ARBEITSPAPIERE ZUR
MATHEMATISCHEN WIRTSCHAFTSFORSCHUNG

Some Remarks about the Usage of Asymmetric Correlation Measurements for the Induction of Decision Trees

Andreas Hilbert

Heft 180/2002

Institut für Statistik und Mathematische Wirtschaftstheorie
Universität Augsburg

Dr. Andreas Hilbert
Universitätsstr. 16, D - 86159 Augsburg

Telefon +49 821 598-4155
Telefax +49 821 598-4226
EMail andreas.hilbert@wiwi.uni-augsburg.de

1 Introduction

Decision trees are used very successfully for the identification resp. classification task of objects in many domains like marketing (e.g. Decker, Temme (2001)) or medicine. Other procedures to classify objects are for instance the logistic regression, the logit- or probit analysis, the linear or squared discriminant analysis, the nearest neighbour procedure or some kernel density estimators. The common aim of all these classification procedures is to generate classification rules which describe the correlation between some independent exogenous variables resp. attributes and at least one endogenous variable, the so called class membership variable.

If there are exclusively metric scaled exogenous attributes the procedures often try to aggregate these attributes in a way that the so built new quantity describes the class membership as good as possible. The accuracy of this identification procedure is often measured by variance based measurements. The regression based procedures use the least squares approach and serve especially for the classification of binary scaled membership variables. If they are above all nominal scaled exogenous attributes the procedures divide the objects in a way that the so generated partitions are as homogeneous as possible. The homogeneity itself is measured by some deviation measurements like the Entropy measure or by some generalized variance based measurements like the Gini index. Only the CHAID algorithm by Kaas (1980), a special decision tree procedure, uses a correlation measure, the χ^2 correlation measurement, to generate some classification rules in order to describe the correlation between the involved attributes and the class membership.

Although the proper task of the classification procedures is to identify and explain the correlation between at least one membership variable and in general several exogenous attributes, only one algorithm actually uses a correlation measurement to do that. Furthermore, it is noteworthy that this correlation measurement is symmetric in its nature although the classification task is asymmetric: at least one exogenous attributes should explain at least one endogenous variate and not vice versa.

Thus, the possibility to classify objects in the manner of decision trees using asymmetric correlation measures should be analyzed. It will be shown that some well-known decision tree algorithms like ID3, C4.5 or CART can be understood as special versions of a generalized decision tree based on asymmetric correlation measurements. But in contrast to these procedures the measure to be proposed offers the chance to do some inferential statistics as well.

2 Decision Trees

The construction of decision trees is described, among others, by Breiman et al. (1984), who present an important and well-known monograph on classification trees. A number of standard techniques have been developed, for example like the basic algorithm ID3 by Quinlan (1986), the C4.5 algorithm by Quinlan (1993), the above mentioned χ^2-based algorithm CHAID by Kaas (1980) or the CART algorithm of Breiman et al. (1984). A very interesting *multi-disciplinary* survey of the construction of decision trees and related topics is presented by Murthy (1998).

Broadly speaking, a decision tree is built from a set of data having attributes $X_1, \ldots, X_n$ and a class or membership variable Y. The result of the process is represented as a flow-chart-like tree in which each internal node specifies a decision on an attribute and each branch denotes an outcome of these decisions. Furthermore, each end node or leaf of the tree corresponds to an subset of objects with the same class or to objects for which the homogeneity is as good as requested. Thus the leaf nodes represent classes or class distributions.

The basic algorithm for the induction of a decision tree itself is a "greedy algorithm that generates decision trees in a top-down recursive divide-and-conquer manner" (Han, Kamber (2001)). The **basic strategy** consists of the following steps:

- The tree starts with a single node, the so called *root*, representing the whole considered data set.
- If the objects all belong to the same class, then the node becomes to a leaf and is labeled with that class.
- Otherwise, the algorithm uses a *split criterion* for selecting the attribute that will best separate the set of objects into individual subsets or classes. This attribute becomes the *decision attribute* for the given node. Depending on the used split criterion, all attributes, categorical as well as continuous-valued attributes, resp. only categorical or artificially categorized attributes could be used (*Attribute Selection Measure*).
- A branch is created for each known value of the decision attribute and the data set (of the considered node) is partitioned accordingly.
- The algorithm repeats the same procedure recursively to form a decision tree for all subsets of each partition. Once an attribute has occurred at a node, it need not be considered in any of the node’s descendents.

