Goodness-of-Fit Measures for Induction Trees

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

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

Abstract. This paper is concerned with the goodness-of-fit of induced decision trees. Namely, we explore the possibility to measure the goodness-of-fit as it is classically done in statistical modeling. We show how Chi-square statistics and especially the Log-likelihood Ratio statistic that is abundantly used in the modeling of cross tables, can be adapted for induction trees. Not only is the Log-likelihood Ratio statistic suited for testing the goodness-of-fit. It allows also to test the significance of the fit between two nested trees. In addition, we derive from it pseudo R^2's. We propose also adapted forms of the Akaike (AIC) and Bayesian (BIC) information criteria that prove useful in selecting the best compromise model between fit and complexity.

Keywords. Induction tree, goodness-of-fit, Chi-square tests, models comparison.

1 Introduction

Decision tree induction has become since Breiman et al. [1] one of the most popular supervised learning tool. It consists in seeking, trough successive splits of the learning data set, some optimal partition of the predictor space for predicting the value of the response variable. Once the membership to a class of the partition is established, the prediction is given by the majority rule that assigns the most frequently observed value of the response variable in that class. Though the primary use of decision trees is classification, they provide a useful description of how the distribution of the response variable is conditioned by the values of the predictors. A tree may exhibit for instance how attributes like age, gender, education level and profession influence the probability of solvency of customers. In this respect, induced trees model the effect of the predictors upon the response variable in the same way as linear or logistic regression. In this paper, we focus on this descriptive modeling feature of induction trees. We introduce and discuss criteria for measuring the reliability of the description provided by a tree.

It is worth mentioning that the criteria we are interested in are not intended to measure the classification performance of the tree. The error rate that focuses on the fit of individual values is well suited for this purpose and has already been largely studied. For the reliability of the description, individual predictions do not matter. Rather, we focus on the posterior distribution of the response

variable, i.e. on the distribution conditioned by the values of the predictors. Our concern is thus to measure how well a tree may predict these conditional distributions. This is a goodness-of-fit issue very similar to that encountered in the statistical modeling of multiway cross tables. According to our knowledge, however, it has not been addressed so far for induced trees. Textbooks, like [2] for example, do not mention it, and, as far as this model assessment issue is concerned, statistical learning focuses almost exclusively on the statistical properties of the classification error rate (see for example [3], chap. 7).

In statistical modeling, e.g. linear regression, logistic regression or more generally generalized linear models (GLM), the goodness-of-fit is usually assessed by two kinds of measures. On the one hand, indicators like the coefficient of determination R^2 or pseudo R^2's tell us how better the model does than some naive baseline model. On the other hand we measure, usually with divergence Chi-square statistics, how well the model reproduces some target or, in other words, how far we are from the target.

Our contribution is a trick that permits to use this statistical machinery with induced trees. The trick allows us to propose, among others, an adapted form of the Likelihood Ratio deviance statistic with which we can test statistically the significance of any expansion of a tree. Other criteria discussed are R^2 like measures and the powerful model selection AIC and BIC criteria.

The paper is organized as follows. Section 2 enlightens different kinds of fit that make sense in supervised learning. In Section 3 we formalize our fit issue in terms of table comparison and describe the trick for induced trees. Section 4 proposes and discusses goodness-of-fit measures. Finally, the concluding Section 5 presents some development perspectives.

2 Descriptive Ability versus Classification Performance

The goodness-of-fit of a statistical model refers to its capacity to reproduce the data. In a predictive framework, it is measured by a decreasing function of the prediction error, i.e. of the discrepancy between the observed values y_α and the predicted states $\hat{y}_\alpha = f(\mathbf{x}_\alpha)$ of the response variable, with $\alpha = 1, \ldots, n$, n being the number of cases. In classification, the response variable y is categorical with say r values and we have to distinguish between two kinds of predictions. For given values $\mathbf{x} = (x_1, \ldots, x_p)$ of the predictors, we may be interested in predicting the class or in predicting the probability to be in a given class. This suggests to distinguish between the two following models

- the descriptive classification model $\hat{\mathbf{p}}(\mathbf{x})$
- the classifier itself $f(\mathbf{x}) = g\big(\hat{\mathbf{p}}(\mathbf{x})\big)$

with $\hat{\mathbf{p}}(\mathbf{x}) = \big(\hat{p}_1(\mathbf{x}), \ldots, \hat{p}_r(\mathbf{x})\big)$ being the prediction of the probability distribution $\mathbf{p}(\mathbf{x}) = \big(p(Y = y_1|\mathbf{x}), \ldots, p(Y = y_r|\mathbf{x})\big)$ of the response variable y given $\mathbf{x}$, and $g(\cdot)$ denoting the majority vote rule $g\big(\hat{\mathbf{p}}(\mathbf{x})\big) = \arg\max_i \hat{p}_i(\mathbf{x})$.

