Inducing and Evaluating Classification Trees with Statistical Implicative Criteria

Gilbert Ritschard[1], Vincent Pisetta[2], and Djamel A. Zighed[2]

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

Abstract. Implicative statistics criteria have proven to be valuable interestingness measures for association rules. Here we highlight their interest for classification trees. We start by showing how Gras' implication index may be defined for rules derived from an induced decision tree. This index is especially helpful when the aim is not classification itself, but characterizing the most typical conditions of a given conclusion. We show that the index looks like a standardized residual and propose as alternatives other forms of residuals borrowed from the modeling of contingency tables. 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 relies upon the strength of implication for assigning the most appropriate conclusion to each leaf of the induced tree. We demonstrate the practical usefulness of this statistical implicative view on decision trees through a full scale real world application.
Classification tree, Implication strength, Class assignment, Rule relevance, Typical profile, Targeting

1 Introduction

Implicative statistics was introduced by the French mathematician Régis Gras [**? ? ?**] as a tool for data analysis and has, since the late 90's, 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" [**? ? ? ?**]. The basic idea behind implicative statistics is that a statistically observed relationship is of interest only if the number of counter-examples is less than expected by chance, and that the larger the difference, the more implicative it is.

We see two major motivations for this concept of statistical implication. On the one hand, logic implication, does not admit any counter-example. Hence, it is too strong and leaves no place for dealing with the random content of statistical relationships. On the other hand, the classical confidence, which measures the chances of matching the conclusion when the condition is satisfied, is not able to tell us whether or not the conclusion is more probable than it would in case of independence from the condition. For instance, assume that the conclusion B

is true for 95% of all the cases. Then, a rule with a confidence of 90% would do worse than simple chance, i.e. than deciding that B is true for all cases without taking care of the condition A. But why looking at counter-examples and not just at positive examples? Indeed, this is formally equivalent (see Section **??**), and hence is just a matter of taste. Looking for the rarity of counter-examples makes the reasoning closer to what is done with logic rules, i.e. invalidating the rule when there are (too many) negative examples.

Though, as we will show, this concept of strength of implication is applicable in a straightforward manner to classification rules, only a little attention has been paid to this appealing idea in the framework of supervised learning. The aim of this article is to discuss the scope and limits of implicative statistics for supervised classification and especially for classification trees. One difference between classification rules and association rules is that the consequent of the former has to be chosen from an a priori set list of classes (the possible states of the response variable), while the consequent for the latter can concern any event not involved in the premise, since there is no a priori outcome variable. A second difference is that unlike the premises of association rules, those of a set of classification rules define a partition of the data set, meaning that there is one and only one rule applicable to each case. These aspects, however, do not intervene in anyway in the definition of the implication index which just requires a premise and a consequent. Hence, implication indexes are technically applicable without restrictions to classification rules. There remains, nevertheless, the question of whether they make sense in the supervised learning setting.

The implication index measures how typical the condition of the rule is for the conclusion, i.e. how much more characteristic than pure chance it is for the selected conclusion. Indeed, we are only interested in conditions under which the probability to match the conclusion is higher than the marginal proportion corresponding to pure chance. A condition with a probability lower than the marginal proportion would characterize atypical situations for the conclusion, i.e. situations in which the proportion of cases matching the conclusion is less than in the whole data set. It would thus be characteristic of the negation of the conclusion, not the conclusion itself. Looking at typical conditions for the negation of the conclusion could be useful too. Nevertheless, it does not require any special attention since it can simply be handled by looking at the implication strength of the rule in which we would have replaced the conclusion by its negation.

The information on the gain of performance over chance provided by the implication index usefully complements the knowledge provided for instance by the classical raw misclassification rate. However, we may go a step further and, by considering a so called targeting or condition typicality paradigm instead of the classification paradigm, resort to implication indexes for selecting the conclusion of a rule. Moreover, we could even imagine methods for growing trees that would optimize the implication strength of the resulting rules. Such a targeting paradigm will be adopted, for instance, by a physician who is more interested in knowing the typical profile of persons who develop a cancer than in predicting

for each patient whether or not he has a cancer. Likewise, a tax-collector may be more interested in characterizing groups in which he has increased chances to find fakers than in predicting for each taxpayer whether or not he commits fraud. The most frequent class, commonly called the 'majority class' in the decision tree literature, is obviously the best choice for minimizing classification errors. However, we will see that for the targeting paradigm, the highest quality conclusion, i.e. that for which the rule has the highest implication strength, is not necessarily this majority class.

The paper is organized as follows. Section **??** shows how Gras' implication index 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 **??** discusses the use of implication strength for the individual validation of each classification rule. In Section **??** we adopt the aforementioned typical profile paradigm and consider using the implication indexes for selecting the most relevant conclusion in a leaf of a classification tree. We also briefly describe different approaches for growing trees from that typical profile standpoint. Section **??** reports experimental results that highlight the behavior of the implication strength indexes and illustrates their potential on a real world application from social sciences. Finally, we present concluding remarks in Section **??**.