- The recursive partitioning is finished when a so called *stop criterion*, which depends on the used split criterion and/or the underlying type of induction procedure, is fulfilled (*Pre-Pruning*).

This basic strategy could be found in most of the well-known algorithms for induction of decision tree like for instance in ID3, C4.5 or CART as well as in CHAID. But only the split criterion resp. the attribute selection measure and the way of pruning separate the different algorithms.

The most important **attribute selection measures** are the *information gain*, the *information gain ratio* (both are based on the well-known *entropy*), the *Gini index*, the *twoing value* and the χ^2*-based measure of correlation*. All of them attempt to partition the data set in such a way that the resulting subsets are as homogeneous as possible with respect to the class membership of the objects. Thus, the general aim is to minimize the so called *impurity* of the partition.

When decision trees are built, many of the branches may reflect anomalies in the data set due to noise or outliers. Tree **pruning** methods address this problem of *overfitting* the data. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify unknown objects. To prune a tree there are two common approaches, the so called pre-pruning technique and the so called post-pruning technique.

In the **pre-pruning** approach, the tree is pruned by halting its construction early, for instance by deciding not to further split or partition the subset of objects at a given node. This decision could base on measures like χ^2, information gain, and so on, which are used to assess the goodness of a split. If the partitioning of the objects of a node would result in a split that falls below a prespecified threshold, then further dividing of the given subset is halted. There are some difficulties, however, in choosing an appropriate threshold. High threshold could result in oversimplified trees, while low threshold could result in trees with a high probability of overfitting. Other possibilities to control the halting of the induction are for instance the (too small) number of objects in a node, the fact that all objects belong to the same class or that all objects are identical with respect to the given attributes.

The second approach, the **post-pruning**, attempts to remove branches and nodes from a "fully grown" tree in such a way that the resulting pruned tree optimizes some special accuracy measures. The best-known techniques in this framework are the error-complexity-pruning by Breiman et al. (1984), the pessimistic-error-pruning by Quinlan (1986) and

finally the error-based-pruning by Quinlan (1993). While the procedure of Breiman et al. is based upon an accuracy measure that is calculated for a so called *test data set*, which have to be also available, the algorithms of Quinlan only use the given (training) data set to prune the tree.

Alternatively, pre-pruning and post-pruning may be interleaved for a combined approach. Post-pruning requires more computation than pre-pruning, but leads to a more reliable tree, in general.

3 Attribute Selection and Correlation Measures

The main task of all classification procedures is to explain the correlation between at least one dependent, also called endogenous membership variable Y and in general several independent, exogenous attributes $X_1, \ldots, X_n$. Regarding now the induction of decision trees, this classification task can also be seen as an iterative bivariate analysis of pairs (X_i, Y) for a given data set, i.e. objects belonging to a given node. To consider such pairs (X_i, Y) in the manner of bivariate analysis the techniques of correlation analysis are available. Following Hilbert (1998) there exist several types of correlations:

- Type 1 measures the deviation from the stochastical independence, is symmetric in the way to use the attributes and is based on the χ^2 measure.
- Type 2 compares the conditional distributions of one attribute, given the values of another, distinguish between cause and effect, i.e. is asymmetric, and can be considered as an unweighted variety of type 1.
- Type 3 considers the *reduction of the prediction error* for one attribute, given the value of another attribute, is also asymmetric and known as *predictive association*.
- Type 4 is based on the concept of pairwise comparisons, does not distinguish between cause and effect, and is above all suitable for at least ordinal scaled attributes.