Many supervised classification tools, among which logistic regression, but also the induction trees, follow a two steps process: First predict the class probabilities

by fitting a descriptive model $\mathbf{p}(\mathbf{x})$, then classify according to the majority rule $g(\hat{\mathbf{p}}(\mathbf{x}))$. It makes then sense to assess the fit of both the descriptive and the classification models.

If the final goal is to predict the class, the classification error rate based on the discrepancy between the y_α's and $\hat{y}_\alpha$'s provides undoubtedly the relevant information on how well the classifier fits the data.

Nevertheless, users may be interested in the description provided by the descriptive model rather than in classification itself. Indeed, logistic regressions as well as induction trees or graphs typically provide useful insights on how predictors jointly affect the probabilities to be in given classes. It is then this representation that goodness-of-fit measures should assess. Hence, the measures should report how well the model $\hat{\mathbf{p}}(\mathbf{x})$ predicts the conditional distributions $\mathbf{p}(\mathbf{x})$ of the response variable, rather than the classification performance of the resulting classifier.

With n cases, the classifier $f(\mathbf{x})$ has to fit n values of y. The target of the descriptive model $\hat{\mathbf{p}}(\mathbf{x})$ is quite different. It has to fit a probability distribution $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^r$ for each of the c different[3] observed profiles $\mathbf{x}$. Hence it has to predict rc probabilities (indeed only the non zero probabilities). Notice that $c \leq n$ is often much smaller than n, especially when all predictors are categorical and n is large. The target of the descriptive model is thus a $r \times c$ cross table that synthesizes the relevant information contained in the data. Goodness-of-fit should here assess how well the model reproduces this synthetic representation.

The fit of a synthetic representation is very similar to what is done, for instance, in the statistical log-linear modeling of multidimensional cross tables where goodness-of-fit has to do with the ability of the model to reproduce the cross-classification rather than the exact classification of each case. It can also be compared with structural equation modeling where the goal is to fit the observed covariance or correlation matrix of the observed variables rather than the exact individual values of the endogenous variables. In these two latter examples, the target is assimilated to the exact predictions generated by a so called saturated model, i.e. a model with as much independent parameters as independent values in the synthetic data representation. Similarly, a saturated tree will be a tree that predicts exactly the target $r \times c$ cross table.

In the next section, we specify the notions of target and predicted table in the case of decision trees. We introduce also the trick that will allow the machinery of statistical tests and goodness indicators work on decision trees.

3 Target Table and Predicted Table

When all variables are discrete, the empirical counterpart of the conditional distributions $\mathbf{p}(\mathbf{x})$ can be derived from the $r \times c$ contingency table $\mathbf{T}$ that cross classifies the r values of Y with the c profiles. Table 1, for example, synthesizes 100 data used for predicting the *marital status* (yes, no) with the two predictors

[3] If each predictor x_s, $s = 1, \ldots, p$ has c_s different values, the number c of possible different profiles, and hence conditional distributions, is at most $\prod_{s=1}^{p} c_s$.

Table 1. Example of a response × predictors contingency table **T**

	male			female			
married	primary	secondary	tertiary	primary	secondary	tertiary	total
no	11	14	15	0	5	5	50
yes	8	8	9	10	7	8	50
total	19	22	24	10	12	13	100

gender (M = male, F = female) and *activity sector* (P = primary, S = secondary, T = tertiary). Letting n_{ij} denote an element of table **T** and $n_{.j}$ the total of column j, the maximum likelihood estimation of $\mathbf{p}_{|j} = \mathbf{p}(\mathbf{x}_j)$ is indeed the vector of the observed frequencies $n_{ij}/n_{.j}, i = 1, \ldots, r$. Each column of the table **T** corresponds to the terminal node of a so called *saturated tree*, i.e. the tree that exhausts all splits and generates the finest partition for the retained predictors (see Figure 1, left.)