We start our presentation by adopting a classical classification standpoint.

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 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 or decision or dependent) variable, while sex and activity sector are the predictors (or condition or independent variables). The data set is composed of the 273 cases described by Table **??**.

2.1 Trees and Rules

Classification rules can be induced from data using classification trees 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 criterion becomes true. In a second step, once the tree is grown, classification rules are derived by choosing the most relevant value, usually the majority class (the most frequent), in each leaf (terminal node) of the tree.

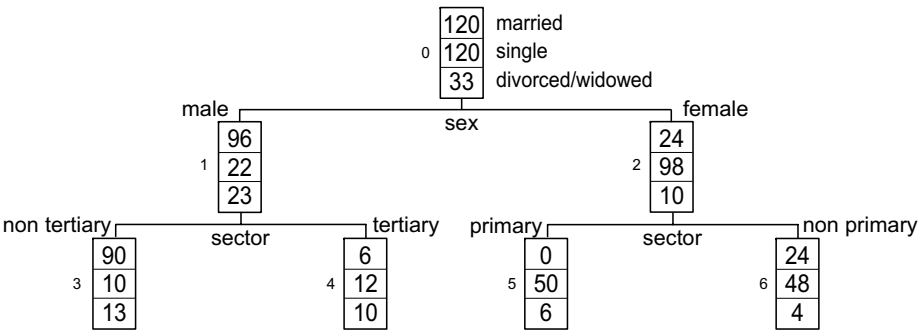


Fig. 1. Example: Induced tree for civil status (married, single, divorced/widowed)

Figure **??** shows the tree induced with the CHAID method [**?**], using a 5% significance level and a minimal node size fixed at 20. The same tree is obtained with CART [**?**] using a minimal .02 gain value. The three numbers in each node represent the counts of individuals who are respectively 'married', 'single', and 'divorced or widowed'. The tree partitions the predictor space 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 **??** is thus associated to the tree of Figure **??**.

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

As mentioned, 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 **??**, the column headings define the premises of the rules, the conclusion being given, for each column, by the row containing the greatest count. Using this approach, the four following rules are derived from the tree shown in Figure **??**:

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

In contrast to association rules, classification rules have the following characteristics: i) The conclusions of the rules 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. Nonetheless, they are rules and we can then apply 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 **?** , p19] of a rule is defined from the number of *counter-examples*, i.e. of cases that match the premise but not the conclusion. In our case, for each leaf (represented by a column in Table **??**), the count of counter-examples is 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 $n_{\bar{b}j} = n_{\cdot j} - n_{bj}$. 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 $n_{\cdot j}$ cases. Under H_0, letting $n_{b\cdot}/n$ be the marginal proportion of cases in the conclusion class b of rule j and setting $n_{\bar{b}\cdot} = n - n_{b\cdot}$, $N_{\bar{b}j}$ follows a binomial distribution

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

$\text{Bin}(n_{\cdot j}, n_{\bar{b}\cdot}/n)$, or, when $n_{\cdot j}$ is not fixed a priori, a Poisson distribution with parameter $n^e_{\bar{b}j} = n_{\bar{b}\cdot} n_{\cdot j}/n$ [**?**]. In the latter case, the parameter $n^e_{\bar{b}j}$ is both the mathematical expectation $\text{E}(N_{\bar{b}j} \mid H_0)$ and the variance $\text{var}(N_{\bar{b}j} \mid H_0)$ 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 according to the marginal distribution, i.e. that of the root node (right margin in Table **??**).

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,

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

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

Let us make the calculation of the index explicit for our example. We define for that the variable "predicted class", denoted *cpred*, which 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 **??** where the first row gives the number $n_{\bar{b}j}$ of counter-examples for each rule and the second row the number n_{bj} of examples.

Likewise, Table **??** gives the expected numbers $n^e_{\bar{b}j}$ and n^e_{bj} of negative examples (counter-examples) and positive examples obtained by distributing the $n_{j\cdot}$ covered cases according to the marginal distribution. Note that these counts cannot be computed from the margins of Table **??**. They are obtained by first dispatching the column total using the marginal distribution of Table **??** and

Table 5. Contributions to the Chi-square measuring divergence between Tables **??** and **??**

	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

then separately aggregating each resulting column according to its corresponding observed majority class (not the expected one!). This explains why Tables **??** and **??** do not have the same right margin.

From these two tables, we can easily get the implication indexes using formula (**??**). They are reported in the first row of Table **??**. 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 the second rule, the implication index is positive, which tells us that the rule is less powerful than pure chance since it generates more counter-examples than would classifying without taking account of the condition.