While the correlation measures based on type 1 as for instance the χ^2 measure itself and its derived measures like the ϕ coefficient, Tschuprow's contingency measure T or Cramer's V are very popular, correlation measures of type 2 to 4 are not used very often. The main reason why these type 1 measures are used so often is the knowledge of the (asymptotical) distribution of these measures and as a result the possibility for doing some inferential statistics in order to test the correlation between the attributes. But "the fact that an excellent test of independence may be based on χ^2 does not at all mean

that χ^2, or some simple function of it, is an appropriate measure of degree of association" (Goodman, Kruskal (1954), 740). One difficulty with the use of these traditional measures is that it is difficult to compare meaningfully their values of two pairs (X_i, Y) and (X_j, Y) resp. to interpret their values in an operational way.

Thus, Goodman and Kruskal (1954) proposed another concept to measure the correlation between two attributes, which is based upon an idea of Guttman (1941) and well-known as **predictive association**. Their concept is able to reflect the extent of the ability of an attribute to predict the values of another attribute, for instance of a class membership variable.

To construct a correlation measurement that follows this concept the following has to be done: Defining $\mathbf{PE}(Y)$ as the prediction error of an attribute Y with values $y_1, \ldots, y_m$ and $\mathbf{PE}(Y|X_i)$ as the equivalent prediction error of the same attribute Y given an attribute X_i with values $x_{i1}, \ldots x_{im_i}$. Then, it can easily be seen that the following equation holds:

$$\mathbf{PE}(Y|X_i) = \sum_{j=1}^{m_i} \mathbf{P}_{X_i}(x_j) \cdot \mathbf{PE}(Y|X_i = x_j) \tag{1}$$

$\mathbf{P}_{X_i}(x_j)$ denotes the probability or (in case of a sample) the relative frequency of an attribute X_i having the value x_j. Using this denotation the measure of **predictive association** is defined by

$$\mathbf{CM}^{\mathrm{PRE}}_{(X_i \to Y)} := \frac{\mathbf{PE}(Y) - \mathbf{PE}(Y|X_i)}{\mathbf{PE}(Y)}. \tag{2}$$

This quantity is a general *asymmetric measurement of correlation* between the two involved attributes with X_i as cause and Y as effect. Because of its characteristic to reflect the extent of the ability of an attribute to predict the values of another attribute it is also called **P**roportional-**R**eduction-of-**E**rror or PRE coefficient. Based upon this very general definition of a PRE coefficient, only the operational form of $\mathbf{PE}(\cdot)$ has to be specified to obtain a concrete measurement of correlation. If considering, however, the definition and meaning of the coefficient, it is easy to understand why every deviation coefficient is a good choice for the prediction error $\mathbf{PE}(\cdot)$: the smaller the deviation of an attribute the better the prediction of it.

The well-known deviation coefficients for nominal scaled attributes are *Shannon's entropy* **H** (Shannon (1948)), the *deviation coefficient* $\mathbf{S}_H$ by Herfindahl and a measure that is based on the probability resp. frequency of the mode of the distribution of the considered attribute, the so called *modality measurement* **M** (Hilbert (1998), 115–122).

Shannon's entropy

If Shannon's entropy **H** is used to declare the prediction error, under the known assumption the following definition holds:

$$\mathbf{H}(Y) := -\sum_{k=1}^{m} \mathbf{P}_Y(y_k) \cdot \mathrm{ld}\, \mathbf{P}_Y(y_k) \tag{3}$$

An equivalent definition of the entropy of Y given X_i can be obtained by the use of (1) and (3). Both definitions consequently lead to the following PRE coefficient of correlation for two nominal scaled attributes:

$$\mathbf{CM}^{\mathrm{PRE},\mathbf{H}}_{(X_i \to Y)} := \frac{\sum_{k=1}^{m} \mathbf{P}_Y(y_k) \cdot \mathrm{ld}\, \mathbf{P}_Y(y_k) - \sum_{k=1}^{m} \sum_{j=1}^{m_i} \mathbf{P}_{(X_i,Y)}(x_j,y_k) \cdot \mathrm{ld}\, \mathbf{P}_{(Y|X_i=x_j)}(y_k)}{\sum_{k=1}^{m} \mathbf{P}_Y(y_k) \cdot \mathrm{ld}\, \mathbf{P}_Y(y_k)} \tag{4}$$

This predictive association coefficient is, except for the denominator, equal to the well-known **information gain** measure by Quinlan (1986). The denominator which ensures the *proportional* part of the coefficient has the task to normalize the measure and can be neglected for the induction of a decision tree: the attribute Y corresponds to the membership variable in a node and thus, all considered pairs $(X_1,Y),\ldots,(X_n,Y)$ have the same endogenous variable resp. the same denominator while calculating the correlation coefficient resp. the split criterion at a node. Thus, the information gain can be treated as an asymmetric correlation measure and the proposed split criterion is nothing but a predictive association coefficient.