As will be shown, an induced tree provides a prediction $\hat{\mathbf{T}}$ of **T**. Measuring the (descriptive) goodness-of-fit of the tree consists then in measuring how well $\hat{\mathbf{T}}$ fits **T**. To explain how we get $\hat{\mathbf{T}}$ from an induced tree, we consider the following rebuilding model where $\hat{\mathbf{T}}_j$ stands for the j-th column of $\hat{\mathbf{T}}$

$$\hat{\mathbf{T}}_j = n\, a_j \hat{\mathbf{p}}_{|j}, \quad j = 1, \ldots, c \tag{1}$$

The parameters are the total number of cases n, the proportions a_j of cases in column (terminal node) $j = 1, \ldots, c$ and the c column distribution vectors $\mathbf{p}_{|j}$. The a_j's are naturally estimated by $n_{.j}/n$. The only trick required concerns the estimation of the $\mathbf{p}_{|j}$'s. Indeed, the induced tree has generally $q < c$ terminal nodes which generate a $r \times q$ table $\mathbf{T}^a$ not conformable with **T**. To render table $\mathbf{T}^a$ conformable, we have to extend it or equivalently extend the induced tree.

Definition 1. *The* maximal extension of an induced tree *is obtained by maximally further splitting each terminal node $k = 1, \ldots, q$ of the tree and by distributing the cases in each new node according to the distribution $\mathbf{p}^a_{|k}$ of its parent terminal node of the induced tree. (See Figure 1, right.)*

Formally, letting $\mathcal{X}_k$ denote the subset of profiles that belong to the group defined by the terminal node k, the maximally extended tree leads to the following estimations of $\mathbf{p}_{|j}$

$$\hat{\mathbf{p}}_{|j} = \mathbf{p}^a_{|k} \quad \text{for all } \mathbf{x}_j \in \mathcal{X}_k \;\; k = 1, \ldots, q \tag{2}$$

For example, if the induced tree is the tree with the white nodes in the right part of Figure 1, its maximal extension is obtained by completing the tree with the grayed nodes. The distributions in the six terminal nodes of the extension

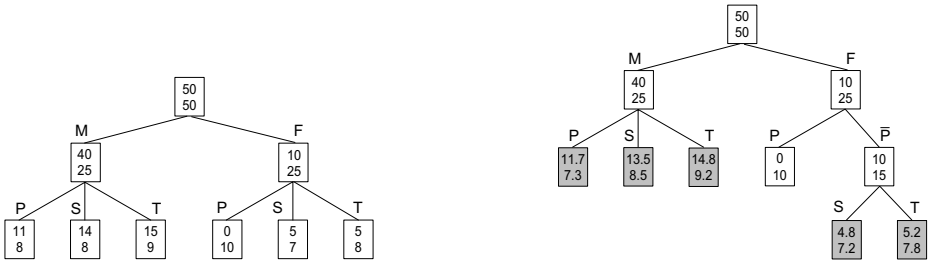


Fig. 1. (left) saturated tree and (right) an induced tree (white nodes) together with its maximal extension (white + grey nodes). The predictors are the gender (M, F) and the sector of activity (P=primary, S=secondary, T=tertiary) and the response variable is the marital status (yes, no). Observe that the distribution in the grayed nodes is the same as in their parental white node.

Table 2. The predicted table $\hat{\mathbf{T}}$

	male			female			
married	primary	secondary	tertiary	primary	secondary	tertiary	total
no	11.7	13.5	14.8	0	4.8	5.2	50
yes	7.3	8.5	9.2	10	7.2	7.8	50
total	19	22	24	10	12	13	100

follow from those of the three (white) leaves of the induced tree:

$$\hat{\mathbf{p}}_{|MP} = \hat{\mathbf{p}}_{|MS} = \hat{\mathbf{p}}_{|MT} = \mathbf{p}^a_{|M} = \begin{pmatrix} 40/65 \\ 25/65 \end{pmatrix}$$

$$\hat{\mathbf{p}}_{|FP} = \mathbf{p}^a_{|FP} = \begin{pmatrix} 0/10 \\ 10/10 \end{pmatrix}$$

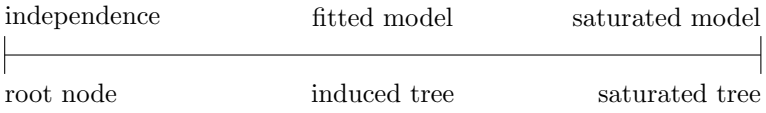
$$\hat{\mathbf{p}}_{|FS} = \hat{\mathbf{p}}_{|FT} = \mathbf{p}^a_{|F\bar{P}} = \begin{pmatrix} 10/25 \\ 15/25 \end{pmatrix}$$

From the leaves of the extension we derive the predicted table $\hat{\mathbf{T}}$ depicted in Table 2.

