Attributauswahlmaße für die Induktion von Entscheidungsbäumen: Ein Überblick

Christian Borgelt and Rudolf Kruse

Institut für Informations- und Kommunikationssysteme
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, D-39106 Magdeburg

e-mail: borgelt@iik.cs.uni-magdeburg.de

Zusammenfassung Die Induktion von Entscheidungsbäumen mit Hilfe eines Top-Down-Verfahrens ist eine bekannte und weit verbreitete Technik zur Bestimmung von Klassifikatoren. Der Erfolg dieser Methode hängt stark von dem Auswahlmaß ab, mit dem beim Aufbau des Entscheidungsbaums das nächste zu testende Attribut bestimmt wird. In diesem Aufsatz geben wir einen Überblick über eine Reihe von Auswahlmaßen, die in der Vergangenheit für die Induktion von Entscheidungsbäumen vorgeschlagen wurden. Wir erläutern die den Maßen zugrundeliegenden Ideen und vergleichen die betrachteten Maße anhand experimenteller Ergebnisse.

1 Einleitung

Entscheidungsbäume sind eine sehr bekannte Form von Klassifikatoren. Klassifikatoren wiederum sind Programme, die einen Fall oder ein Objekt automatisch klassifizieren, d.h. ihn bzw. es anhand seiner Merkmale einer von mehreren vorgegebenen Klassen zuordnen. Wenn es sich z.B. bei den Fällen um Patienten handelt, sind die Merkmale Angaben über die Eigenschaften der Patienten (Geschlecht, Alter etc.) und ihre Symptome (Fieber, hoher Blutdruck etc.), die Klassen etwa Krankheiten oder zu verabreichende Medikamente.

Entscheidungsbäume sind Klassifikatoren, die — wie der Name schon sagt — eine baumartige Struktur haben. Jedem Blatt ist eine Klasse zugeordnet, jedem inneren Knoten ein Attribut (oder Merkmal), wobei es mehrere Blätter zu einer Klasse und mehrere innere Knoten zu einem Attribut geben kann. Die Nachfolger der inneren Knoten werden über Kanten erreicht, denen jeweils ein Wert des zu dem Knoten gehörenden Attributes zugeordnet ist. Jeder Blattknoten stellt eine Entscheidung „Der betrachtete Fall gehört zur Klasse c.“ dar, wobei c die dem Blatt zugeordnete Klasse ist. Jeder innere Knoten entspricht einer Anweisung „Teste Attribut A und folge der Kante, der der festgestellte Wert zugeordnet ist.“, wobei A das dem Knoten zugeordnete Attribut ist. Die Klassifikation eines Falles mit einem Entscheidungsbaum wird so vorgenommen, daß man an der Wurzel startet und die Anweisungen in den jeweils erreichten inneren Knoten ausführt, bis der Fall durch einen Blattknoten klassifiziert wird.

Nun sind Entscheidungsbäume zwar sehr leicht anzuwenden, deutlich schwieriger aber ist es, sie „per Hand“ zu erzeugen. Insbesondere wenn die Zahl der

möglichen Testattribute groß und das Wissen um die Zusammenhänge vage ist, kann die Konstruktion eines Entscheidungsbaumes aufwendig und langwierig sein. Liegt jedoch eine Datenbank bereits klassifizierter Fälle vor, kann man eine automatische Erzeugung versuchen [Breiman et al. 1984, Quinlan 1986, Quinlan 1993]. Dies geschieht gewöhnlich mit Hilfe eines Top-Down-Verfahrens, das nach dem Prinzip „teile und herrsche“ (divide and conquer) arbeitet und außerdem Attribute „gierig“ (greedy) nach dem Wert auswählt, der ihnen von einem Auswahlmaß zugeschrieben wird. Man spricht auch von „top down induction of decision trees“ (TDIDT). In Abschnitt 2 wird dieses Verfahren anhand eines einfachen Beispiels erläutert.

Der Erfolg eines solchen Induktionsverfahrens hängt stark von dem Auswahlmaß ab, mit dem beim Aufbau des Entscheidungsbaums das nächste zu testende Attribut bestimmt wird. In Abschnitt 3 geben wir einen Überblick über eine ganze Reihe von Attributauswahlmaßen, die auf zum Teil sehr unterschiedlichen Ideen beruhen. Da jedoch am Ende nicht die theoretische Begründung sondern der praktische Erfolg ausschlaggebend ist, vergleichen wir die betrachteten Maße in Abschnitt 4 anhand einiger experimenteller Ergebnisse auf Datensätzen aus dem UC Irvine Machine Learning Repository [Murphy und Aha 1994]. Es zeigt sich, daß eine Rangfolge der Maße nach ihrer Qualität nicht angegeben werden kann. Daher kann es sich lohnen, jedes der beschriebenen Maße auszuprobieren.

2 Induktion von Entscheidungsbäumen

Die Induktion von Entscheidungsbäumen ist zwar eine sehr bekannte Technik zur Erzeugung von Klassifikatoren, dennoch mag es Leser geben, die mit dieser Methode noch nicht vertraut sind. In diesem Abschnitt geben wir daher anhand eines einfachen Beispiels eine kurze Einführung. Leser, die das Prinzip der Induktion von Entscheidungsbäumen bereits kennen, können diesen Abschnitt (ggf. bis auf einen kurzen Blick in die Beschreibung der in diesem Aufsatz verwendeten Notation in Abschnitt 2.2) überspringen und mit Abschnitt 3 fortfahren.

Das Prinzip der Induktion von Entscheidungsbäumen ist, wie bereits in der Einleitung erwähnt, ein „teile und herrsche“ Verfahren (divide and conquer) mit „gieriger“ (greedy) Auswahl der zu testenden Attribute: In einer gegebenen Menge von klassifizierten Fallbeschreibungen werden die bedingten Häufigkeitsverteilungen der Klassen unter den einzelnen zur Beschreibung verwendeten Attributen bestimmt und mit Hilfe eines Auswahlmaßes bewertet. Dasjenige Attribut, das die beste Bewertung erhält, wird als Testattribut ausgewählt. Dies ist der „gierige“ Teil des Algorithmus. Anschließend werden die Fallbeschreibungen gemäß der verschiedenen Werte des Testattributes aufgeteilt, und das Verfahren wird rekursiv auf die sich ergebenden Teilmengen angewandt. Dies ist der „teile und herrsche“ Teil des Algorithmus. Die Rekursion bricht ab, wenn entweder alle Fälle einer Teilmenge zu der gleichen Klasse gehören, wenn kein Attribut zu einer Verbesserung der Klassifikation führt, oder keine weiteren Attribute für einen Test zur Verfügung stehen.[1]

[1] In vielen Entscheidungsbaum-Lernprogrammen wird der konstruierte Baum an-

Im folgenden illustrieren wir das geschilderte Vorgehen anhand eines einfachen Beispiels, erläutern dann die in diesem Aufsatz verwendete Notation und geben den Induktionsalgorithmus im Pseudocode an.