The deviation coefficient by Herfindahl

If, however, the deviation coefficient $\mathbf{S}_H$ by Herfindahl is used to define the prediction error, the following definition holds:

$$\mathbf{S}_H(Y) := 1 - \sum_{k=1}^{m} \mathbf{P}_Y(y_k)^2 \tag{5}$$

An equivalent procedure leads to the following PRE coefficient:

$$\mathbf{CM}^{\mathrm{PRE},\mathbf{S}_H}_{(X_i \to Y)} := \frac{\sum_{j=1}^{m_i} \mathbf{P}_{X_i}(x_j) \sum_{k=1}^{m} \mathbf{P}_{(Y|X_i=x_j)}(y_k)^2 - \sum_{k=1}^{m} \mathbf{P}_Y(y_k)^2}{1 - \sum_{k=1}^{m} \mathbf{P}_Y(y_k)^2} \tag{6}$$

Similar to the above situation, the proposed PRE coefficient is equal to a well-known split criterion, the **Gini index** by Breiman et al. (1984) - except again for the denominator for standardization. Furthermore, all the above mentioned remarks are valid as well. By the way, the standardization of the Gini index proposed by Zhou and Dillon (1991) differs from this and should not be recommended because of the fact that their procedure generates a correlation or split measure which is not asymmetric but symmetric.

The modality measurement

Considering the information gain and the Gini index as special correlation measures it can easily be realized that the interpretation of these quantities is very difficult and in common without a concrete meaning. Only the relative comparison of the different measures is possible and leads to a best splitting attribute. An absolute appraisal of the values, however, is not possible. On the other hand, this very characteristic is one of the most important advantages of the third deviation measurement, the modality measure **M**. The use of this measurement to describe the prediction error for an attribute leads to a definition according to

$$\mathbf{M}(Y) := 1 - \max_{k=1,\ldots,m} \mathbf{P}_Y(y_k) \tag{7}$$

and respectively to the following PRE coefficient:

$$\mathbf{CM}^{\mathrm{PRE},\mathbf{M}}_{(X_i \to Y)} := \frac{\sum_{j=1}^{m_i} \max_{k=1,\ldots,m} \mathbf{P}_{(X_i,Y)}(x_j, y_k) - \max_{k=1,\ldots,m} \mathbf{P}_Y(y_k)}{1 - \max_{k=1,\ldots,m} \mathbf{P}_Y(y_k)} \tag{8}$$

This coefficient is known as the predictive association coefficient λ by Goodman and Kruskal (1954) and has some similarities to the split criterion *theta* by Messenger and Mandell (1972) who propose a less known decision tree algorithm THAID based on that theta coefficient. Also Breiman et al. (1984) analyze a similar coefficient in the framework of reducing missclassification costs.

In contrast to the other predictive association measurements Goodman and Kruskal's λ can easily be interpreted. If nothing is known about the distribution of the values of an attribute Y, the best prediction of Y is the mode y_{mod} of the attribute with a prediction error $1 - \mathbf{P}_Y(y_{\mathrm{mod}})$. Given an exogenous attribute X_i this prediction error can be calculated according to (1) to describe the support of X_i to predict Y. Thus, λ corresponds to the proportional reduction of the prediction error of Y as long as the mode is the best prediction for an attribute. For instance, a value of 0.2 means that there is a 20 percent opportunity to improve the prediction of an attribute Y using an exogenous attribute X_i. In other words, the probability to make an error while predicting a value of an attribute Y by selecting the mode of this attribute decreases by 20 percent using the exogenous attribute. Further characteristics of λ are:

- If and only if $\lambda = 0$ then the knowledge of a value of X_i is no help for predicting the value of Y.
- If and only if $\lambda = 1$ then the knowledge of a value of X_i leads definitely to one value of Y.
- If Y and X_i are stochastical independent then $\lambda = 0$. The reversal doesn't hold.