2.3 Implication Index and Residuals

In its formulation (**??**), the implication index looks like a standardized residual, namely as the (signed square root of) the contribution to the Pearson Chi-square [see for example **?** , p 224]. The implication index is indeed related to the Chi-square that measures the divergence between Tables **??** and **??**. The contributions of each cell to this Chi-square are depicted in Table **??**, 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 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 square root of the contribution (in absolute value) to the likelihood ratio Chi-square [**?** , pp 136-137].
- *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 [**?** , p 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 [**?** , p 224].

Table 6. The various residuals as alternative implication indexes

Residual		Rule R1	Rule R2	Rule R3	Rule R4
Standardized (Gras' index)	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

There are thus different ways of measuring the departure from the expected number of counter-examples. It is always instructive to cross-compare values produced by such alternatives. When they are concordant, as they should be, comparison reinforces the reliability of the outcome. Divergences, on the other hand, flag situations for which we should be more cautious before drawing any conclusion from the numerical value of a given index. Section **??** provides some highlights on the specific behavior of each of the four alternatives considered here.

Table **??** exhibits the values of these alternative implication indexes for each of the four rules derived from the tree in Figure **??**. We observe that they are concordant as expected. The standardized residual is known to have a variance that may be lower than one. 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 (**??**) leads to underestimating the strength. The deviance, adjusted and Freeman-Tukey's residuals are better 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 see in our example that the standardized residual, i.e. Gras' implication index, tends to give lower absolute values than the three alternatives. The only exception is rule R3, for which the deviance residual provides a slightly smaller value than Gras' index. Note that R3 admits only six 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, because the difference might be for example as large as 2.6 percent when $n^e_{\bar{b}j} = 100$. Letting $\phi(.)$ denote the standard normal distribution, we have $p(N_{\bar{b}j} \leq n_{\bar{b}j} \mid H_0) \simeq \phi\Big((n_{\bar{b}j} + 0.5 - n^e_{\bar{b}j})/\sqrt{n^e_{\bar{b}j}}\Big)$.

The *implication intensity* can be defined as the complement of such a p-value. Gras [see for instance **?**] defines it in terms of the normal approximation, but

without the correction for continuity. We compute it as

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

In either case, 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 **??** gives these intensities for our four rules. It shows also the complement 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. Notice that provided probabilities below 50% correspond to positive values of the indexes, i.e. bad ones, and those above 50% to negative ones. This is a direct consequence of taking the probabilities from the normal distribution, which is symmetric.

3 Individual Rule Relevance

The implication intensity and its variants are 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 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 **??**. We exclude indeed further splitting of the node, since we assume that a stopping criterion has been matched. 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 sibling leaf, which is equivalent to pruning the corresponding branch. In our example, this leads to merging rules R1 and R2 into a new rule "Man ⇒ 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 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.

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

Residual		Rule R1	Rule R2	Rule R3	Rule R4
Standardized (Gras)	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

Table 8. Implication index penalized for the rule complexity

Rule	res_d	k	Imp_{pen}
R1	-6.826	2	-3.75
R2	0.788	2	3.37
R3	-4.456	2	-1.62
R4	-4.847	2	-1.90
Man ⇒ married	-7.119	1	-4.89
Woman ⇒ single	-7.271	1	-5.06

The error rate is thus the percentage of counter-examples among the cases covered by the rule, i.e. $\text{err}(j) = n_{\bar{b}j}/n_{\cdot j}$, which is also equal to $1 - n_{bj}/n_{\cdot j}$, the complement to one of the confidence. The error rate suffers that from the same drawbacks as the confidence. For instance, it does not tell us how better the rule does than a classification done independently 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 better 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 [**?**] 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 **??**.

These penalized values confirm the ranking of the initial rules, which here all have the same length $k_j = 2$. In addition, the penalized index is useful for validating results of merging the two rules R1 and R2. Table **??** highlights the superiority of the merged rule "Man ⇒ married" over both rules R1 and R2. It gives a clear signal in favor of merging.

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 better in generalization than assigning randomly the cases according to the root node distribution, i.e. independently of any condition. For our example, this confirms once again the badness of rule R2.

4 Adopting a Typical Profile Paradigm

To this point, we have assumed that the conclusion of the rule was simply the majority class. This is justified when the pursued aim is classification. However, as already mentioned in the introduction, there are situations where the typical profile paradigm is better suited. Remember the example of the physician primarily interested in the characteristics of those patients who develop a cancer, and that of the tax-collector who wants to know the groups of tax payers who are at most risk of committing fraud. Social sciences, where the concern is most often to understand phenomena rather than to predict values or classes, is also a distinctive domain to which the typical profile paradigm suits well. For example, sociologists of the family may be interested in determining the profiles in terms of education, professional career, parenthood, etc. that increase chance of divorce, and in Section **??**, we present an application where the goal is to characterize the profiles of those students who are at most risk of repeating their first year. In such situations, the majority class rule is no longer the best choice. Indeed, from this typical profile standpoint, it is more natural to search for rules with the highest possible implication strength than to minimize the misclassification rate.