2.1 Ein einfaches Beispiel

Tabelle 1 zeigt Merkmale von zwölf Patienten — das Geschlecht, das Alter und eine qualitative Angabe des Blutdrucks — zusammen mit einem Medikament, das bei diesen Patienten bei der Behandlung einer nicht näher spezifizierten Krankheit wirksam war. Ohne Betrachtung der Patientenmerkmale läßt sich offenbar das richtige Medikament nur mit einer Erfolgswahrscheinlichkeit von 50% bestimmen, da sowohl Medikament A als auch Medikament B in sechs Fällen wirksam war. Da eine solche Situation für zukünftige Behandlungen nicht gerade günstig ist, soll ein Entscheidungsbaum gefunden werden, der es erlaubt, das wirksame Medikament aus den Patientenmerkmalen abzuleiten.

Zu diesem Zweck betrachten wir alle bedingten Verteilungen der wirksamen Medikamente unter den zur Verfügung stehenden Merkmalen. Diese bedingten Verteilungen sind in Tabelle 2 dargestellt. Es zeigt sich, daß das Geschlecht des Patienten offenbar keinen Einfluß hat, da sowohl für männliche als auch für weibliche Patienten Medikament A wie Medikament B in der Hälfte der Fälle wirksam war. Man kann folglich (wie ohne Berücksichtigung von Patientenmerkmalen) nur mit 50% Erfolgswahrscheinlichkeit das wirksame Medikament bestimmen. Bessere Ergebnisse erbringt die Betrachtung des Alters des Patienten. Unter vierzig Jahren war Medikament A in vier von sechs Fällen, über vierzig Jahren Medikament B in vier von sechs Fällen wirksam. Die Erfolgsquote liegt daher bei 67%. Am besten läßt sich jedoch mit Hilfe des Blutdrucks das wirksame Medikament bestimmen. Ist er hoch, war es Medikament A, ist er niedrig, war es Medikament B. Nur bei normalem Blutdruck kann die Bestimmung nicht verbessert werden. Insgesamt liegt die Erfolgsquote bei 75%.

Da der Blutdruck die beste Bestimmung des wirksamen Medikamentes erlaubt, wird er als erstes zu testendes Attribut ausgewählt. Die Fälle der Tabelle werden gemäß der Werte, die sie für dieses Attribut aufweisen, unterteilt. Da das wirksame Medikament für Fälle mit hohem oder niedrigem Blutdruck eindeutig ist, brauchen diese Fälle nicht weiter betrachtet zu werden. Für die Fälle, in denen normaler Blutdruck vorlag, prüfen wir erneut die bedingte Verteilung des wirksamen Medikamentes bezüglich des Alters des Patienten. Diese Verteilung ist in Tabelle 3 gezeigt. Da eine Unterteilung bei einem Alter von vierzig Jahren die Fälle in denen Medikament A wirksam war, von jenen, in denen Medikament B wirksam war, trennt, ist ein Verfahren zur sicheren Bestimmung des wirksamen Medikamentes gefunden. Der zugehörige Entscheidungsbaum, dessen Struktur sich direkt aus Tabelle 3 ablesen läßt, ist in Abbildung 1 gezeigt.

---

schließend gestutzt („Pruning“), d.h. einige Entscheidungsknoten, die nur geringen Anteil an der Klassifikationsgüte haben, werden wieder entfernt. Dies geschieht, um den Klassifikator zu vereinfachen und um eine Überanpassung an die zufälligen Besonderheiten der Lerndaten („Overfitting“) zu vermeiden. Auf diese zusätzliche Phase gehen wir jedoch nicht ein.

No	Geschlecht	Alter	Blutdruck	Medikament
1	männlich	20	normal	A
2	weiblich	73	normal	B
3	weiblich	37	hoch	A
4	männlich	33	niedrig	B
5	weiblich	48	hoch	A
6	männlich	29	normal	A
7	weiblich	52	normal	B
8	männlich	42	niedrig	B
9	männlich	61	normal	B
10	weiblich	30	normal	A
11	weiblich	26	niedrig	B
12	männlich	54	hoch	A

Tabelle 1. Diese Tabelle zeigt Patientendaten zusammen mit einem wirksamen Medikament (wirksam in bezug auf eine nicht näher spezifizierte Krankheit). Um einen Entscheidungsbaum zu finden, der das wirksame Medikament aus den Patientendaten bestimmt, werden die bedingten Verteilungen des wirksamen Medikamentes unter den verschiedenen Patientenmerkmalen untersucht (siehe dazu die Tabellen 2 und 3).

No	Geschlecht	Med.
1	männlich	A
6	männlich	A
12	männlich	A
4	männlich	B
8	männlich	B
9	männlich	B
3	weiblich	A
5	weiblich	A
10	weiblich	A
2	weiblich	B
7	weiblich	B
11	weiblich	B

No	Alter	Med.
1	20	A
11	26	B
6	29	A
10	30	A
4	33	B
3	37	A
8	42	B
5	48	A
7	52	B
12	54	A
9	61	B
2	73	B

No	Blutdruck	Med.
3	hoch	A
5	hoch	A
12	hoch	A
1	normal	A
6	normal	A
10	normal	A
2	normal	B
7	normal	B
9	normal	B
4	niedrig	B
8	niedrig	B
11	niedrig	B

Tabelle 2. Die bedingten Verteilungen des wirksamen Medikamentes gegeben das Geschlecht (links), das Alter (mitte, mit einer Einteilung in „bis vierzig Jahre“ und „über vierzig Jahre“) und den Blutdruck (rechts) des Patienten.

No	Blutdruck	Alter	Medikament
3	hoch	37	A
5	hoch	48	A
12	hoch	54	A
1	normal	20	A
6	normal	29	A
10	normal	30	A
7	normal	52	B
9	normal	61	B
2	normal	73	B
11	niedrig	26	B
4	niedrig	33	B
8	niedrig	42	B

Tabelle 3. Das Hinzufügen des Alters in den Fällen, in denen der Blutdruck normal ist, führt zu einer perfekten Bestimmung des wirksamen Medikamentes. Der zugehörige Entscheidungsbaum, der sich direkt aus dieser Tabelle ablesen läßt, ist in Abbildung 1 dargestellt.

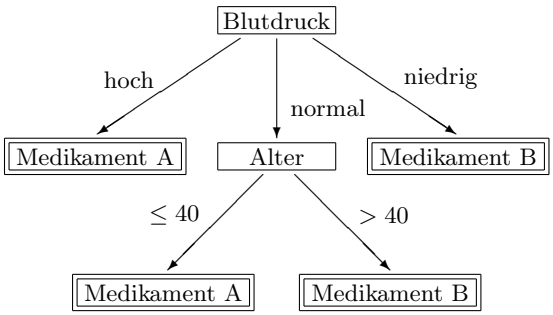

Abbildung 1. Der gefundene Entscheidungsbaum. Zuerst wird das Attribut Blutdruck getestet. Ist der Wert hoch oder niedrig, kann das wirksame Medikament sofort angegeben werden. Ist der Blutdruck normal, wird das Alter des Patienten geprüft, wodurch auch in diesem Fall das wirksame Medikament bestimmt werden kann.