With respect to the sample form $\hat{\lambda}$ of the coefficient, Goodman and Kruskal (1963) prove that $\hat{\lambda}$ is asymptotically unbiased and asymptotically normal distributed. Furthermore, they give the following expression for the asymptotical variance of $\hat{\lambda}$:

$$\mathbf{Var}(\hat{\lambda}) = \frac{1 - \sum_{j=1}^{m_i} \mathbf{P}_{(X_i,Y)}(x_j, y_{k_j^*})}{N \cdot \left(1 - \mathbf{P}_Y(y_{k^*})\right)^3} \tag{9}$$
$$\cdot \left(\sum_{j=1}^{m_i} \mathbf{P}_{(X_i,Y)}(x_j, y_{k_j^*}) + \mathbf{P}_Y(y_{k^*}) - 2 \cdot \sum_{\substack{j=1 \\ k_j^*=k^*}}^{m_i} \mathbf{P}_{(X_i,Y)}(x_j, y_{k_j^*}) \right)$$

N is the sample size (at a node), $\mathbf{P}(\cdot)$ the known probability or the sample estimator of the unknown probability of the (common) distribution of Y and/or X_i, k^* the index of the mode of Y and k_j^* the index of the mode of Y given $X_i = x_j$. Using these sampling properties it is now possible to calculate some *confidence intervals* for $\lambda(X_i \rightarrow Y)$, to do some one- or two-sided *inferential tests* for $\lambda(X_i \rightarrow Y)$ or to test $\lambda(X_i \rightarrow Y)$ against $\lambda(X_j \rightarrow Y)$ for some i, j.

Comparison of the different split criteria

Beside this very interesting and important property to do some inferential statistics for this correlation measure resp. split criterion it has to be analyzed which further advantages or disadvantages – compared to the Gini-index or the information gain measure – Goodman and Kruskal's λ has. To do so, it is useful to have a look at the functional form of the measurements, described for a binary membership attribute Y (figure 1).

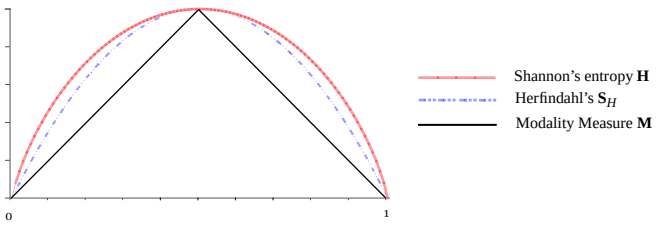

Fig. 1. Deviation measures for a binary attribute Y

First of all, it can be seen that Shannon's entropy **H** and Herfindahl's deviation measurement $\mathbf{S}_H$ are very similar. This also explains why the results of the induction of a decision tree using the Gini-index (which uses $\mathbf{S}_H$) and the information gain measure (which uses **H**) are often identical. Furthermore, **H** and $\mathbf{S}_H$ assess all distributions, which are more

or less similar to an equal distribution, in the same manner. Here, differences are very difficult to analyze. The modality measure **M**, however, reacts in such a situation much more sensitive. On the other hand, **M** has some difficulties in the discrimination of distributions which are similar to an extrem unbalanced one. Here, the other measures have the advantage to react very sensitive to a variation of the probabilities. This also means that the modality measure, taken as a split criterion, does not prefer the so called end-cut splits. But if this is an advantage or disadvantage can not easily be answered, even though Breiman et al. (1984) do not feel very well about a decision tree algorithm which has not this end-cut-split feature. Here, some further research and simulation studies are necessary. Finally, it can be shown that the modality measure **M** and the measures **H** and $\mathbf{S}_H$ generate a different order with respect to the deviation of two distributions. While, for instance, the entropy measurement and Herfindahl's deviation measure resp. their analog split version both prefer the distribution on the right-hand side in figure 2 (see also table 1), the modality measure prefers the left-hand distribution which is – in the author's opinion – much more desirable in the framework of decision trees.

