

**Attribute Selection for Decision Tree Induction:
 An Alternative Formulation**

Decision trees, or more generally class-probability trees, are constructed by a greedy, divide-and-conquer algorithm, which at each step has the goal of selecting from a set of attributes the one that best discriminates a set of examples according to the classification. There is choice, and no consensus in the literature, about the criterion for deciding which attribute to select, and a possible experiment is concerned with comparing five such criteria, which we now discuss.

Suppose we have a problem with k classes C_1, \dots, C_k , and we wish to quantify the discriminatory utility of an attribute A having m distinct values a_1, \dots, a_m . We can show the cross-classification of class and attribute values in a contingency table of the form [10]:

| | | C_1 | C_2 | \dots | C_k |
|--|-----------------------|-----------------------|-----------------------|---------|-----------------------|
| $N = \sum_{i=1}^m \sum_{j=1}^k n_{ij}$ | | $\sum_{i=1}^m n_{i1}$ | $\sum_{i=1}^m n_{i2}$ | \dots | $\sum_{i=1}^m n_{ik}$ |
| a_1 | $\sum_{j=1}^k n_{1j}$ | n_{11} | n_{12} | \dots | n_{1k} |
| a_2 | $\sum_{j=1}^k n_{2j}$ | n_{21} | n_{22} | \dots | n_{2k} |
| \vdots | \vdots | \vdots | \vdots | \dots | \vdots |
| a_m | $\sum_{j=1}^k n_{mj}$ | n_{m1} | n_{m2} | \dots | n_{mk} |

The following proportions can then be defined:

$$p_{ij} = n_{ij}/N,$$

$$p_i = \left(\sum_{j=1}^k n_{ij}\right)/N,$$

$$p_j = \left(\sum_{i=1}^m n_{ij}\right)/N$$

We can define the information content [9] of each cell, class, and attribute as:

$$H_{cell} = - \sum_{i=1}^m \sum_{j=1}^k p_{ij} \log_2 p_{ij},$$

$$H_{class} = - \sum_{j=1}^k p_j \log_2 p_j,$$

$$H_{attribute} = - \sum_{i=1}^m p_i \log_2 p_i$$

From these, information gain is defined as: $Gain = H_{class} + H_{attribute} - H_{cell}$. Information gain means information about class membership which is conveyed by attribute value.

Information gain is used to determine which attribute to employ at a given stage during tree induction, but gain is far from the only such splitting criteria to be found in the literature.

Among the alternatives are:

1. The information gain ratio [3, 7, 8]: $Gainratio = Gain/H_{attribute}$.

2. Distance measure $1 - d_N$ [4]: $1 - d_N = Gain/H_{cell}$
3. The p value obtained from the G statistic [5, 6]: $G\text{-statistic} = 2 * N * Gain * \log_e 2$. Just as with the more familiar χ^2 statistic, this p value gives the probability that an observed co-occurrence between attribute value and classification occurred purely by chance.
4. The p value obtained from the classical χ^2 statistic, where

$$\chi^2 = \sum_i \sum_j \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

Here, $O_{ij} = n_{ij}$ in the table, and $E_{ij} = p_i p_j N$.

Any of these alternative measures can be used as a splitting criterion in tree and rule induction, with varying effects on the form and content of the resulting models.

As an example, here is the contingency table breakdown for the attribute “squamous metaplasia” in a database of 652 experiments [1]. (We use this attribute as an illustrative example only, not because of any special significance.)

| | POSITIVE | NEGATIVE |
|----------------------------------|----------|----------|
| none: | 258 | 374 |
| urinary bladder: | 4 | 0 |
| nasal cavity: | 5 | 3 |
| larynx & nasal cavity: | 0 | 2 |
| larynx & nasal cavity & trachea: | 0 | 1 |
| salivary gland: | 1 | 3 |
| prostate: | 1 | 0 |

The class and attribute sums are computed:

| | | |
|-----|-----|-----|
| 652 | 269 | 383 |
| 632 | 258 | 374 |
| 4 | 4 | 0 |
| 8 | 5 | 3 |
| 2 | 0 | 2 |
| 1 | 0 | 1 |
| 4 | 1 | 3 |
| 1 | 1 | 0 |

The proportions are therefore:

| | | |
|-------|-------|-------|
| 1.000 | 0.413 | 0.587 |
| 0.969 | 0.396 | 0.574 |
| 0.006 | 0.006 | 0.000 |
| 0.012 | 0.008 | 0.005 |
| 0.003 | 0.000 | 0.003 |
| 0.002 | 0.000 | 0.002 |
| 0.006 | 0.002 | 0.005 |
| 0.002 | 0.002 | 0.000 |

and the expected values for χ^2 purposes are:

| | | |
|---------|---------|---------|
| 652.000 | 269.000 | 383.000 |
| 632.000 | 260.748 | 371.252 |
| 4.000 | 1.650 | 2.350 |
| 8.000 | 3.301 | 4.699 |
| 2.000 | 0.825 | 1.175 |
| 1.000 | 0.413 | 0.587 |
| 4.000 | 1.650 | 2.350 |
| 1.000 | 0.413 | 0.587 |

Finally, the various splitting criteria can be compared as follows:

| Gain | Gain Ratio | 1-Dn | G stat p-value | χ^2 p-value |
|-------|------------|-------|----------------|------------------|
| 0.016 | 0.058 | 0.013 | 0.050 | 0.100 |

To assess whether to install the attribute “squamous metaplasia” in a tree or ruleset model, the value of the splitting criteria we had chosen to use would then be compared to that value for all other unused attributes [2]. Attributes are selected to minimize the p values; the other criteria are maximized.

References

- [1] Bahler, D. and D. W. Bristol 1993. A Quantitative Comparison of the Utility of Characteristics for Predicting Chemical Carcinogenesis. *4th Annual Keck Symposium on Computational Biology*, Pittsburgh.
- [2] Bahler, D. and D. W. Bristol 1993. The Induction of Rules for Predicting Chemical Carcinogenesis in Rodents. In L. Hunter, J. Shavlik, and D. Searls (eds.), *Intelligent Systems for Molecular Biology*, Cambridge, MA: AAAI/MIT Press.
- [3] Bahler, D. 1992. Methods of Decision Tree Induction. *4th North Carolina Symposium on Art. Intell. and Advanced Computing Tech.*, Raleigh.
- [4] Lopez de Mantaras, R. 1991. A Distance-Based Attribute Selection Measure for Decision Tree Induction. *Machine Learning 6*, 1991, 81-92.
- [5] Mingers, J. 1989. An Empirical Comparison of Selection Measures for Decision-Tree Induction. *Machine Learning 3*: 319-342.
- [6] Mingers, J. 1987. Expert Systems – Rule Induction with Statistical Data. *Journal of the Operations Research Society 38*, 39-47.
- [7] Quinlan, J.R. 1986. Induction of Decision Trees. *Machine Learning 1*: 81-106.
- [8] Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- [9] Shannon, C. E. 1949. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.
- [10] White, A. and W.Z. Liu 1994. Bias in Information-Based Measures in Decision Tree Induction. *Machine Learning 15*, 321-329.