 2	Rule 3	Rule 4
standardized	res_s	1.000	0.419	1.000	0.985
deviance	res_d	1.000	0.099	1.000	1.000
Freeman-Tukey	res_{FT}	1.000	0.350	1.000	0.988
adjusted	res_a	1.000	0.373	1.000	1.000

Intensity $< 0.5 \Leftrightarrow$ more counter-examples than expected under H_0.

$\Rightarrow$ Rule 2 irrelevant, since it makes worse than chance for predicting "single".

3 Individual rule relevance

In classification and especially with trees, the performance of the classifier is usually evaluated globally for the whole set of rules, by means for instance of the overall classification error in generalization.

The *implication intensity* and its variants may be used for validating the individual relevance of the rules.

In our example

- R1, R3 et R4 are clearly relevant
- R2 is not

What shall we do with non relevant rules? (Remember that the set of rule conditions must define a partition of the data set)

Error rate and implication index

number of errors = number of counter-examples

Error rate for rule j:

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

$\Rightarrow$ error rate has same drawbacks as confidence

Does not tell us if the rule makes better than chance (independent from any condition)!

For our example:

	Rule 1	Rule 2	Rule 3	Rule 4	Root node
error rate	0.20	0.57	0.11	0.36	0.56

Should be compared with error (.56) at root node.

Residuals, and hence implication indexes, account for this comparison.

Implication index in generalization

Practically, the error rate is computed on generalization
(on validation data) or through cross-validation.

Implication indexes can likewise be computed in generalization or by means of cross-validation.

Alternatively, in the spirit of BIC or MDL criteria, we could think to

implication indexes penalized for the rule complexity

computed on the learning data.

Penalized implication index

complexity = length k_j of rule j (branch of the tree)

$$\mathrm{Imp}_{pen}(j) \quad = \quad res_d(j) + \sqrt{k_j \ln(n_j)}$$

Rule	res_d	$\ln(n_j)$	k_j	Imp_{pen}
R1	-6.826	4.727	2	-3.75
R2	0.788	3.332	2	3.37
R3	-4.456	4.025	2	-1.62
R4	-4.847	4.331	2	-1.90
Man $\Rightarrow$ married	-7.119	4.949	1	-4.89
Woman $\Rightarrow$ single	-7.271	4.883	1	-5.06

Confirms that Rule 2 is irrelevant ($\mathrm{Imp}_{pen} = 0$ for root node).

Rule of 1st level look more relevant than those of level 2.

What to do with irrelevant rules?

1. Merge with an other rule.
2. Change the conclusion of the rule.

Merging rules

To respect tree structure, merge with sister rule.

In example, merge irrelevant rule R2 with sister rule R1

Residual	Rule 1+2	Rule 1	Rule 2	Rule 3	Rule 4
standardized	-3.8	-5.1	0.1	-4.5	-2.2
deviance	-7.1	-6.8	0.8	-4.5	-4.8
Freeman-Tukey	-8.3	-6.3	0.1	-6.2	-2.4
adjusted	-4.3	-10.0	0.1	-7.7	-3.9

4 Selecting the conclusion in each leaf

IS-optimal conclusion:

class with which we get the maximal implication strength.

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

Example: selecting conclusion for rule R2

Residual		Indexes			Intensities		
		married	single	div./w	married	single	div./w
standardized	res_s	1.6	0.1	**-1.3**	0.043	0.419	**0.891**
deviance	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**
adjusted	res_a	2.4	0.1	**-2.0**	0.005	0.379	**0.968**

Conclusion "divoced/widow" is more typical than "single" (modal class) for rule R2.

R2 becomes relevant with this conclusion.