4 Goodness-of-Fit Measures for Induction Trees

Having defined the target table $\mathbf{T}$ and the one $\hat{\mathbf{T}}$ predicted by the induced tree, we can now apply the statistical tests and goodness indicators used in the statistical modeling of cross tables (see for instance [4]). We first discuss statistics that measure the divergence between the predicted and target tables or, equivalently, between the induced and saturated trees. We then address the issue of trees comparison and propose some R^2 like indicators that measure the

improvement over a baseline model. We complete the Section with AIC and BIC information criteria for induced trees.



4.1 Chi-Square Statistics

The most popular divergence Chi-square statistics are the Pearson X^2 and the deviance G^2 statistics. Under some regularity conditions (see for instance [5] chap. 14) these statistics have, when the induced tree is correct, an asymptotical Chi-square distribution. In our case, the deviance G^2 reads

$$G^2 = 2\sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij} \ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right) .$$

For practical use of this statistics, we have to determine the associated degrees d of freedom. These are given by the number of independent constraints (2) on the posterior probability vectors $\mathbf{p}_{|j}$'s. Each vector $\mathbf{p}_{|j}$ has $r-1$ independent terms and there are only $q \leq c$ different distributions among the c vectors. Hence the degrees of freedom are: $d = (r-1)(c-q)$ The independence model I between the predictors and the response corresponds to the tree with the sole root node. In this case, we have $q = 1$ and get the well known number $d_I = (r-1)(c-1)$ of degrees of freedom for the independence test. Likewise, for the saturated model S, we have $q = c$, which gives $d_S = 0$.

For the white induced tree M in the right panel of Figure 1, we get $G^2(M) \simeq 0.18$. Since $r = 2$, $c = 6$ and $q = 3$, we have $d_M = 3$ degrees of freedom. The value of the G^2 is in this example very small. Its p-value is about 98%, which indicates an excellent fit.

Though we should theoretically be able to test statistically the goodness-of-fit of trees with Chi-square statistics, it is well known that the scope of the test is limited when the number of data is large. Further, the required regularity conditions, especially interiority that requires non zero expected frequencies, may not hold when the number of variables becomes large. Nevertheless, the G^2 statistic proves in any case useful for model comparison.

Thanks to an additive property, G^2 permits to test the difference between nested models. Let M_2 be a restricted form of model M_1. Then, the deviance between the two models is (see [4] p. 211)

$$G^2(M_2|M_1) = G^2(M_2) - G^2(M_1)$$

which, if M_2 is correct, has an asymptotic Chi-square distribution with $d_2 - d_1$ degrees of freedom.

For induction trees, the deviance between nested trees, i.e. between a tree and the same tree after the pruning of a subtree of interest, provides a natural way to test the statistical relevance of the subtree. This way of testing a whole part of the tree clearly complements the information provided by the criteria locally optimized at each split. In the example of Figure 1, we can for instance test if the activity sector has a significant role, by comparing the induced tree $M_1 = M$ with the tree M_2 that includes only the split by gender. For the latter, we get $G^2(M_2) = 8.41$ with $d_2 = 4$ degrees of freedom. The divergence between the two trees is thus

$$G^2(M_2|M_1) = 8.41 - 0.18 = 8.23 \quad \text{with} \quad d_2 - d_1 = 4 - 3 = 1$$

Its p-value is 0.4%. This is clearly less than the usual 5% threshold from which we conclude that the split is statistically significant.

4.2 Comparison with a Baseline Model

This Section is devoted to R^2-type indicators that measure the relative quality improvement over a baseline tree. For trees, a natural choice for the naive baseline model is the tree with the sole root node, i.e. the independence case. We discuss successively the relative reduction in the error rate, the reduction in the entropy and the relative fit improvement.

Remark on the proportion of reduction in the error rate. Though we are not interested in the prediction of individual values, it is worth mentioning that R^2-type measures make also sense for the error rate. The proportion of reduction of the learning error rate corresponds indeed to the Goodman and Kruskal [6] first association measure $\lambda_{y|\text{partition}}$. The literature [7] [8] gives formulas of its asymptotic variance that can be used to test its statistical significance. We will not go in further details, since our purpose is here to measure the descriptive quality of a tree and not its prediction performance.

Information Gain. The gain in information is typically measured by the reduction in the entropy of the response variable. Since we are interested in quality measures of the whole tree, we consider the global reduction achieved by the induced tree when compared with the root node. This should not be confused with the local entropy reduction that some tree growing algorithms maximize at each node.