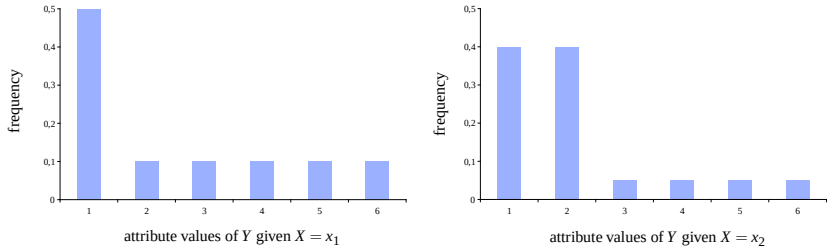

Fig. 2. Distributions of Y given two different values of an exogenous attribute X

Another critical situation arises by using the entropy resp. Herfindahl's measure if the distributions to compare with are more or less equal. Figure 3 shows two such distributions. Four of the six probabilities of $Y|X = x_1$ and $Y|X = x_3$ are equal, only the conditional probability of $Y = 2$ and $Y = 3$ differs. Even though the structure of the distributions are similar, **H** und $\mathbf{S}_H$ prefer the one on the right-hand side (see also table 1). Thinking about a slice changing in the frequency of a node, depending for instance on the drawn sample, the induction of a decision tree will lead by using $\mathbf{S}_H$ or **H** to two different classification rules – a situation that looks like the problem of overfitting the tree. In contrast to this behavior, both distributions of Y are assessed in the same manner when using the modality measure **M**. A variation of non-mode probabilities has no effect to the quantity of **M**. Thus, **M** resp. Goodman and Kruskal's λ leads to an induction of much more stable trees,

which are less dependent on small variations in the probabilities caused by random effects of the drawn sample.

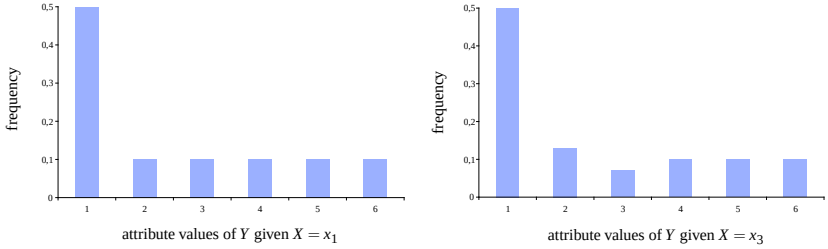

Fig. 3. Two similar distributions of Y given an exogenous attribute X

In summary it may be said that the modality measure has the ability to induct a decision tree which differs from the ones produced by the Gini-index or the information gain measure. And in contrast to the opinion of Breiman et al. (1984) the modality measure seems to have some advantages for which it is worthwhile to analyze this measure in more detail. Considering as well the opportunity to do some inferential statistics the introduction of Goodman and Kruskal's λ as a split criterion is a chance to get an algorithm for the induction of a decision tree which might generate trees and classification rules which are, probably, much more adequate than the known ones.

	$Y\|X = x_1$	$Y\|X = x_2$	$Y\|X = x_3$
Modality measure **M**	0.50	0.60	0.50
Herfindahl's $\mathbf{S}_H$	0.70	0.67	0.69
Entropy **H**	1.49	1.33	1.48

Table 1. Deviation measures of Y given $X = x_1$, $X = x_2$ resp. $X = x_3$

4 The procedure ASCAID

Based upon these remarks about the predictive association measure λ, a new algorithm for the induction of a decision tree should be proposed in the following. This algorithm takes care of the asymmetric character of the classification task, in contrast to the well-known CHAID algorithm, and nevertheless allows to do some inferential statistics, in contrast to the other asymmetric versions of decision tree algorithms like ID3 or CART. And according to the first algorithm for the induction of classification rules, the AID procedure by Sonquist et al. (1971), this procedure should be called **ASCAID**: Using an **AS**ymmetric **C**orrelation Measure for **A**utomatic **I**nteraction **D**etection.