5 Application: Students Enroled at the ESS Faculty in 1998

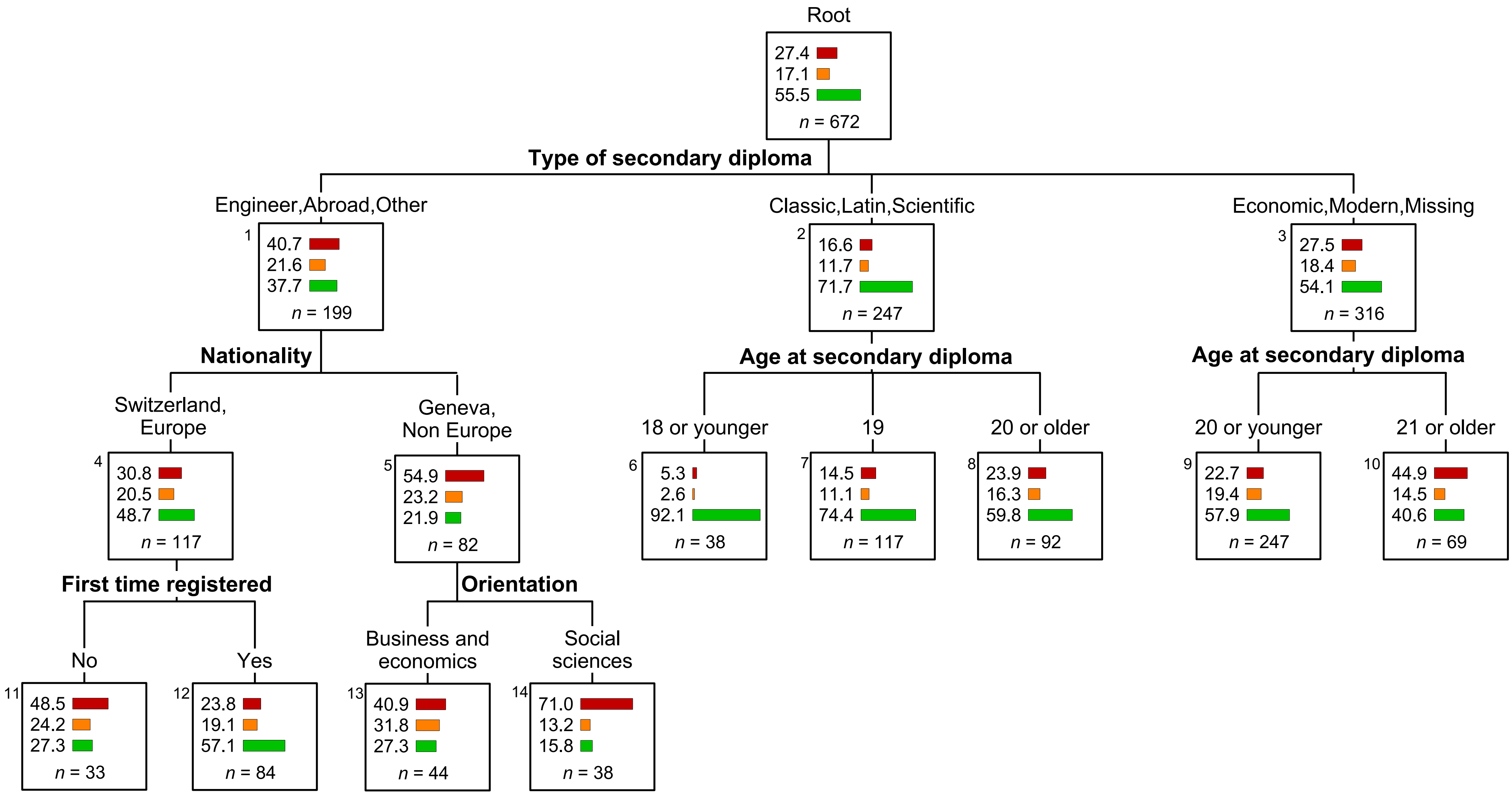
Response variable:

- Situation in October 1999 (eliminated, repeating 1st year, passed)

Predictors:

- Age
- Registration Date
- Selected Orientation (Business and Economics, Social Sciences)
- Type of Secondary Diploma Obtained
- Place of Obtention of Secondary Diploma
- Age at Obtention of Secondary Diploma
- Nationality
- Mother's Living Place

What is typical profile of those who repeat 1st year? Of those who are eliminated?



State assigned by the various criteria

Leaf	6	7	8	9	10	11	12	13	14
Majority class	3	3	3	3	1	1	3	1	1
Standardized residual	3	3	3	3	1	1	3	$\boxed{2}$	1
Freeman-Tukey residual	3	3	3	3	1	1	$\boxed{2}$	$\boxed{2}$	1
Deviance residual	3	3	3	$\boxed{2}$	1	1	$\boxed{2}$	$\boxed{2}$	1
Adjusted residual	3	3	3	$\boxed{2}$	1	1	$\boxed{2}$	$\boxed{2}$	1

Without correction for continuity, only one conclusion changes.

And we get no changes when the counts are multiplied by 1.4!

6 Conclusion

- Implication statistics applicable to and useful for classification trees.
- Index and intensity of implication usefully complement classical tree quality measures.
- They give valuable indications on the individual relevance of the rules.
- Interpreting the implication index as residuals suggests best suited variants borrowed from contingency table modeling.
- IS-optimal conclusion shows that the modal class is not necessarily the best from the typical profile paradigm standpoint.

Future researches

- Growing trees using IS criteria (typical profile paradigm).
- Further theoretical and experimental investigation of the penalized index.

THANK YOU

References

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.