The relative gain of information with respect to the root node can for instance be measured with the following indexes:

$$\hat{\tau}_{M|I} = \frac{n \sum_i \sum_j \frac{\hat{n}_{ij}^2}{n_{\cdot j}} - \sum_i n_{i\cdot}^2}{n^2 - \sum_i n_{i\cdot}^2} \tag{3}$$

$$\hat{u}_{M|I} = \frac{\sum_i \frac{n_{i\cdot}}{n} \log_2 \frac{n_{i\cdot}}{n} - \sum_j \frac{n_{\cdot j}}{n} \sum_i \frac{\hat{n}_{ij}}{n_{\cdot j}} \log_2 \frac{\hat{n}_{ij}}{n_{\cdot j}}}{\sum_i \frac{n_{i\cdot}}{n} \log_2 \frac{n_{i\cdot}}{n}} \tag{4}$$

The $\hat{\tau}_{M|I}$ is the second nominal association measure of Goodman and Kruskal [6]. It measures the proportion of reduction in the quadratic entropy $H_Q(\mathbf{p}) = \sum_i p_i(1-p_i)$ also known as the Gini variation index. The $\hat{u}_{M|I}$ is known in statistics as the Theil uncertainty coefficient [9] [10]. It measures the proportion of reduction in Shannon's entropy $H_S(\mathbf{p}) = -\sum_i p_i \log_2 p_i$.

For our induced tree M in Figure 1, we have for example $\hat{\tau}_{M|I} = 0.145$ and $\hat{u}_{M|I} = 0.132$. These values indicate a relative gain of information of about 14%. This should in turn be compared with the gains achieved by the saturated tree that are respectively $\hat{\tau}_{S|I} = 0.146$ and $\hat{u}_{S|I} = 0.134$, from which it appears that almost the induced tree captures almost the whole potential gain we can get from the predictors.

For testing the statistical significance of the information gain, i.e. the hypotheses $H_0 : \tau_{M|I} = 0$ et $H_0 : u_{M|I} = 0$, there are two possibilities. We can use available asymptotic variances [8] and a gaussian approximation. It is most powerful, however, to use the following transformations of the indexes:

$$C(I|M) = (n-1)(r-1)\,\hat{\tau}_{M|I} \tag{5}$$

$$G^2(I|M) = \big(-2\sum_i n_{i\cdot} \log(n_{i\cdot}/n)\big)\,\hat{u}_{M|I} \tag{6}$$

The first has been suggested by Light and Margolin [11]. In a setting corresponding to the case where the tree is the saturated tree $M = S$, these authors show that under the independence hypothesis ($\tau_{S|I} = 0$), the quantity $C(I|S)$ follows asymptotically a χ^2 with d_I degrees of freedom. Replacing S with a restraint model M just requires adjusting the degrees of freedom. Hence, for the general case, $C(I|M)$ has a χ^2 distribution with $d_I - d_M$ degrees of freedom.

The transformation of $\hat{u}_{M|I}$ shows that testing the significance of $u_{M|I}$ is equivalent to testing the difference in fit between I and M with $G^2(I|M) = G^2(I) - G^2(M)$. Both transformations (5) and (6) have thus, under H_0, the same asymptotic χ^2 distribution with $d_I - d_M$ degrees of freedom.

For our illustrative induced tree, we get $C(I|M) = 14.32$ and $G^2(I|M) = 18.36$. These are large values for the degrees of freedom $d_I - d_M = 5 - 3 = 2$. They confirm the statistical significance of the global information gain.

Pseudo R^2. The goal is here to measure the proportion of the divergence between the baseline and saturated trees that can be catched by the induced tree. If we measure the divergence with the G^2 statistic, we can use for example the pseudo R^2:

$$R^2 = 1 - \frac{G^2(M)}{G^2(I)}$$

or its form adjusted for the degrees of freedom

$$R^2_{\text{ajust}} = 1 - \frac{G^2(M)/d_M}{G^2(I)/d_I}$$

For our example, we have $G^2(I) = 18.55$, $d_I = 5$, $G^2(M) = .18$ et $d_M = 3$ from which we get $R^2 = .99$ and $R^2_{\text{ajust}} = .984$. Again, these values confirm that the induced tree catches almost the entire lack of fit of the independence tree.

