Implication Strength of Classification Rules

Gilbert Ritschard[1] and Djamel A. Zighed[2]

[1] Dept of Econometrics, University of Geneva, CH-1211 Geneva 4, Switzerland
gilbert.ritschard@themes.unige.ch
[2] Laboratoire ERIC, University of Lyon 2, C.P.11 F-69676 Bron Cedex, France
abdelkader.zighed@univ-lyon2.fr

Abstract. This paper highlights the interest of implicative statistics for classification trees. We start by showing how Gras' implication index may be defined for the rules derived from an induced decision tree. Then, we show that residuals used in the modeling of contingency tables provide interesting alternatives to Gras' index. We then consider two main usages of these indexes. The first is purely descriptive and concerns the a posteriori individual evaluation of the classification rules. The second usage, considered for instance by Zighed and Rakotomalala [15], relies upon the intensity of implication to define the conclusion in each leaf of the induced tree.

1 Introduction

Implicative statistics has been introduced by the French mathematician Gras [6, 8, 9] as a tool for data analysis and has, more recently, been exploited for deriving valuable interestingness measures for association rules of the form "If A is observed, then we are very likely to observe B too" [3, 5, 10, 14]. The basic idea behind implicative statistics is that the fewer counter-examples a statistically observed relationship admits, the more implicative it is. It states also that a rule is irrelevant when the observed number of counter-examples exceeds the number expected in case of independence between the premise and the conclusion. Though, as we will show, this concept of strength of implication is applicable in a straightforward manner to classification rules for instance, only little attention has been paid to this appealing idea in the framework of supervised learning.

The aim of this paper is to discuss the scope and limits of implicative statistics for supervised classification and especially for classification trees. Section 2 shows how Gras' implication indexes can be applied to classification rules derived from an induced decision tree. It proposes alternatives to Gras' index inspired from residuals used in the modeling of multiway contingency tables. Section 3 discusses the use of implication strength for the individual validation of each classification rule, while Section 4 shows that the implication strength provides a useful alternative to the majority rule for selecting the most relevant conclusion in a leaf. Section 5 proposes concluding remarks and perspectives of development.

2 Classification trees and implication indexes

For our discussion, we consider a fictional example where we are interested in predicting the civil status (married, single, divorced/widowed) of individuals from their sex (male, female) and sector of activity (primary, secondary, tertiary). The civil status is the outcome or response variable, while sex and activity sector are the predictors. The data set is composed of the 273 cases described by Table 1.

Table 1. The illustrative data set

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

2.1 Trees and rules

Classification trees induce classification rules from data in two steps. First, the tree is grown by seeking, through recursive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class. Each split is done according to the values of one predictor. The process is greedy. It starts by trying all predictors to find the "best" split of the whole learning data set. Then, the process is repeated at each new node until some stopping rule is reached. In a second step, once the tree is grown, classification rules are derived by choosing the most relevant value, usually the majority class, in each leaf (terminal node) of the tree.

Figure 1 shows the tree induced with the CHAID method [11], a 5% significance level and a minimal node size fixed to 20. The same tree is obtained with CART [4] and a minimal .02 gain value. The tree partitions the predictor space

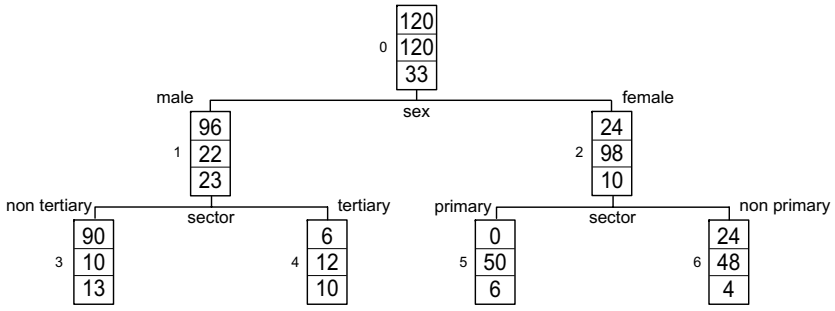

Fig. 1. Example: Induced tree for civil status (married, single, divorced/widowed)

into groups such that the distribution of the outcome variable, the civil status, differs as much as possible from one group to the other. For our discussion, it is convenient to represent the four resulting distributions into a table that cross classifies the outcome variable with the set of profiles (the premises of the rules) defined by the branches. Table 2 is thus associated to the tree of Figure 1.