2.2 Notation

Die in diesem Aufsatz verwendete Notation ist in der folgenden Liste erläutert. Es wird eine Menge S von Fallbeschreibungen angenommen, die aus der Angabe einer Klasse $c_i \in \mathrm{dom}(C)$ und m Attributwerten $a_{ij}^{(j)} \in \mathrm{dom}(A^{(j)})$, $j \in \{1, \ldots, m\}$, besteht. Da stets nur Häufigkeitsverteilungen der Klassen unter den Werten eines Attributes betrachtet werden, vernachlässigen wir den Attributindex. Die zur Berechnung verwendeten (relativen) Häufigkeiten lassen sich dann leicht durch zwei Indizes kennzeichnen. Wir verwenden stets den Index i für Klassen und den Index j für Attributwerte. Es bedeuten:

S	eine Menge von Fall- bzw. Objektbeschreibungen
C	das Klassenattribut
$A^{(1)}, \ldots, A^{(m)}$	andere Attribute (Index im folgenden weggelassen)
$\mathrm{dom}(C)$	$= \{c_1, \ldots, c_{n_C}\}$, n_C: Anzahl Klassen
$\mathrm{dom}(A)$	$= \{a_1, \ldots, a_{n_A}\}$, n_A: Anzahl Attributwerte
$N_{\cdot\cdot}$	Gesamtzahl Fall- bzw. Objektbeschreibungen, d.h. $N_{\cdot\cdot} = \lvert S \rvert$
$N_{i\cdot}$	absolute Häufigkeit der Klasse c_i
$N_{\cdot j}$	absolute Häufigkeit des Attributwertes a_j
N_{ij}	absolute Häufigkeit der Kombination der Klasse c_i und des Attributwertes a_j. Es ist $N_{i\cdot} = \sum_{j=1}^{n_A} N_{ij}$ und $N_{\cdot j} = \sum_{i=1}^{n_C} N_{ij}$.
$p_{i\cdot}$	relative Häufigkeit der Klasse c_i, $p_{i\cdot} = \frac{N_{i\cdot}}{N_{\cdot\cdot}}$
$p_{\cdot j}$	relative Häufigkeit des Attributwertes a_j, $p_{\cdot j} = \frac{N_{\cdot j}}{N_{\cdot\cdot}}$
p_{ij}	relative Häufigkeit der Kombination der Klasse c_i und des Attributwertes a_j, $p_{ij} = \frac{N_{ij}}{N_{\cdot\cdot}}$
$p_{i\lvert j}$	relative Häufigkeit der Klasse c_i in den Fällen mit dem Attributwert a_j, $p_{i\lvert j} = \frac{N_{ij}}{N_{\cdot j}} = \frac{p_{ij}}{p_{\cdot j}}$
$p_{j\lvert i}$	relative Häufigkeit des Attributwertes a_j in den Fällen mit der Klasse c_i, $p_{j\lvert i} = \frac{N_{ij}}{N_{i\cdot}} = \frac{p_{ij}}{p_{i\cdot}}$

```
function grow_tree (S : set of cases) : node;
begin
    best_v := WORTHLESS;
    for all untested attributes A do
        compute frequencies N_ij, N_i., N_.j
            for 1 ≤ i ≤ n_C and 1 ≤ j ≤ n_A;
        compute value v of a selection measure
            using N_ij, N_i., N_.j;
        if v > best_v
        then best_v := v;
            best_A := A;
        end;
    end
    if best_v = WORTHLESS
    then create leaf node n;
        assign majority class of S to n;
    else create test node n;
        assign test on attribute best_A to n;
        for all a ∈ dom(best_A) do
            n.child[a] := grow_tree(S|_{best_A=a});
        end;
    end;
    return n;
end;   /* grow_tree() */
```

Abbildung 2. Der Induktionsalgorithmus. Der einfacheren Darstellung wegen wird angenommen, daß alle Attribute eine endliche Anzahl symbolischer Werte haben. Ganzzahlige und reellwertige Attribute werden behandelt, indem alle möglichen Trennwerte gebildet werden. Für jeden Trennwert wird das (künstliche) symbolische Attribut mit den Werten „größer als Trennwert“ und „kleiner-gleich Trennwert“ untersucht. Im Unterschied zu symbolischen Attributen gilt ein ganzzahliges oder reellwertiges Attribut weiter als ungetestet, wenn ein Testknoten zu diesem Attribut erzeugt wurde, da der Wertebereich in einem späteren Schritt (anders als bei symbolischen Attributen) weiter unterteilt werden kann.

2.3 Der Induktionsalgorithmus

Das allgemeine Schema des Entscheidungsbaum-Induktionsalgorithmus ist in Abbildung 2 in Pascal-ähnlicher Schreibweise dargestellt. Im ersten Teil des Algorithmus wird für jedes Attribut die gemeinsame Häufigkeitsverteilung seiner Werte und der Klassen bestimmt und aus dieser der Wert eines Attributauswahlmaßes berechnet. Das am besten bewertete Attribut wird in der Variablen *best_A* abgelegt. Im zweiten Teil wird abhängig vom Ergebnis des ersten Teils entweder ein Blatt- oder ein Entscheidungsknoten angelegt. Im letzteren Fall werden anschließend die Fallbeschreibungen gemäß der Werte, die sie für das Testattribut aufweisen, unterteilt, und für jede Teilmenge wird die Funktion grow_tree rekursiv wieder ausgeführt.

Zur Vereinfachung der Darstellung wird angenommen, daß alle Attribute eine endliche Anzahl symbolischer Werte haben. Ganzzahlige und reellwertige Attribute können behandelt werden, indem man die auftretenden Werte sortiert und für jedes Paar benachbarter Werte einen zwischen ihnen liegenden Trennwert wählt (z.B. das arithmetische Mittel der beiden Werte). Mit diesem Trennwert wird dann das (künstliche) symbolische Attribut mit den Werten „größer als Trennwert“ und „kleiner-gleich Trennwert“ gebildet. Es wird derjenige Trennwert gewählt, dessen zugehöriges (künstliches) symbolisches Attribut von dem verwendeten Auswahlmaß am besten bewertet wird.

Während des rekursiven Abstiegs werden getestete symbolische Attribute

$$P(x_1) = 0.40, \quad P(x_2) = 0.10, \quad P(x_3) = 0.15, \quad P(x_4) = 0.15, \quad P(x_5) = 0.20$$

Shannonsche Entropie: 2.15 Bit/Symbol

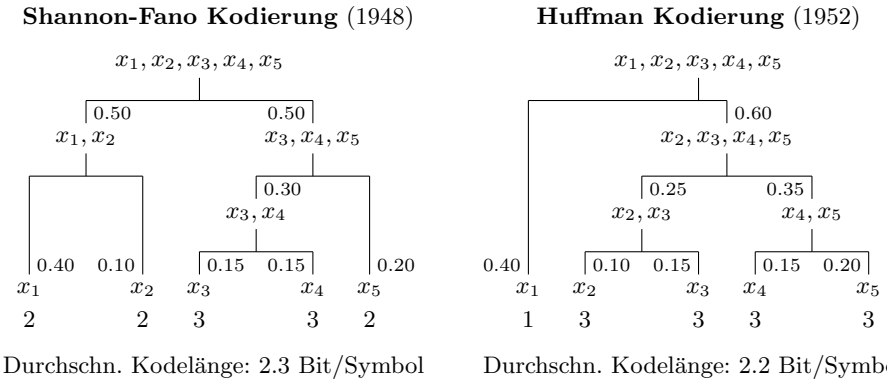

Abbildung 3. Zwei Frage- bzw. Kodierungsschemata für die oben angegebene Wahrscheinlichkeitsverteilung über den fünf Attributwerten x_1, x_2, x_3, x_4 und x_5. Die Zahlen unter den Attributwerten geben die Anzahl Ja/Nein-Fragen an, die man stellen muß, um den Wert mit dem entsprechenden Frageschema zu identifizieren, bzw. die Anzahl Bit, in denen der Wert für eine Übertragung kodiert werden muß.