Like all the other induction procedures, ASCAID is a top-down approach as well and consists of the following steps:

- STARTING: All objects are assigned to the root node.
- MERGING and SPLITTING: Using the split criterion λ the set of objects at a node will be divided in further, but not necessarily all subsets which will be assigned to further nodes as well.
- STOPPING: The new generated nodes will be treated in the same manner until nodes will be created for which at least one of some well-defined stop criteria is true. These nodes will be called leaves.
- LABELING: Each leaf will be labeled with the corresponding mode of the membership attribute in the leaf.

To prevent the problem of overfitting the well-known techniques of pre- and post-pruning, which have to be adopted to λ, will be used, too.

The Merging Step

The aim of this optional stage is to merge the categories of the exogenous attributes to prevent a partitioning with too many subsets and/or to prevent the creation of subsets which are too similar. Both approaches lead to a decision tree which is much more suitable to be used for other data sets (less overfitting). The procedure itself is similar to the one of the CHAID procedure and consists of the following steps:

- Instead of the χ^2 measure to analyze the correlation of Y given $X = x_i$ and Y given $X = x_j$, Goodman and Kruskal's λ should be used.

- The decision for the fusion of two categories of an attribute will be made by using the PRE coefficient λ and a well-drawn threshold λ^*. The assessment by using the significance level of the sample quantity of λ, however, is not directly possible. This is due to the degeneration of the distribution of $\hat{\lambda}$ (Goodman, Kruskal (1963)). Here, some further work is necessary, like for instance in the manner of solving the same problem for the coefficient of determination.

The Splitting Step

Based upon the suitably merged categories of all exogenous attributes X_i, now the assessment of these attributes with respect to their discrimination power for the membership attribute Y should be analyzed. Therefore, the following steps are necessary:

- The coefficient $\hat{\lambda}$ should be calculated for all possible pairs (X_i, Y), i.e. $\hat{\lambda}(X_i \to Y) := \hat{\lambda}(i)$.
- All exogenous attribute X_i with a value $\hat{\lambda}(i) > \lambda_{\min}$, which have to be well-drawn as well, will be chosen as a potential split criterion.
- If there is more than one potential split criterion, three different procedures are possible.
  - If there is only one attribute X_{i^*} with a large value $\hat{\lambda}(i)$ (relative to the others), then that attribute will be chosen as split criterion.
  - If there are two such attributes, a test with the null hypothesis $H_0 : \lambda(i) = \lambda(j)$ (or $<$ resp. $>$ as well) will be analyzed. The corresponding test statistic T is asymptotically normal distributed and defined as follows (Goodman and Kruskal (1963)):

$$T := \frac{\hat{\lambda}(i) - \hat{\lambda}(j)}{\sqrt{Var(\hat{\lambda}(i)) + Var(\hat{\lambda}(j))}} \tag{10}$$

    If then the adequate null hypothesis can not be rejected, there are strong hints for the overfitting of the tree. Another sample, drawn from the same population, could lead to another tree, the induction will be very unstable. In such a case the induction of the tree, however, should be continued, for instance by choosing the attribute with the largest sample value $\hat{\lambda}$ as split criterion, but handled with care.
  - If there are more than two such potential split attributes, say $X_1, \ldots, X_k$, an adopted test for the null hypothesis $H_0 : \lambda(i_1) = \ldots = \lambda(i_k)$ should be analyzed. The cor-

responding test statistic T is now asymptotically χ^2 distributed with $k-1$ degrees of freedom and defined as follows (Goodman and Kruskal (1963)):

$$T := \sum_{i=1}^{k} \frac{\left(\hat{\lambda}(i) - \overline{\lambda}\right)^2}{Var(\hat{\lambda}(i))} \qquad \text{with} \qquad \overline{\lambda} := \frac{\sum_{i=1}^{k} \frac{\hat{\lambda}(i)}{Var(\hat{\lambda}(i))}}{\sum_{i=1}^{k} \frac{1}{Var(\hat{\lambda}(i))}} \tag{11}$$

For the treatment of the results of this test, the same rules as for the two attribute case above are valid as well.