Having this optimal implication strength goal in mind, we successively discuss the assignment of the most relevant conclusion to the premises defined by a given grown tree, and the use of implication strength criteria in the tree growing process.

4.1 Maximal Implication Strength versus Majority Rule

For a given grown tree, maximizing the implication strength is simply achieved by assigning to each leaf the conclusion for which the rule gets its highest implication intensity. Though **?**], pp 282-287 have already considered this way of proceeding, they do not provide a sound justification for the approach. Note also that the method has not, to the best of our knowledge, been implemented so far in any tree growing software.

To illustrate the principle, we give in Table **??** the values of the alternative indexes and intensities of implication for each of the three possible conclusions that may be assigned to rule R2 of our example. The conclusion labeled “single” corresponds to the majority class. However, considering the strength of implication, the best conclusion is “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. Indeed, to be a man working in the tertiary sector is not typical of single people since the rule would in that case generate more counter-examples than expected by chance. Concluding to "divorced or widowed" is better in that respect since the number of positive examples is in that case larger than expected by chance. Again we can notice that Gras' index seems to slightly under-estimate the implication intensity.

An important point is that unlike the majority rule, seeking the maximal implication strength favors the variability of conclusions among rules, meaning that we have more chances to create at least one rule for each value of the outcome variable. In our example, using the majority class we do not create any rule that concludes with divorced/widowed, while with the implication strength at least one rule concludes with each of the three outcome states. Indeed, we need at least as many different profiles as outcome classes if we want at least one rule concluding with each outcome state, i.e. we should have $r \leq q$ with r the number of outcome classes and q the number of rules.

By definition, if we assign the same conclusion to all rules, any negative departure from the expected number of counter-examples of a rule should be compensated for a positive departure for an other rule. Likewise, for a given rule, any negative departure from the expected number of counter examples for one of the possible conclusions should be compensated for a positive one for an other conclusion. Formally we have

$$n_{\bar{\imath}j} < n^e_{\bar{\imath}j} \;\Rightarrow \begin{cases} \text{there exists } k \neq i \text{ such that } n_{\bar{k}j} > n^e_{\bar{k}j} \text{ and} \\ \text{there exists } h \neq j \text{ such that } n_{\bar{\imath}h} > n^e_{\bar{\imath}h} \end{cases}$$

$$n_{\bar{\imath}j} > n^e_{\bar{\imath}j} \;\Rightarrow \begin{cases} \text{there exists } k \neq i \text{ such that } n_{\bar{k}j} < n^e_{\bar{k}j} \text{ and} \\ \text{there exists } h \neq j \text{ such that } n_{\bar{\imath}h} < n^e_{\bar{\imath}h} \end{cases}$$

As a consequence, all the rules cannot attain their maximal implication strength for the same conclusion, which favors indeed the diversity of the conclusions among rules. A second consequence is that at each leaf we may assign a conclusion such that the rule gets a non positive implicative index or, equivalently, an implication intensity greater or equal to 50%.

Table 9. Implication indexes and intensities of rule R2 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**

4.2 Growing Trees with Implication Strength Criteria

Let us now look at the tree growing procedure and assume that the rule conclusions are selected so as to maximize the implication strength of the rules. The question is whether there is a way to split a node so as to maximize the strength of the resulting rules. The difficulty here is that a split results indeed in more than one rule. Hence, we face a multicriteria problem, namely the maximization over sets of implication strengths.

To get simple solutions, one can think to transform the multidimensional optimization problem into a one dimensional one by focusing on some aggregated criterion. The following are three possibilities:

- A weighted average of the concerned optimal implication indexes, taking weights proportional to the number of concerned cases.
- The maximum over the strengths of the rules belonging to the set.
- The minimum over the strengths of the rules belonging to the set.

The first criterion is of interest when the goal is to achieve good strengths on average. The second one should be adopted when we look for a few rules with high implication strengths without bothering too much for the other ones, and the latter is of interest when we want the highest possible implication strength for the poorest rule.

We have not yet experimented tree growing with these criteria. It is worthwhile however to say that, from the typical profile paradigm standpoint methods such as CHAID that attempt to maximize association seem preferable to those based on entropies. Indeed, maximizing the strength of association between the resulting nodes and the outcome variable leads to distributions that departure as much as possible from that in the parent node, and hence from that of the root node corresponding to independence. We may thus expect the most significant departures from independence and hence rules with strong implication strength. Methods based on entropy measures, on the other hand, favor departures from the uniform, or equiprobable, distribution and are therefore more in line with the classification standpoint.