4.3 Goodness-of-fit and complexity

Clearly, the more we grow a tree the better is the fit. For description purposes the tree should on the contrary be as simple as possible. Parsimony is not only an interpretability requirement. It is also essential to ensure the stability of the description. Thus when selecting a descriptive tree we have to trade off between fit and complexity. This is precisely what the Akaike (AIC) [12] and Bayesian (BIC) [13] [14] information criteria are intended for. Their principle is simply to penalize the fit for the complexity. In their definition the AIC and BIC measure the complexity in terms of number of independent parameters of the model. For trees, we have to count the independent parameters of the rebuilding formula (1). This gives $qr - q + c$ and the AIC and BIC read thus

$$\text{AIC}(M) = G^2(M) + 2(qr - q + c)$$
$$\text{BIC}(M) = G^2(M) + (qr - q + c)\log(n)$$

The BIC penalizes complexity increasingly with n and more strongly than does AIC.

There are alternative forms of the BIC. Raftery [15], for example, proposes $BIC = G^2 - d\log(n)$, where d is the number of degrees of freedom which is in our case $d = (r-1)(c-q)$. Since d is reduced by one unity each time we add an independent parameter, the penalization remains indeed the same. Both formulations differ just by the constant translation factor cr.

These information criteria are specially useful for model selection. Among several trees, the one that minimizes the criteria provides the best compromise between fit and adjustment.

To illustrate, let us compare our induced tree M with the variant M^* where the node "female" is split into the three sectors P, S, I instead of the binary split into primary P and not primary $\bar{P}$. For both trees we have $n = 100$, $c = 6$ and $c = 2$. For M, we have $q = 3$ and hence $(qr - q + c) = 9$, and for M^*, $q = 4$ and $(qr - q + c) = 10$. Since $G^2(M) = 0.18$ and $G^2(M^*) = .16$, we get $\text{AIC}(M) = 18.18$ and $\text{AIC}(M^*) = 20.16$. Likewise, we get $\text{BIC}(M) = 41.63$ and $\text{BIC}(M^*) = 46.21$. Both criteria indicate the the simpler tree M should be preferred to M^*. The fit improvement of M^* over M is not sufficient to justify the increase in complexity. Note that the induced tree outperforms both the independence model ($\text{AIC}(I) = 32.55$, $\text{BIC}(I) = 50.78$) and the saturated tree ($\text{AIC}(S) = 24$, $\text{BIC}(S) = 55.26$).

5 Conclusion

We have addressed the issue of evaluating the descriptive quality of an induced tree. The central point in the approach presented is the trick of the extended

tree that allows comparison with the target saturated tree. The approach is quite simple and its implementation as an a posteriori quality measure in tree building softwares should be straightforward. Its main limitation is the number of profiles involved by the predictors that should not be too huge.

Criteria like the AIC and BIC are powerful model selection tools. We plan therefore to use them at the tree building stage. Two aspects merit here special attention. First, we should establish some local criteria that when optimized (at each node) provides a solution equivalent to the global minimization of the AIC or BIC. Second, there is the issue of continuous attributes. Induction tree methods usually discretize them dynamically, i.e. optimally during the tree building process. We have then to solve the question of the a priori target table to use for the computation of the divergence Chi-squares.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification And Regression Trees. Wadsworth International Group, Belmont, CA (1984)
2. Han, J., Kamber, M.: Data Mining: Concept and Techniques. Morgan Kaufmann, San Francisco (2001)
3. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
4. Agresti, A.: Categorical Data Analysis. Wiley, New York (1990)
5. Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis. MIT Press, Cambridge MA (1975)
6. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. Journal of the American Statistical Association **49** (1954) 732–764
7. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications iv: simplification of asymptotic variances. Journal of the American Statistical Association **67** (1972) 415–421
8. Olszak, M., Ritschard, G.: The behaviour of nominal and ordinal partial association measures. The Statistician **44** (1995) 195–212
9. Theil, H.: Economics and Information Theory. North-Holland, Amsterdam (1967)
10. Theil, H.: On the estimation of relationships involving qualitative variables. American Journal of Sociology **76** (1970) 103–154
11. Light, R.J., Margolin, B.H.: An analysis of variance for categorical data. Journal of the American Statistical Association **66** (1971) 534–544
12. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In Petrox, B.N., Caski, F., eds.: Second International Symposium on Information Theory. Akademiai Kiado, Budapest (1973) 267
13. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6** (1978) 461–464
14. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90** (1995) 773–795
15. Raftery, A.E.: Bayesian model selection in social research. In Marsden, P., ed.: Sociological Methodology. The American Sociological Association, Washington, DC (1995) 111–163