markiert, da ein erneuter Test dieses Attributes offenbar sinnlos ist, denn die Unterteilung der Fallbeschreibungen bewirkt ja, daß in der nächsten Rekursionsstufe alle Fälle den gleichen Wert für dieses Attribut haben. Ganzzahlige und reellwertige Attribute werden dagegen nicht markiert, da in der nächsten Rekursionsstufe ein anderer Trennwert gewählt und so der Wertebereich weiter unterteilt werden kann.

3 Auswahlmaße

Wie schon in der Einleitung erwähnt, hängt der Erfolg des Entscheidungsbaumlernens stark von dem verwendeten Attributauswahlmaß ab. In diesem Abschnitt geben wir einen Überblick über eine Reihe von Auswahlmaßen und erläutern die ihnen zugrundeliegenden Ideen.

3.1 Informationsgewinn

Das wohl bekannteste Auswahlmaß ist der sogenannte *Informationsgewinn* (information gain) [Quinlan 1986, Quinlan 1993]. Er mißt, wieviel Information man durch das Feststellen des Wertes des Testattributes über die Klasse gewinnt. Der hierbei verwendete Informationsbegriff basiert auf der von [Shannon 1948] definierten *Entropie* H einer Wahrscheinlichkeitsverteilung, $H = -\sum_{i=1}^{n} p_i \log_2 p_i$. Der Informationsgewinn ist nichts anderes als die Entropieverminderung beim

Übergang zur bedingten Verteilung und definiert als

$$I_{\text{gain}}(C,A) = H_C - H_{C|A} \;=\; H_C + H_A - H_{CA}$$
$$= -\sum_{i=1}^{n_C} p_{i\cdot} \log_2 p_{i\cdot} - \sum_{j=1}^{n_A} p_{\cdot j} \log_2 p_{\cdot j} + \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} p_{ij} \log_2 p_{ij}.$$

H_C ist die Entropie der Klassenverteilung, H_A die Entropie der Attributwerteverteilung und H_{CA} die Entropie der gemeinsamen Verteilung.

Die diesem Auswahlmaß zugrundeliegende Idee kann man sich am besten klar machen, indem man die Entropie einer Wahrscheinlichkeitsverteilung über einer Menge von Werten interpretiert als die durchschnittliche Anzahl von Ja/Nein-Fragen, die nötig sind, um einen Wert aus dieser Menge zu bestimmen. Dies ist in Abbildung 3 illustriert. Gegeben sei die oben in dieser Abbildung angegebene Wahrscheinlichkeitsverteilung über den Werten $x_1, \ldots, x_5$. Der in einem konkreten Fall vorliegende Wert soll durch Ja/Nein-Fragen ermittelt werden. Ein besseres Verfahren als das Fragen „der Reihe nach“ findet man leicht: Die Werte werden in zwei annähernd gleich große Teilmengen zerlegt und dann wird nach der Zugehörigkeit zu einer der beiden Teilmengen gefragt. Eine der beiden Teilmengen kann so ausgeschlossen werden, die andere wird wieder unterteilt, bis die übrigbleibende Menge nur noch ein Element enthält. Die Anzahl Fragen, die man bei diesem Verfahren braucht, bezeichnet man auch als *Hartley-Information* der Wertemenge [Hartley 1928].[2] Wir werden ihr bei den possibilistischen Maßen wieder begegnen. Hier lassen wir sie jedoch beiseite, da sich das verwendete Frageschema bei ungleicher Häufigkeit der Werte (wie sie auch in unserem Beispiel gegeben ist) noch verbessern läßt.

Ein naheliegender Verbesserungsvorschlag stammt von Shannon und Fano [Shannon 1948]: Teile die Wertemenge nicht in annähernd *gleich große*, sondern in annähernd *gleichwahrscheinliche* Teilmengen (siehe Abbildung 3). Dadurch wird die nötige Anzahl Fragen für häufige Werte verringert und für seltene Werte erhöht. Im Durchschnitt sinkt so die Anzahl der benötigten Fragen (in unserem Beispiel auf 2.30).

Aber auch dieses Verfahren ist noch nicht das bestmögliche. In unserem Beispiel erkennt man dies schon daran, daß auf der zweiten Ebene der linken Seite des Fragebaum die Unterteilung notgedrungen zu zwei sehr ungleich wahrscheinlichen Teilmengen führt. Um dies zu vermeiden, geht man in dem (beweisbar) optimalen Verfahren von [Huffman 1952] umgekehrt vor: Statt die Wertemenge immer weiter zu unterteilen, wird mit einelementigen Mengen begonnen und es werden stets die beiden Mengen zusammengefaßt, die die geringste Wahrscheinlichkeit besitzen (siehe Abbildung 3). Tatsächlich reduziert sich so auch in unserem Beispiel die Zahl der im Durchschnitt notwendigen Fragen auf 2.20.

[2] In unserem Beispiel sind dies $\log_2 5 \approx 2.32$ Fragen. Allerdings gilt dieser Wert nur bei Gleichwahrscheinlichkeit der Symbole. Bei ungleichen Wahrscheinlichkeiten kann er höher sein, in unserem Beispiel etwa kann er bei ungünstiger Einteilung bis auf 2.6 Fragen steigen.

Schon von [Shannon 1948] wurde für den allgemeinen Fall gezeigt, daß die Untergrenze für die Zahl der notwendigen Fragen die Entropie ist. Diese kann zwar nicht in jedem Fall erreicht (das Huffman-Verfahren liefert stets das optimale Frageschema), aber nie unterschritten werden. Der Einfachheit halber rechnet man daher i.a. mit dieser Untergrenze, um sich die Konstruktion des Fragebaums zu sparen.

Für die folgenden Abschnitte, insbesondere den über die auf dem Prinzip der minimalen Beschreibungslänge basierenden Maße, bemerken wir, daß sich die durchschnittliche Frageanzahl auch als Kodelänge deuten läßt. Durch zusätzliche Festlegungen kann man nämlich die Konstruktion des Fragebaums bei gegebener Verteilung eindeutig machen. Als Kode zur Übertragung eines Wertes wählt man dann einfach die Ja/Nein-Antworten, die bei dem dann festliegenden Frageschema zu diesem Wert führen.