5 Experimental Results

We present here a series of experimental results that provide additional insights into the behavior and scope of the original implication index and the three variants we introduced. First, we study the behavior of the indexes. We then present an application, which also serves as a basis for experimental investigations regarding the effect of the continuity correction and the consequences of using maximal implication strength rules instead of the majority rule on classification accuracy, recall and precision.

5.1 Compared Behavior of the 4 Indexes

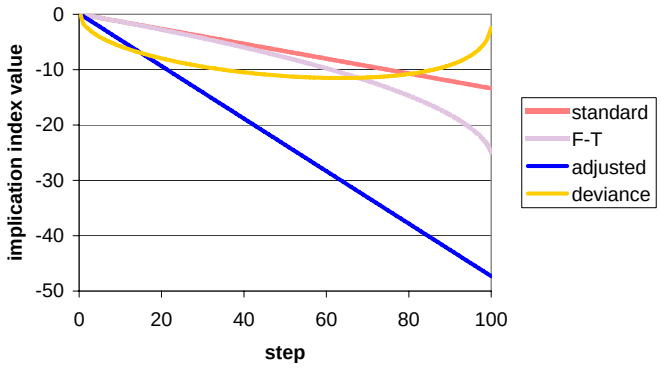
In order to gain better understanding on how the different implication indexes behave, we ran a simulation to see how they evolve when the number of counter-examples is progressively decreased from the expected number under independence to 0. At independence we expect a null implication strength, while when no counter-examples are observed we should have high implication strength.

The simulation design is as follows. We consider a dataset of size 1000 and a rule defined from a leaf containing 200 cases (20%). We suppose that a proportion p of the 1000 cases belongs to the outcome class selected as conclusion for the rule. Starting with a proportion $f = f_0$ of cases of the leaf that fall in the conclusion class, we progressively increase f in 100 constant steps until the maximum $f = 100\%$ is reached. The initial starting point corresponds to independence and the final point to a pure distribution with no counter-examples. At each step we compute, applying the continuity correction, the value of each of the 4 indexes, namely the standardized, Freeman-Tukey, adjusted and deviance residuals.

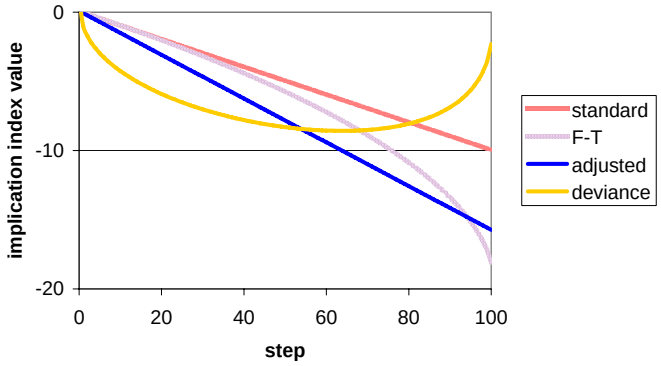
Figure **??** shows the results for $p = 10\%$, 50% and 90%. Notice the difference of scale between the three plots: The implication strengths are higher when the class of interest is infrequent in the population, i.e. when p is small. We observe that the standardized and adjusted residuals evolve linearly between independence and purity, while the increase in Freeman-Tukey's residual tends to accelerate when we approach purity. The deviance residual evolves curiously in a parabolic way. It dominates the other indexes in the neighborhood of independence, it reaches a maximum (in absolute terms) and diminishes (in absolute terms) when we approach purity. This decreasing behavior when the number of counter-examples tends to 0 disqualifies the deviance residual as a good measure of the rule implication strength. The linear evolution of the standadized and adjusted residuals makes them our prefered measures, the latter having in addition the advantage of being the most reliably comparable with standard normal thresholds.

5.2 Application on a Student Administrative Dataset

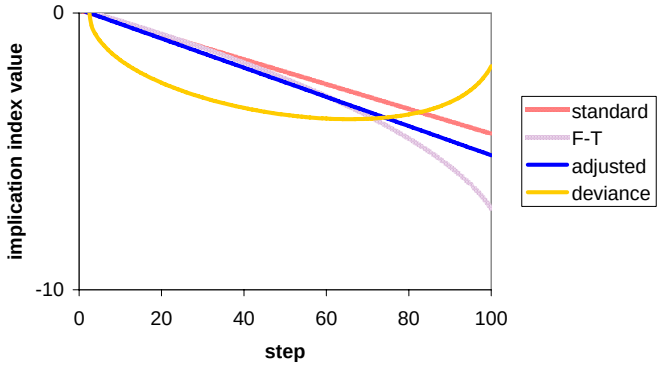
We consider administrative data about the 762 first year students who were enrolled in fall 1998 at the Faculty of Economic and Social Sciences (ESS) of the University of Geneva [**?**]. The goal is to learn rules for predicting the situation (1. eliminated, 2. repeating first year, 3. passed) of each student after the first year, or more precisely to discover the typical profile of those students who are either eliminated or have to repeat their first year. For the learning data, the response variable is thus the student situation in October 1999. The predictors retained are age, first time registered at University of Geneva, chosen orientation (Social Sciences or Business and Economics), type of secondary diploma achieved (classic, latin, scientific, economics, modern, other), place where secondary diploma was obtained (Geneva, Switzerland outside Geneva, Abroad), age when secondary diploma was obtained, nationality (Geneva, Swiss except