Classification rules are usually derived from the tree by assigning the majority class of the leaf to the branch that leads to it. For example, a man working in the secondary sector belongs to leaf 3 and will be classified as married, while a man of the tertiary sector (leaf 4) will be classified as single. In Table 2, the column headings define the premises of the rules, the conclusion being given, for each column, by the row containing the greatest count. The tree defines thus the four following rules:

R1: Man of primary or secondary sector ⇒ married
R2: Man of tertiary sector ⇒ single
R3: Woman of primary sector ⇒ single
R2: Woman of secondary or tertiary sector ⇒ single

Classification rules as compared with association rules have the following characteristics: i) Their conclusions can only be values (classes) of the outcome variable, and ii) the premises of the rules are mutually exclusive and define a partition of the predictor space. Anyway, they are rules and we can then apply

Table 2. Table associated to the induced tree

	Man		Woman		
Civil Status	primary or secondary	tertiary	primary	secondary or tertiary	Total
Married	**90**	6	0	24	120
Single	10	**12**	**50**	**48**	120
Div./Widowed	13	10	6	4	33
Total	113	28	56	76	273

to them concepts such as support, confidence and, which is here our concern, implication indexes.

2.2 Counter-examples and implication index

The index of implication [see for instance 7, p. 19] of a rule is defined from the number of counter-examples, i.e. of cases that verify the premise but not the conclusion. In our case, it is in each leaf (column of Table 2) the number of cases that are not in the majority class. Letting b denote the conclusion (row of the table) of rule j and n_{bj} the maximum in the jth column, the number of counter-examples is $n_{\bar{b}j} = n_{.j} - n_{bj}$. The index of implication is a standardized form of the deviation between this number and the number of counter-examples expected when assuming that the distribution of the outcome values is independent of the premise.

Formally, the independence hypothesis H_0 states that the number $N_{\bar{b}j}$ of counter-examples of rule j results from a random draw of $n_{.j}$ cases. Under H_0, letting $n_{b.}/n$ be the marginal proportion of cases in the conclusion class b of rule j, $N_{\bar{b}j}$ follows a binomial distribution $\mathrm{Bin}(n_{.j}, n_{b.}/n)$, or, when $n_{.j}$ is not fixed a priori, a Poisson distribution with parameter $n^e_{\bar{b}j} = n_{\bar{b}.} n_{.j}/n$ [12]. In the latter case, the parameter $n^e_{\bar{b}j}$ is both the mathematical expectation $\mathrm{E}(N_{\bar{b}j} \mid H_0)$ and the variance $\mathrm{var}(N_{\bar{b}j} \mid H_0)$ of the number of counter-examples under H_0. It is the number of cases in leaf j that would be counter-examples if they were distributed among the outcome classes with the marginal distribution, i.e. that of the root node (right margin in Table 2).

Gras' *implication index* is the difference $n_{\bar{b}j} - n^e_{\bar{b}j}$ between the observed and expected numbers of counter-examples, standardized by the standard deviation, i.e., if we retain the Poisson model,

$$\mathrm{Imp}(j) = \frac{n_{\bar{b}j} - n^e_{\bar{b}j}}{\sqrt{n^e_{\bar{b}j}}} \ , \tag{1}$$

which can also be expressed in terms of the number of cases verifying the rules as $\mathrm{Imp}(j) = -(n_{bj} - n^e_{bj})/\sqrt{n_{.j} - n^e_{bj}}$.

Let us explicit the calculation of the index for our example. We consider for that the variable "predicted class", denoted *cpred*, that takes value 1 for each case (example) belonging to the majority class of its leaf and 0 otherwise (counter-example). By cross-classifying this variable with the premises of the rules, we get Table 3 where the first row gives for each rule its number $n_{\bar{b}j}$ of counter-examples and the second row its number n_{bj} of examples.