Kehren wir nun zum Informationsgewinn zurück. Indem wir die bedingte Entropie

$$H_{C|A} \;=\; -\sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} \frac{p_{ij}}{p_{.j}} \log_2 \frac{p_{ij}}{p_{.j}} \;=\; -\sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} p_{i|j} \log_2 p_{i|j}$$

betrachten, läßt sich mit den obigen Erläuterungen die Idee des Informationsgewinns leicht einsehen. $H_{C|A}$ ist offenbar die zu erwartende durchschnittliche Anzahl von Fragen, die nach Bekanntwerden des Wertes des Attributes A noch nötig sind, um die Klasse zu bestimmen. Indem man diese von der Zahl der Fragen H_C abzieht, die ohne Kenntnis des Wertes des Attributes A notwendig sind, erhält man die Ersparnis in der Anzahl der Fragen: den Informationsgewinn. In der Interpretation als Kodelänge mißt der Informationsgewinn die zu erwartende Verringerung der Nachrichtenlänge beim Übergang von einer unbedingten Kodierung der Klassen zu einer bedingten Kodierung bei bekanntem Attributwert.

Schon sehr früh bemerkte man, daß der Informationsgewinn zu einer Bevorzugung von Attributen mit vielen Werten neigt. Auf die Gründe für diese Verzerrung können wir hier leider nicht näher eingehen. Sie läßt sich im wesentlichen erklären durch die Tatsache, daß bei Teilung eines Attributwertes der Informationsgewinn höchstens zunehmen kann, und durch Quantisierungseffekte, die durch die endiche Anzahl von Fallbeschreibungen hervorgerufen werden.

Zum Ausgleich der Bevorzugung von Attributen mit vielen Werten wurden Normalisierungen des Informationsgewinns vorgeschlagen, etwa das *Informationsgewinnverhältnis* (information gain ratio) [Quinlan 1986, Quinlan 1993]

$$I_{\mathrm{gr}}(C,A) = \frac{I_{\mathrm{gain}}(C,A)}{H_A} = \frac{H_C + H_A - H_{CA}}{H_A},$$

bei dem durch die Entropie der Häufigkeitsverteilung über den Attributwerten geteilt wird, um so so etwas wie den „Gewinn an nutzbarer Information“ zu bestimmen.

Das Informationsgewinnverhältnis ist nicht symmetrisch, d.h. im allgemeinen ist $I_{\mathrm{gr}}(C,A) \neq I_{\mathrm{gr}}(A,C)$, was sich u.U. als nachteilig erweisen kann. Das *symmetrische Informationsgewinnverhältnis* [Lopez de Mantaras 1991]

$$I^{(1)}_{\mathrm{sgr}}(C,A) = \frac{I_{\mathrm{gain}}(C,A)}{H_{CA}} = \frac{H_C + H_A - H_{CA}}{H_{CA}},$$

bei dem durch die Entropie der gemeinsamen Verteilung geteilt wird, besitzt nicht diesen Nachteil. Weiter gibt es, obwohl diese Variante in der Literatur bisher nicht untersucht zu sein scheint, die Möglichkeit, ein symmetrisches Informationsgewinnverhältnis durch eine Division durch die Summe von Attribut- und Klassenentropie zu erhalten:

$$I^{(2)}_{\mathrm{sgr}}(C,A) = \frac{I_{\mathrm{gain}}(C,A)}{H_A + H_C} = \frac{H_C + H_A - H_{CA}}{H_A + H_C}.$$

Der Grund für die Vernachlässigung dieses Maßes liegt wahrscheinlich darin, daß es bei einem reinen Anwenden von Attributauswahlmaßen zu den gleichen Ergebnissen führt wie $I^{(1)}_{\mathrm{sgr}}$. In der Praxis hat man jedoch zu berücksichtigen, daß außer dem Informationsgewinn den ein Attribut einbringt, auch die Wahrscheinlichkeit mit der dieser Gewinn tatsächlich genutzt werden kann, berücksichtigt werden muß. Denn ein Attributwert kann ja auch unbekannt sein und in diesem Fall gewinnt man natürlich keine Information. [Quinlan 1993] schlug vor, dies dadurch zu behandeln, daß man das Auswahlmaß mit dem relativen Anteil der bekannten Werte des betrachteten Attributes gewichtet. Bei einem solchen Vorgehen ergeben sich nun aber Unterschiede zwischen $I^{(1)}_{\mathrm{sgr}}$ und $I^{(2)}_{\mathrm{sgr}}$.

3.2 Gini-Index

Der im vorhergehenden Abschnitt behandelte Informationsgewinn basiert, wie beschrieben, auf der Shannonsche Entropie. Diese kann gesehen werden als ein Spezialfall einer *verallgemeinerten Entropie* [Daróczy 1970]

$$H_{\mathrm{gen}}(\beta) = \sum_{i=1}^{n} p_i \, \frac{2^{\beta-1}}{2^{\beta-1}-1} \, \left(1 - p_i^{\beta-1}\right).$$

Aus dieser ergibt sich die Shannonsche Entropie als

$$H = \lim_{\beta \to 1} H_{\mathrm{gen}}(\beta) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Für das Lernen von Entscheidungsbäumen wird aber auch die sogenannte *quadratische Entropie*

$$H^2 = H_{\mathrm{gen}}(\beta = 2) = 2\sum_{i=1}^{n} p_i(1-p_i) = 2\left(1 - \sum_{i=1}^{n} p_i^2\right)$$

verwendet. Indem man sie in gleicher Weise einsetzt wie die Shannonsche Entropie, gelangt man zum *Gini-Index* [Breiman et al. 1984]

$$\begin{aligned}\mathrm{Gini}(C,A) &= \frac{1}{2}\left(H_C^2 - H_{C|A}^2\right) \\ &= 1 - \sum_{i=1}^{n_C} p_{i.}^2 - \sum_{j=1}^{n_A} p_{.j}\left(1 - \sum_{i=1}^{n_C} p_{i|j}^2\right) \\ &= \sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} p_{i|j}^2 - \sum_{i=1}^{n_C} p_{i.}^2.\end{aligned}$$

Anschaulich kann der Gini-Index gedeutet werden als die zu erwartende Verringerung der Fehlklassifikationswahrscheinlichkeit. Stellen wir uns dazu vor, daß ein vorliegender Fall zufällig klassifiziert werde, und zwar mit der Wahrscheinlichkeit p_i als zur Klasse c_i gehörig. Dies wird offenbar mit Wahrscheinlichkeit $(1 - p_i)$ falsch sein. Folglich gibt $\sum_{i=1}^{n} p_i(1 - p_i)$ die Wahrscheinlichkeit an, daß ein Fall mit diesem Verfahren falsch klassifiziert wird. Indem man nun die zu erwartende Fehlklassifikationswahrscheinlichkeit bei Kenntnis des Wertes des Attributes A abzieht, erhält man, analog zum Informationsgewinn, die Steigerung der Wahrscheinlichkeit einer richtigen Klassifikation.