(a) 10% of cases in selected outcome class



(b) 50% of cases in selected outcome class



(c) 90% of cases in selected outcome class

Fig. 2. Behavior of the 4 indexes between independence (Step 0) and purity (Step 100). Values reported include the continuity correction.

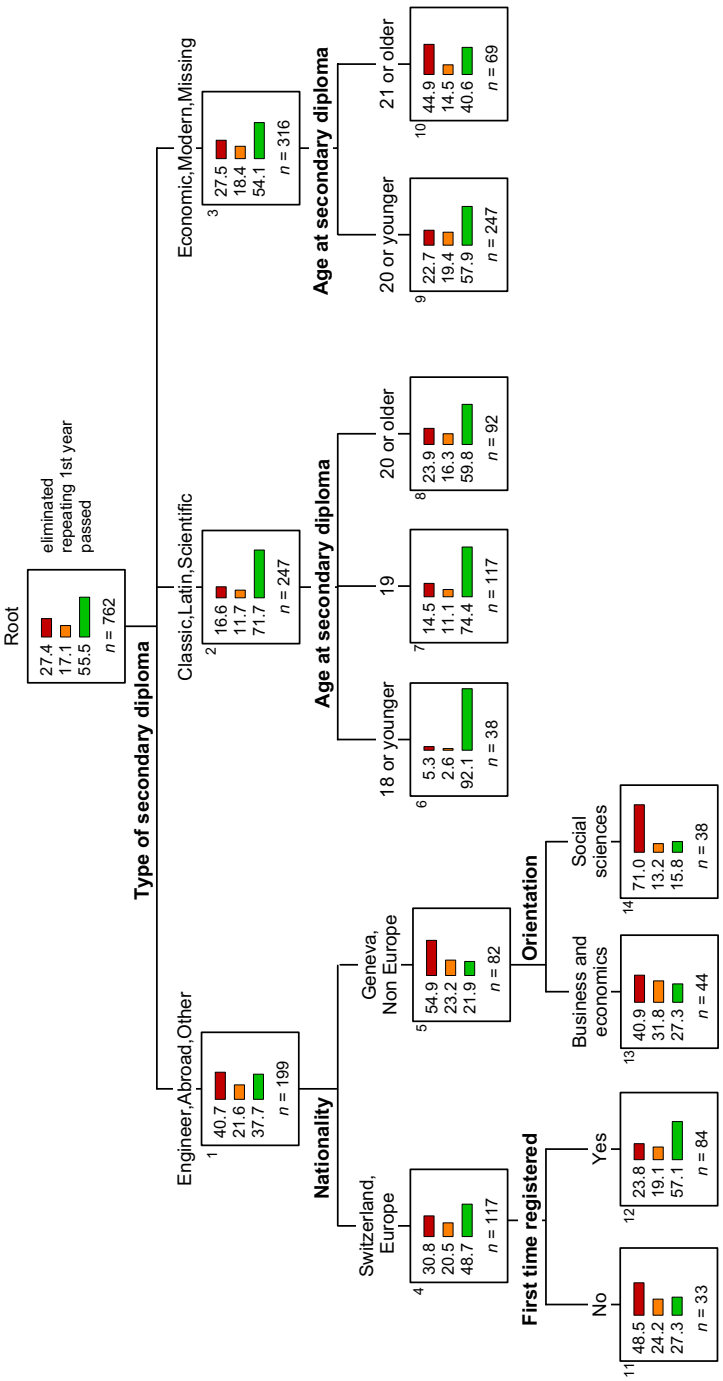


Fig. 3. CHAID induced tree for the ESS Student data. Outcome states are from top to down: eliminated, repeating 1st year, passed. Figures next to the bars are percentages.

Table 11. 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	**2**	1
Freeman-Tukey residual	3	3	3	3	1	1	**2**	**2**	1
Deviance residual	3	3	3	**2**	1	1	**2**	**2**	1
Adjusted residual	3	3	3	**2**	1	1	**2**	**2**	1

Geneva, Europe, Non Europe) and mother's living place (Geneva, Switzerland outside Geneva, Abroad).