Likewise, Table 4 gives the expected numbers $n^e_{\bar{b}j}$ and n^e_{bj} of counter-examples and examples obtained by distributing the $n_{j.}$ covered cases with the marginal distribution. Note that these counts cannot be computed from the margins of Table 3. They are obtained by first dispatching the column total using the marginal distribution of Table 2 and aggregating then separately each resulting column

Table 3. Observed numbers $n_{\bar{b}j}$ and n_{bj} of counter-examples and examples

	Man		Woman		
Predicted class *cpred*	primary or secondary	tertiary	primary	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

Table 4. Expected numbers $n^e_{\bar{b}j}$ and n^e_{bj} of counter-examples and examples

	Man		Woman		
Predicted class *cpred*	primary or secondary	tertiary	primary	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

according to its corresponding observed majority class (not the expected one!). This explains why Tables 3 and 4 do not have the same right margin.

From these two tables, we get easily the implication indexes using formula (1). They are reported in the first row of Table 5. For the first rule, the index equals $\text{Imp}(1) = -5.068$. This negative value indicates that the number of observed counter-examples is less than the number expected under the independence hypothesis, which stresses the relevance of the rule. For rule 2, the implication index is positive, which tells us that the rule is irrelevant since it generates more counter-examples than classifying without taking account of the condition.

2.3 Implication index and residuals

In its formulation (1), the implication index looks like a standardized residual, namely as the (signed root square of) the contribution to the Pearson Chi-square [see for example 1, p 224]. Here, it is indeed the Chi-square that measures the divergence between Tables 3 and 4. These contributions are depicted in Table 5, those of the first row being the implication indexes.

This interpretation of Gras' implication index in terms of residuals (residuals for the fitting of the counts of counter-examples by the independence model) suggests that other forms of residuals used in the framework of the modeling of

Table 5. Contributions to the Chi-square measuring divergence between Tables 3 and 4

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

Table 6. The various residuals as alternative implication indexes

Residual		Rule R1	Rule R2	Rule R3	Rule R4
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

the counts in multiway contingency tables could also prove useful for measuring the strength of rules. These include:

The *deviance residual*, $res_d(j) = \text{sign}(n_{\bar{b}j} - n^e_{\bar{b}j})\sqrt{|2n_{\bar{b}j}\log(n_{\bar{b}j}/n^e_{\bar{b}j})|}$, which is the signed square root of the contribution (in absolute value) to the likelihood ratio Chi-square [2, pp 136-137]).

Haberman's adjusted residual, $res_a(j) = (n_{\bar{b}j} - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}(n_{b\cdot}/n)(1 - n_{\cdot j}/n)}$, which is the Pearson standardized residual divided by its standard error [1, p 224].

Freeman-Tukey's residual, $res_{FT}(j) = \sqrt{n_{\bar{b}j}} + \sqrt{1 + n_{\bar{b}j}} - \sqrt{4n^e_{\bar{b}j} + 1}$, which results from a variance-stabilizing transformation [2, p 137].

Table 6 exhibits the values of theses alternative implication indexes for each of our four rules. We observe that they are concordant as expected. The standardized residual is known to have a less than unity variance. This is because the counts $n_{b\cdot}$ and $n_{\cdot j}$ are sample dependent and hence themselves random. Thus $n^e_{\bar{b}j}$ is only an estimation of the Poisson parameter. Ignoring the randomness of the denominator in formula (1) leads to underestimate the strength. The deviance, adjusted and Freeman-Tukey's residuals are best suited for this situation and are known to have in practice a distribution closer to the standard normal $N(0, 1)$ than the simple standardized residual. We can check in our example that the standardized residuals, i.e. Gras' implication index tends to give lower absolute values than the three alternatives. The only exception is rule R3 which admits only 6 counter-examples.

2.4 Implication intensity and *p*-value

In order to evaluate the statistical significance of the computed implication strength, it is natural to look at the p-value, i.e. at the probability $p(N_{\bar{b}j} \leq n_{\bar{b}j} \mid H_0)$. When $n^e_{\bar{b}j}$ is small, this probability can be obtained, conditionally on $n_{b\cdot}$ and $n_{\cdot j}$, with the Poisson distribution $P(n^e_{\bar{b}j})$. For large $n^e_{\bar{b}j}$, the normal distribution gives a good approximation. A correction for the continuity may be necessary however, the difference mighting be for example as large as 2.6 points of percentage when $n^e_{\bar{b}j} = 100$. Letting $\phi(.)$ denote the standard normal distribution, we have $p(N_{\bar{b}j} \leq n_{\bar{b}j} \mid H_0) \simeq \phi\Big((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}}\Big)$.