Auch der Gini-Index läßt sich zur Beseitigung einer Bevorzugung von Attributen mit vielen Werten normalisieren. Wir geben hier nur die von [Zhou und Dillon 1991] vorgeschlagene Form an.

$$\begin{aligned}\mathrm{Gini}_{\mathrm{sym}}(C,A) &= \frac{H_C^2 - H_{C|A}^2 + H_A^2 - H_{A|C}^2}{H_C^2 + H_A^2} \\ &= \frac{\sum_{i=1}^{n_C} p_{i.} \sum_{j=1}^{n_A} p_{j|i}^2 + \sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} p_{i|j}^2 - \sum_{i=1}^{n_C} p_{i.}^2 - \sum_{j=1}^{n_A} p_{.j}^2}{2 - \sum_{i=1}^{n_C} p_{i.}^2 - \sum_{j=1}^{n_A} p_{.j}^2}\end{aligned}$$

Die vom symmetrischen Informationsgewinnverhältnis abweichende Form des Zählers erklärt sich aus der inhärenten Unsymmetrie des Gini-Index. Denn im Gegensatz zum Informationsgewinn, für den immer $I_{\mathrm{gain}}(C,A) = I_{\mathrm{gain}}(A,C)$ gilt, ist im allgemeinen $\mathrm{Gini}(C,A) \neq \mathrm{Gini}(A,C)$.

3.3 Relevanz

Für eine gute Klassifikation erscheint es als günstig, wenn ein Attributwert eindeutig eine Klasse anzeigt. Dies wurde schon aus dem Beispiel in Abschnitt 2.1 deutlich. Ein Attributauswahlmaß, das explizit nach solchen eindeutigen Anzeichen für eine Klassenzugehörigkeit sucht, ist das von [Baim 1988] vorgeschlagene

Relevanzmaß

$$R(C,A) = 1 - \frac{1}{n_C - 1} \sum_{j=1}^{n_A} \sum_{i=1, i \neq i_{\max}(j)}^{n_C} \frac{p_{ij}}{p_{i.}}$$
$$= 1 - \frac{1}{n_C - 1} \sum_{j=1}^{n_A} p_{.j} \left(\sum_{i=1}^{n_C} \frac{p_{i|j}}{p_{i.}} - \max_i \frac{p_{i|j}}{p_{i.}} \right),$$

wobei $c_{i_{\max}(j)}$ die Mehrheitsklasse in den Fällen mit dem Attributwert a_j ist. Daß $R(C,A)$ nach eindeutigen Anzeichen sucht, ersieht man daraus, daß $R(C,A)$ umso größer ist, je größer die $\max_i \frac{p_{i|j}}{p_{i.}}$ sind, d.h. je besser die Attributwerte a_j mit einzelnen Klassen verknüpft sind.

3.4 χ^2-Maß

Wir haben in Abschnitt 3.1 den Informationsgewinn gedeutet als die zu erwartende Verringerung der Anzahl der Fragen, die zur Identifikation der wahren Klasse notwendig sind. Wenn man ihn etwas anders schreibt, kann er aber auch gedeutet werden als ein Maß für den Unterschied der vorliegenden gemeinsamen Verteilung und der Verteilung, die sich bei Annahme der Unabhängigkeit der Attributwerte und der Klassen aus den Randverteilungen berechnen läßt. Für diesen Vergleich wird der logarithmus dualis der Verhältnisse einander entsprechender Wahrscheinlichkeiten der beiden Verteilungen summiert. Diese Form nennt man üblicherweise *wechselseitige Information* (mutual information).

$$I_{\text{mutual}}(C,A) = \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 \frac{p_{ij}}{p_{i.}\, p_{.j}}$$
$$= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 p_{ij} - \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 p_{i.} - \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 p_{.j}$$
$$= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} p_{ij} \log_2 p_{ij} - \sum_{i=1}^{n_C} p_{i.} \log_2 p_{i.} - \sum_{j=1}^{n_A} p_{.j} \log_2 p_{.j}$$
$$= -H_{CA} + H_C + H_A = I_{\text{gain}}$$

Die Rechnung zeigt, daß die wechselseitige Information tatsächlich mit dem Informationsgewinn identisch ist.

Die wechselseitige Information vergleicht die gemeinsame Verteilung mit einer hypothetischen unabhängigen Verteilung mit Hilfe des Quotienten der Wahrscheinlichkeiten. Ein Vergleich läßt sich aber auch durch Bildung des Abstandsquadrats durchführen. Dies führt zu dem in der Statistik wohlbekannten χ^2-*Maß*

$$\chi^2(C,A) = \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{(E_{ij} - N_{ij})^2}{E_{ij}} \qquad \text{mit } E_{ij} = \frac{N_{i.}\, N_{.j}}{N_{..}}$$

$$= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{N_{\cdot\cdot}^2 \left(\frac{N_{i\cdot}}{N_{\cdot\cdot}} \frac{N_{\cdot j}}{N_{\cdot\cdot}} - \frac{N_{ij}}{N_{\cdot\cdot}} \right)^2}{N_{\cdot\cdot} \frac{N_{i\cdot}}{N_{\cdot\cdot}} \frac{N_{\cdot j}}{N_{\cdot\cdot}}}$$
$$= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} N_{\cdot\cdot} \frac{(p_{i\cdot}\, p_{\cdot j} - p_{ij})^2}{p_{i\cdot}\, p_{\cdot j}}.$$

Nach der Umformung sieht man deutlich, daß der Zähler den Abstand der gemeinsamen von der hypothetischen unabhängigen Verteilung darstellt.

3.5 *g*-Funktion

Das Lernen von Entscheidungsbäumen ist eng verwandt mit dem Lernen Bayesscher Netzwerke. In einem Bayesschen Netzwerk wird einem Attribut zusammen mit seinen Elternattributen eine bedingte Wahrscheinlichkeitsverteilung zugeordnet. Diese kann man sich durch einen Entscheidungsbaum dargestellt denken, wobei als zusätzliche Bedingung gefordert wird, daß alle Blätter in der gleichen Baumebene liegen und alle Entscheidungen in einer Baumebene in der gleichen Weise auf dem gleichen Attribut gefällt werden müssen. Die Verwandtschaft Bayesscher Netze zu Entscheidungsbäumen legt die Idee nahe, Bewertungsmaße, die zum Lernen Bayesscher Netze verwendet werden, auch zur Induktion von Entscheidungsbäumen einzusetzen.

Ein solches Bewertungsmaß ist die von [Cooper und Herskovits 1992] hergeleitete *g-Funktion*, die für ein bedingendes Attribut (wie es ja beim Entscheidungsbaumlernen nur auftritt) lautet:

$$g(C, A) = c \cdot \prod_{j=1}^{n_A} \frac{(n_C - 1)!}{(N_{\cdot j} + n_C - 1)!} \prod_{i=1}^{n_C} N_{ij}!.$$

Anschaulich beschreibt diese Funktion (für einen bestimmten Wert von c) die Wahrscheinlichkeit, die gemeinsame Verteilung der Klassen und der Attributwerte in der Menge der Fallbeschreibungen zu finden. Wenn man annimmt, daß alle Abhängigkeiten des Klassenattributes von anderen Attributen a-priori gleichwahrscheinlich sind, und bei gegebener Abhängigkeit alle bedingten Wahrscheinlichkeitsverteilungen a-priori gleichwahrscheinlich sind, schließt die g-Funktion in Bayesscher Weise von der Wahrscheinlichkeit der Daten gegeben die Abhängigkeit auf die Wahrscheinlichkeit der Abhängigkeit gegeben die Daten.