Figure **??** shows the tree induced using CHAID with minimal node size set to 30, minimal parent node size to 50 and a maximal 5% significance for the Chi-square. Table **??** provides the details regarding the counts in the leafs. Here, our interest is not in the growing procedure, but rather in the state assigned to each leaf.

We used successively the majority class rule and each of the four variants of implication indexes for that. Table **??** reports the results. We can see that the 5 methods agree for 6 out of the 9 leaves. The conclusion assigned to leaves number 9, 12 and 13 vary, however, among the 5 methods. All four implication indexes assign state 2, "repeating the first year", to leaf 13 where the majority class is 1, "eliminated". This tells us that belonging to this leaf, i.e having a not typical Swiss college secondary diploma obtained either in Geneva or abroad and having chosen a business and economic orientation, is a typical profile of those who repeat their first year. And this holds, indeed, despite "repeating the first year" is not the majority class of the leaf.

The deviance and adjusted residuals agree about assigning also state 2, "repeating", to leaves number 9 and 12, and the Freeman-Tukey residual agrees also with this conclusion for leaf 12. These leaves also define characteristic profiles of those who repeat their first year, even though the majority class for these profiles is "passed".

5.3 Effect of Continuity Correction

We expect continuity correction, i.e. adding .5 to the observed counts $n_{\bar{b}j}$ of counter-examples, to have only very marginal effects and to be important only in conjunction with small minimal node sizes.

For our application on the ESS student data, the continuity correction changes the conclusion only when we use the Freeman-Tukey residual for leaf 12 (with 84 cases). The conclusions remain the same for all other leaves and for all leaves when we use any of the three other residuals. Furthermore, the effect of the continuity correction vanishes when we multiply all the counts by a factor greater or equal to 1.4, which confirms our expectation.

Nevertheless, we suggest systematically introducing the error correction when computing the indexes. There are two reasons for this: First, it does not change much the index values in case of large counts and produces values best suited for comparison with standard normal thresholds in case of small counts. Secondly, it avoids possible troubles (division by zero for instance) that may occur when some observed counts are zero.

5.4 Recall and Precision

In terms of the overall error rate, selecting the majority class is no doubt the better choice. However, if we are interested in the recall rate, i.e. in the pro-

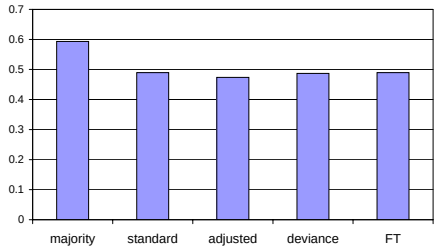


Fig. 4. Correct classification rate, 10-fold CV

portion of cases with a given output value c_k that are detected as having this value, we may expect the implication indexes to outperform the majority rule for infrequent classes. Indeed, highly infrequent outcome states have high chances to never be selected as conclusion by the majority rule. We may therefore expect low recall for them when we select the most frequent class as conclusion. Regarding precision, i.e. the proportion of cases classified as having a value c_k that effectively have this value, expectations are less clear since the relationship between the numerator and denominator seems not linked to the way of choosing the conclusion.

In order to verify these expectations on our ESS student data, we computed for the majority rule and each of the four variants of implication indexes, the 10-fold cross-validation (CV) values of the overall good classification rate, as well as of the recall and precision for each of the three outcome states. As can be shown on Figure **??** the loss in accuracy that results from using maximal implication rules lies between 12% for the adjusted residual and 10% for the standard residual.

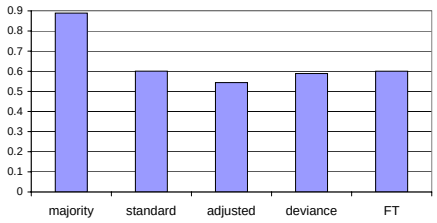
Figure **??** exhibits the CV recall rates obtained for each of the three states. They confirm our expectations: selecting the conclusion according to implication indexes deteriorates the recall for the majority class "passed", but results in an improvement in recall for the two other classes. The improvement is especially important for the last frequent state, i.e. "repeating", for which we get recall

rates ranging between 30% and 40% instead of almost 0% with the majority rule.

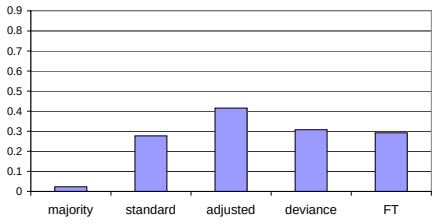
In Figure **??**, we observe an improvement in precision for "passed" (the majority class) and "repeating" (the last frequent class) and a slight deterioration for "eliminated". This illustrates that the choice of the conclusion has apparently no predictable effect on precision. Indeed, the only thing we may notice here is that improvement concerns the two classes with a proportion of cases that is further (on either side) from the equiprobable probability $1/c$, where c is the number of outcome classes.