Table 7. The implication intensity and its variants (with continuity correction)

Residual		Rule R1	Rule R2	Rule R3	Rule R4
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

The *implication intensity* is defined as the complementary to one of this p-value. Gras [see for instance 7] defines it in terms of the normal approximation, without the correction for continuity however. We compute it as

$$\text{Intens}(j) = 1 - \phi\Big((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}}\Big) \ . \tag{2}$$

Anyway, this intensity can be interpreted as the probability of getting, under the independence hypothesis H_0, a higher number of counter-examples than the count observed for rule j. Table 7 gives these intensities for our four rules. It shows also the complementary to 1 of the p-values of the deviance, adjusted and Freeman-Tukey's residuals computed with the continuity correction, i.e. by adding 0.5 to the observed counts of counter-examples.

3 Individual rule relevance

The implication intensity and its variants are naturally useful for validating each classification rule individually. This knowledge enriches the usual global validation of the classifier. For example, among the four rules issued from our illustrative tree, rules R1, R3 and R4 are clearly relevant, while R2, with an implication intensity below 50% should clearly be rejected.

The question is then what shall we do with the cases covered by the conditions of irrelevant rules. Two solutions can be envisaged: i) Merging cases covered by an irrelevant rule with another rule, or ii) changing the conclusion. The possible choice of a more suitable conclusion is discussed in Section 4. As for the merging of rules, if we want to respect the tree structure we have indeed to merge cases of a leaf with those of a sister leaf, which is equivalent to pruning the corresponding branch. In our example, this leads to the merge of rules R1 and R2 into a new rule "Man $\Rightarrow$ married". Residuals for the number of counter-examples of this new rules are respectively $res_s = -3.8$, $res_d = -7.1$, $res_{FT} = -4.3$ and $res_a = -8.3$. Except for the deviance residual, they exhibit a slight deterioration as compared to the implicative strength of rule R1.

It is interesting here to compare the implicative quality with the usual error rate used for validating classification rules. The number of counter-examples considered is precisely the number of errors produced by the rule on the learning set. The error rate is thus the percentage of counter-examples among the cases covered by the rule, i.e. $\text{err}(j) = n_{\bar{b}j}/n_{.j}$, which is also equal to $1 - n_{bj}/n_{.j}$, the complementary to one of the confidence. The error rate suffers that from the

Table 8. Implication index penalized for the rule complexity

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

same drawbacks as the confidence. For instance, it does not tell us how better the rule does than a classification independent of any condition. Furthermore, the error rate is linked with the choice of the majority class as conclusion. For our example, the error rate is respectively for our four rules 0.2, 0.57, 0.11 and 0.36. The second rule is thus also the worst from this point of view. Comparing with the error rate at the root node, which is 0.56, shows that this rate of 0.57 is very bad. Thus, for being really informative about the relevance of the rule, the error rate should be compared with the error rate of some naive baseline rule. This is exactly what the implication index does. Resorting to implication indexes, we get in addition probabilities which permits to distinguish between statistically significant and non significant relevance.

Practically, in order to detect over-fitting, error rates are computed on validation data sets or through cross validation. Indeed, the same can be done for the implication quality by computing the implication indexes and intensities in generalization.

Alternatively, we could consider, in the spirit of the BIC (Bayesian information criteria) or MDL (Minimum message length) principle, to penalize the implication index by the complexity of the condition. Since the lower the implication index of a rule j, the best it is, the index should be penalized by the length k_j of the branch that defines the condition of rule j. The general idea behind such penalization is that the simpler the condition, the lower the risk to assign a bad distribution to a case. As a first proposal we suggest the following penalized form inspired from the BIC [13] and based on the deviance residual

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

For our example, the values of the penalized index are given in Table 8.

This penalized form of the index confirms the ranking of the initial rules, which here have all the same length $k_j = 2$. In addition, it is useful for validating the merge of the two rules R1 and R2. Table 8 highlights the superiority of the merged rule "Man $\Rightarrow$ married" over both rules R1 and R2. It gives a clear signal in favor of the merge.