Da der Wert der g-Funktion stark von der Anzahl der Fallbeschreibungen abhängt, ist es vorteilhaft, die folgende normierte Form zu verwenden:

$$\frac{\log_2(g(C, A))}{N_{\cdot\cdot}} = \frac{1}{N_{\cdot\cdot}} \left(\log_2 c + \sum_{j=1}^{n_A} \left(\log_2 \frac{(n_C - 1)!}{(N_{\cdot j} + n_C - 1)!} + \sum_{i=1}^{n_C} \log_2 N_{ij}! \right) \right)$$

Mit dieser Form des Maßes ist es möglich, Bewertungen zu vergleichen, die auf Datensätzen unterscheidlicher Größe erzielt wurden.

3.6 Minimale Beschreibungslänge

Schon der Informationsgewinn (siehe Abschnitt 3.1) konnte als Reduktion der Beschreibungslänge gedeutet werden. Er nimmt jedoch das Kodierungsschema als bereits bekannt an. Das Prinzip der minimalen Beschreibungslänge stellt dagegen zusätzlich die Kosten für die Beschreibung des Kodierungsschemas in Rechnung.

Die zugrundeliegende Idee ist anschaulich die folgende: Ein Sender S möchte einem Empfänger E eine Nachricht übertragen. Der Empfänger kennt die Symbole, aus denen die Nachricht zusammengesetzt sein kann, weiß jedoch nichts über die Häufigkeit, mit der sie in der von S zu sendenden Nachricht auftreten. S kann daher nicht direkt z.B. einen Huffman-Kode für die Übertragung verwenden, da E diesen wegen der fehlenden Häufigkeitsinformation nicht entschlüsseln kann. Wenn die Nachricht jedoch lang genug ist, kann es sich für S lohnen, zuerst eine Beschreibung des Kodierungsschemas zu senden und dann die nach diesem Schema kodierte Nachricht. Die Summe der zu übertragenden Symbole kann bei diesem Verfahren geringer sein als bei einer naiven, auf gleicher Wahrscheinlichkeit der Symbole basierenden Kodierung.

Angewandt auf Entscheidungsbäume stellt sich das Übertragungsproblem so dar: Sowohl der Sender als auch der Empfänger kennen die Attribute und Attributwerte einer Menge von Fallbeschreibungen, aber nur der Sender kennt die Klassenzuordnungen. Das Ziel besteht darin, unter Verwendung der Attribute und ihrer Werte eine möglichst effiziente Kodierung für die Übertragung der Klassenzuordnungen zu finden. Dazu werden die Kosten einer Kodierung der Klassenzuordnungen allein verglichen mit den Kosten einer Kodierung der Klassenzuordnungen in den durch die Werte eines Attributes bestimmten Teilmengen der Fallbeschreibungen. Die Kosten, die für die Übertragung der bedingten Kodierungsschemata entstehen, können dabei als „Strafe“ dafür gesehen werden, daß der Entscheidungsbaum komplexer gemacht wird.

Für die Kodierung der Klassenzuordnungen können zwei verschiedene Methoden angewandt werden. Erstere kodiert die Klassenzuordnungen auf der Grundlage ihrer relativen, letztere auf der Grundlage ihrer absoluten Häufigkeiten. Für eine Kodierung auf der Grundlage relativer Häufigkeiten sehen die Rechnungen so aus:

$$
\begin{aligned}
L^{(1)}_{\text{prior}}(C) &= \log_2 \frac{(N_{\cdot\cdot} + n_C - 1)!}{N_{\cdot\cdot}!\,(n_C - 1)!} + N_{\cdot\cdot}\, H_C, \\
L^{(1)}_{\text{post}}(C, A) &= \log_2 k + \sum_{j=1}^{n_A} \log_2 \frac{(N_{\cdot j} + n_C - 1)!}{N_{\cdot j}!\,(n_C - 1)!} + \sum_{j=1}^{n_A} N_{\cdot j}\, H_{C|a_j}, \\
L^{(1)}_{\text{gain}}(C, A) &= L^{(1)}_{\text{prior}}(C) - L^{(1)}_{\text{post}}(C, A).
\end{aligned}
$$

$L^{(1)}_{\text{prior}}$ ist die Beschreibungslänge, die sich für die Übertragung der Klassenzuordnungen ohne Zuhilfenahme eines Attributes (vor der Kenntnis seines Wertes) ergibt. Der erste Term dieser Länge beschreibt die Kosten für Übertragung der Klassenhäufigkeiten. Dazu stellt man sich vor, daß alle möglichen Zuordnungen

von $N_{..}$ Fällen zu n_C Klassen in einem Kodebuch verzeichnet sind, eine je Seite. Die Kombinatorik lehrt, daß dieses Buch dann gerade $\frac{(N_{..}+n_C-1)!}{N_{..}!\,(n_C-1)!}$ Seiten haben muß. Nimmt man jede Zuordnung als gleichwahrscheinlich an, braucht man für die Übertragung der Nummer der Seite, auf der die vorliegende Klassenzuordnung steht, gerade die Hartley-Information der Buchseiten, also den logarithmus dualis der Seitenzahl. Die Kosten für die Übertragung der eigentlichen Daten, d.h. der Klassenzuordnungen, werden mit Hilfe der Shannonschen Entropie bestimmt, die ja (siehe Abschnitt 3.1) die Anzahl zu übertragender Bits je Symbol angibt. Diese Kosten beschreibt der zweite Term.

In gleicher Weise wird die Beschreibungslänge in den durch die Werte eines Attributes A bestimmten Teilmengen der Fallbeschreibungen bestimmt. Diese werden dann über die Teilmengen summiert. Außerdem wird der Term $\log_2 k$ hinzugefügt (wobei k die Anzahl der für einen Test zur Verfügung stehenden Attribute ist), der die Kosten für die Mitteilung des verwendeten Attributes angibt. (Im allgemeinen wird dieser Term jedoch vernachlässigt, da er für alle Attribute gleich ist.) Man gelangt so zu der bei Kenntnis des Wertes des Attributes A zu erwartenden Beschreibungslänge $L^{(1)}_{\text{post}}$. Als Attributauswahlmaß verwendet man die Differenz $L^{(1)}_{\text{gain}}$ der beiden Beschreibungslängen.