6 Conclusion

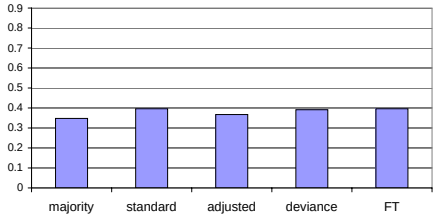
The aim of this article was to demonstrate the usefulness of the concept of implication strength for rules derived from induced decision trees. We have shown that Gras' implication index can be applied in a straightforward manner to classification rules and have proposed three alternatives inspired from residuals used in the statistical modeling of multiway contingency tables, namely the deviance, adjusted and Freeman-Tukey residuals. As for the scope of the implication indexes we have successively discussed their use for evaluating individual rules, for selecting the conclusion of the rule and as criteria for growing trees. We have stressed that implication indexes are a valuable complement to classical error rates as validation tools. They are especially interesting in a targeting framework where the aim is to determine the typical profile that leads to a conclusion



(a) passed

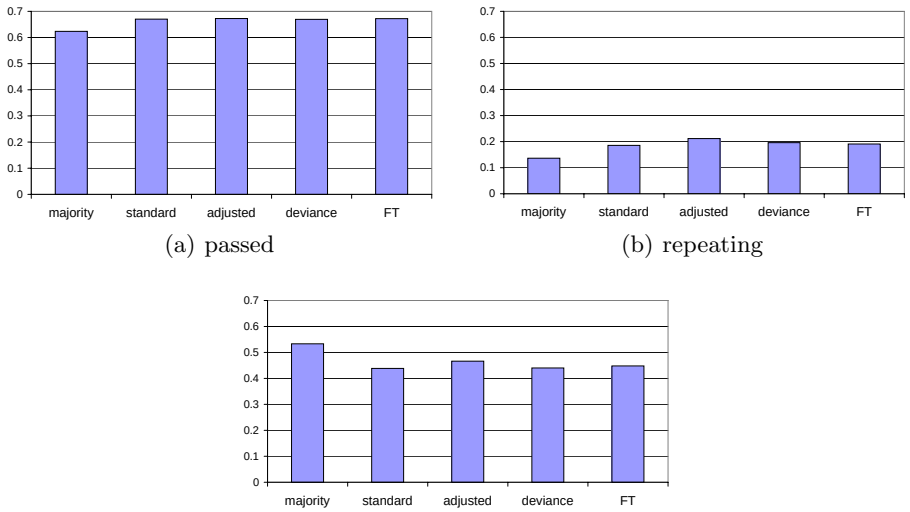


(b) repeating



(c) eliminated

Fig. 5. Recall, 10-fold CV



(a) passed (b) repeating

(c) eliminated

Fig. 6. Precision, 10-fold CV

rather than classifying individual cases. As criteria for selecting the conclusion, they may be a useful alternative to the majority rule in the case of imbalanced data. Their advantage is that in such imbalanced situation and unlike decisions based on the majority class, they favor conclusion diversity among rules as well as recall for poorly represented classes.

Four variants of implication indexes have been discussed. Which one should we use? The simulation study of their behavior has shown that the deviance residual curiously diminishes when the number of counter-examples tends to zero and should therefore be disregarded. The standard residual (Gras' index) and Haberman's adjusted residual both evolve linearly between independence and purity and thus seem to be the better choices. From the theoretical standpoint, if we want to compare the values with thresholds of the standard normal, Haberman's adjusted residual is preferable.

We have also introduced the implication intensity as the probability to get by chance more counter-examples than observed. This is indeed just a monotonic transformation of the corresponding implication index. Hence rankings based on the indexes or on the intensities will necessarily agree. Indexes seem better suited, however, to distinguishing between situations with high implication strengths. The intensities on the other hand, provide additional information about the statistical significance of the implication strength.

It is worth mentioning that, to our knowledge, implication indexes have not so far been implemented in tree growing software. Making them available is essential for popularizing them. We have begun working on implementing the

maximal implication selection process and tree growing algorithms based on implication criteria into Tanagra [**?**] a free open source data mining software, and plan also to make these tools available in Weka.

Beside this implementation task, there are some other issues that would merit further investigation. For instance, the penalized implication index we proposed in Section **??** is not completely satisfactory. In a n-arry tree the paths to the leaves are usually shorter than in a binary tree, even if they define the same leaves. Penalization based on the length of the path as we proposed, would therefore be different for a rule derived from a binary tree than for the same rule derived from a n-arry tree. The use of implication criteria in the tree growing process needs also a deeper reflection.

Despite all which remains to be done, our hope is that this article will contribute to enlarge both the scope of induced decision trees and that of implication statistics.