The Stopping Step

Based upon the best chosen split attribute, a new partition of the sample (of a node) follows. These subsets have to assess if they have to be treated as well or if they become to an end-node or leaf. This decision is made by using some well-known pre-pruning techniques and should help to generate a decision tree which is able to explain the correlation between the involved attributes as good as for the considered sample. The dividing of a node will terminate if for instance at least one of the following, most important rules is valid:

- All objects of a node belong to the same class of the membership attribute.
- All objects of a node have the same values of the considered exogenous attributes.
- The number of objects at a node is smaller than a pre-defined threshold.

If at least one of these rules is valid for a node, then this node becomes a leaf. If all nodes are leaves, the algorithm stops. The fully developed tree should then be treated with some post-pruning techniques to ensure the adaptability to other samples.

The Accuracy of the Tree

To measure the accuracy of the inducted tree, once again the PRE measure λ can be used. Defining the leaves of the tree as values of a dummy attribute X_0, the predictive association $\hat{\lambda}(X_0 \rightarrow Y)$ of that X_0 and the membership attribute Y can be calculated. Additionally, all tests with respect to λ are possible and useful. Thus, different inducted decision trees can be compared as well by using the known test statistics (10) and (11). Furthermore, the value of $\hat{\lambda}$ can be interpreted as (estimated) missclassification rate of the decision tree for the identification of the class membership attribute Y.

5 Outlook

With ASCAID, a decision tree algorithm is proposed which takes care of the asymmetric character of the classification task, allows to do some inferential statistics and is easy to interpret. Furthermore, the accuracy of the whole tree is based on the same measure as the splits of the nodes. Besides these very promising but also very theoretical features of the algorithm some intensive empirical studies are necessary to show what ASCAID is really able to perform. Furthermore, comparisons with the other well-known algorithms are needful as well as studies about the thresholds in the merging and splitting steps.

But also some theoretical work has to be done. Here, the problem of the degeneration of the distribution of λ has to be mentioned as well as the consideration of missclassification costs as well as the consideration of an ordinal endogenous attributes resp. membership attribute. Nevertheless, the new algorithm could offer a chance to solve the classification tasks maybe better than the known ones.

References

BREIMAN, L.; FRIEDMAN, J.H.; Olshen, R.A., and STONE, C.J. (1984): *Classification and Regression Tree*. Statistics/Probability Series. Wadsworth, Belmont.

DECKER, R. and TEMME, T. (2001): CHAID als Instrument der Werbemittelgestaltung und Zielgruppenbestimmung im Marketing. In: H. Hippner, U. Küsters, M. Meyer, and K. Wilde (Eds.): *Handbuch Data Mining im Marketing*. Vieweg, Braunschweig, 671–683.

GOODMAN, L.A. and Kruskal, W.H. (1954): Measures of Association for Cross Classifications. *Journal of the American Statistical Association, Vol. 49, 732 – 764.*

GOODMAN, L.A. and Kruskal, W.H. (1963): Measures of Association for Cross Classifications, III: Approximate Sampling Theory. *Journal of the American Statistical Association, Vol. 58, 310–364.*

GUTTMAN, L. (1941): An Outline of the Statistical Theory of Prediction. In: Horst, P. (Ed.): *The Prediction of Personal Adjustment*. Bulletin 48, Social Science Research Council, New York.

HILBERT, A. (1998): Zur Theorie der Korrelationsmaße. Eul Verlag, Lohmar, Köln.

KAAS, G.V. (1980): An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics, 29, No. 2, 119–127.*

MESSENGER, R.C. and MANDELL, L.M. (1972): A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Society, 67, 768–772*.

MURTHY, S.K. (1998): Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery, 2, 345–389.*

QUINLAN, J.R. (1986): Induction of Decision Trees. *Machine Learning, 1, 81–106*.

QUINLAN, J.R. (1993): *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California.

SHANNON, C.E. (1948): The Mathematical Theory of Communication. *The Bell Systems Technical Journal, Vol. 27, 379–423*.

SONQUIST, J.A.; BAKER, E.L. and MORGAN, J.N. (1971): Searching for Structure. Institute for Social Research, University of Michigan, Ann Arbor, MI.

ZHOU, X. and DILLON, T.S. (1991): A Statistical-Heuristic Feature Criterion for Decision Tree Induction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, 834–841*.