At the root node, both the residual and the number of conditions are zero. Hence, the penalized implication index is zero too. Thus, a positive penalized implication index suggests that we can hardly expect that the rule would do bet-

Table 9. Implication indexes and intensities of rule 2 for each possible conclusion

		Indexes			Intensity		
Residual		married	single	div./wid.	married	single	div./wid.
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**

ter in generalization than a random classification independent of any condition. For our example, this confirms once again the badness of rule R2.

4 Implication strength versus majority rule

A final aspect regarding the implication strength is its link with the choice of the conclusion for each rule. The idea, that has for instance been exploited in [15, pp 282-287], is to chose in each leaf the conclusion for which the rule gets its highest implication intensity. This is indeed an alternative to the majority rule.

To illustrate, we give in Table 9 the values of the alternative indexes and intensities of implication for each of the three possible conclusion that may be assigned to rule 2. The conclusion labeled "single" corresponds to the majority class. We see here that there is a better conclusion from the strength of implication standpoint, namely "divorced or widowed". All four indexes designate this conclusion as the best with an implication intensity that goes from 89.1% for Gras' index to 99.9% for the deviance residual. Again we can notice that Gras' index seems to slightly under-estimate the implication intensity.

5 Conclusion

The aim of the paper was to demonstrate the usefulness of the concept of implication strength for classification rules derived from induced decision trees. We have shown that it may prove helpful for evaluating the individual relevance of rules as well as an alternative to the majority class rule for selecting the rule conclusion. We have also stressed the interest of considering implication indexes penalized for the complexity.

Much remains to be done. The variants to Gras' index, the continuity corrections, the penalized index all would merit further theoretical as well as empirical study of their impact in real word problems. In Section 4, we have stressed the interest of the implication strength approach for selecting the most appropriate conclusion for each leaf. We should even be able to go further and use criteria based on implication indexes in the tree growing process. An other important issue that we plan to investigate is the relationship between the individual performance of the rule and the global classifier efficiency.

References

[1] Agresti, A. *Categorical Data Analysis*. Wiley, New York, 1990.

[2] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. *Discrete Multivariate Analysis*. MIT Press, Cambridge MA, 1975.

[3] Blanchard, J., Guillet, F., Gras, R., Briand, H. Using information-theoretic measures to assess association rule interestingness. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pages 66–73. IEEE Computer Society, 2005. ISBN 0-7695-2278-5.

[4] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. *Classification And Regression Trees*. Chapman and Hall, New York, 1984.

[5] Briand, H., Fleury, L., Gras, R., Masson, Y., Philippe, J. A statistical measure of rules strength for machine learning. In *Proceedings of the Second World Conference on the Fundamentals of Artificial Intelligence (WOCFAI 1995)*, pages 51–62, Paris, 1995. Angkor.

[6] Gras, R. *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, 1979.

[7] Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., Peter, P. Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI*, E-1:3–30, 2004.

[8] Gras, R., Larher, A. L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématique, Informatique et Sciences Humaines*, (120):5–31, 1992.

[9] Gras, R., Ratsima-Rajohn, H. L'implication statistique, une nouvelle méthode d'analyse de données. *RAIRO Recherche Opérationnelle*, 30(3): 217–232, 1996.

[10] Guillaume, S., Guillet, F., Philippé, J. Improving the discovery of association rules with intensity of implication. In Zytkow, J.M., Quafafou, M., editors, *Proceedings of the Eurpoean Conference on Principles of Data Mining and Knowledge Discovery (PKDD 1998)*, volume 1510 of *Lecture Notes in Computer Science*, pages 318–327. Springer, 1998. ISBN 3-540-65068-7.

[11] Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127, 1980.

[12] Lerman, I.C., Gras, R., Rostam, H. Elaboration d'un indice d'implication pour données binaires I. *Mathématiques et sciences humaines*, (74):5–35, 1981.

[13] Raftery, A.E. Bayesian model selection in social research. In Marsden, P., editor, *Sociological Methodology*, pages 111–163. The American Sociological Association, Washington, DC, 1995.

[14] Suzuki, E., Kodratoff, Y. Discovery of surprising exception rules based on intensity of implication. In Zytkow, J.M., Quafafou, M., editors, *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, September 23-26, Proceedings*, pages 10–18. Springer, Berlin, 1998.

[15] Zighed, D.A., Rakotomalala, R. *Graphes d'induction: apprentissage et data mining*. Hermes Science Publications, Paris, 2000.