Bei der Kodierung auf der Grundlage absoluter Häufigkeiten geht man man wie folgt vor:

$$
\begin{aligned}
L^{(2)}_{\text{prior}}(C) &= \log_2 \frac{(N_{..}+n_C-1)!}{N_{..}!\,(n_C-1)!} + \log_2 \frac{N_{..}!}{N_{1.}!\cdots N_{n_C.}!},\\
L^{(2)}_{\text{post}}(C,A) &= \log_2 k + \sum_{j=1}^{n_A} \log_2 \frac{(N_{.j}+n_C-1)!}{N_{.j}!\,(n_C-1)!} + \sum_{j=1}^{n_A} \log_2 \frac{N_{.j}!}{N_{1j}!\cdots N_{n_Cj}!},\\
L^{(2)}_{\text{gain}}(C,A) &= L^{(2)}_{\text{prior}}(C) - L^{(2)}_{\text{post}}(C,A).
\end{aligned}
$$

Der erste Term der Beschreibungslänge $L^{(2)}_{\text{prior}}$ ist der gleiche wie bei einer Kodierung auf der Grundlage relativer Häufigkeiten. Durch ihn werden die Kosten der Übertragung der Klassenhäufigkeitsverteilung beschrieben. Den zweiten Term erhält man diesmal jedoch durch eine ähnliche Überlegung wie den ersten. Offenbar können die Fälle den Klassen zugeordnet werden, indem man $N_{1.}$ Fälle auswählt und sie der Klasse 1 zuordnet, dann aus den verbleibenden Fällen $N_{2.}$ Fälle auswählt und diese der Klasse 2 zuordnet usw. Sämtliche Möglichkeiten solcher Zuordnungen denken wir uns wieder in einem Kodebuch verzeichnet, eine je Seite. Aus der Kombinatorik entnehmen wir, daß dieses Buch $\frac{N_{..}!}{N_{1.}!\cdots N_{n_C.}!}$ Seiten haben muß. Wir nehmen wieder Gleichwahrscheinlichkeit an und erhalten so die Hartley-Information der Buchseiten, d.h. den logarithmus dualis der Seitenzahl als zweiten Term.

Die zu erwartende Beschreibungslänge $L^{(2)}_{\text{post}}$ bei bekanntem Wert des Attributes A erhält man auf analoge Weise. (k ist wieder die Anzahl der für einen Test zur Verfügung stehenden Attribute.) Als Attributauswahlmaß verwendet man die Differenz $L^{(2)}_{\text{gain}}$ der beiden Beschreibungslängen. Wir bemerken hier noch,

daß die Kodierung auf der Grundlage absoluter Häufigkeiten eng mit der g-Funktion verwandt ist, genauer: $\log_2(g(C,A)) = \log_2(c)\left(\log_2 k - L^{(2)}_{\text{post}}(C,A)\right)$.

Wenn einige der Klassenhäufigkeiten nahe bei Null liegen, können die Kosten für die Übertragung des Kodierungsschemas in der Größenordnung der Kosten für die eigentliche Datenübertragung liegen. Dieser Nachteil wird von der sogenannten *stochastischen Komplexität* [Krichevsky und Trofimov 1983, Rissanen 1995]

$$SC = \log_2 \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} + \frac{n-1}{2}\log_2 \frac{N}{2} - NH$$

überwunden. Die Idee ist anschaulich die folgende: Bei den bisher betrachteten Maßen wird jede Verteilung der Klassen a-priori als gleichwahrscheinlich angenommen. Damit liegen offenbar die Kosten für die Übertragung des Kodierungsschemas fest. Wenn man jedoch die A-priori-Wahrscheinlichkeiten so wählt, daß die tatsächliche Klassenverteilung eine hohe Wahrscheinlichkeit hat, kann man offenbar diese Kosten verringern. (Wir sahen ja in Abschnitt 3.1, daß in einer guten Kodierung wahrscheinliche Werte einen kurzen Kode haben, d.h. durch eine geringe Anzahl Fragen bestimmbar sind.) Im günstigsten Fall hat die tatsächlich vorliegende Klassenverteilung die A-priori-Wahrscheinlichkeit 1, so daß die Übertragungskosten entfallen. Das Problem veränderter A-priori-Wahrscheinlichkeiten besteht darin, daß der Empfänger sie natürlich nicht kennt. Daher muß der Sender, wenn er veränderte A-priori-Wahrscheinlichkeiten ausnutzen möchte, diese vorher dem Empfänger mitteilen, was selbst wieder Übertragungskosten verursacht. Man sieht so leicht, daß ein Setzen der A-priori-Wahrscheinlichkeit der vorliegenden Verteilung zu Eins keinen Vorteil bringt, da die Übertragung dieser Information die gleichen Kosten verursacht wie das Übertragen der Verteilung bei Annahme gleicher A-priori-Wahrscheinlichkeiten. Es gibt aber zwischen den beiden Extremen der einzelnen, zu Eins gesetzten A-priori-Wahrscheinlichkeit und den gleichen A-priori-Wahrscheinlichkeiten für alle Verteilungen ein Minimum. Dieses wird gerade durch die stochastische Komplexität beschrieben.

Indem man die stochastische Komplexität in gleicher Weise verwendet wie die oben besprochenen Beschreibungslängenmaße, erhält man

$$\begin{aligned}
L^{(3)}_{\text{prior}}(C) &= \log_2 \frac{\pi^{\frac{n_C}{2}}}{\Gamma(\frac{n_C}{2})} + \frac{n_C-1}{2}\log_2 \frac{N_{..}}{2} - N_{..}\, H_C, \\
L^{(3)}_{\text{post}}(C,A) &= \log_2 k + n_A \log_2 \frac{\pi^{\frac{n_C}{2}}}{\Gamma(\frac{n_C}{2})} + \frac{n_C-1}{2}\sum_{j=1}^{n_A}\log_2 \frac{N_{.j}}{2} - N_{..}\, H_{C|A}, \\
L^{(3)}_{\text{gain}}(C,A) &= L^{(3)}_{\text{prior}}(C) - L^{(3)}_{\text{post}}(C,A),
\end{aligned}$$

wobei k wieder die Anzahl der für einen Test zur Verfügung stehenden Attribute ist. Als Auswahlmaß wird die Differenz $L^{(3)}_{\text{gain}}$ der beiden Längen verwendet.

3.7 Spezifizitätsgewinn

Viele der bisher vorgestellten Maße basierten auf der Wahrscheinlichkeitstheorie. Diese ist aber nicht das einzige Kalkül zur Quantifizierung von Möglichkeit. Eine beachtenswerte Alternative stellt die Possibilitätstheorie dar, die eng mit der Theorie der Fuzzy-Mengen verwandt ist. In dieser Theorie übernimmt die *Nichtspezifizität* (nonspecificity) einer Possibilitätsverteilung die Rolle der Shannonschen Entropie einer Wahrscheinlichkeitsverteilung. Die Nichtspezifizität einer Possibilitätsverteilung π ist allgemein definiert als

$$\mathrm{nsp}(\pi|P_1, P_2) = -\int_0^{\sup(\pi)} \sum_{\omega\in\Omega} P_1(\omega|[\pi]_\alpha) \log_2 P_2(\omega|[\pi]_\alpha) d\alpha,$$

wobei P_1 und P_2 bedingte Wahrscheinlichkeitsverteilungen sind. Wählt man für P_1 und P_2 Gleichverteilungen, ergibt sich das *U-Unsicherheitsmaß der Nichtspezifizität* (U-uncertainty measure of nonspecificity) [Higashi und Klir 1982]

$$\mathrm{nsp}(\pi) = \int_0^{\sup(\pi)} \log_2 |[\pi]_\alpha| d\alpha,$$

das als Verallgemeinerung der Hartley-Information [Hartley 1928] (siehe Abschnitt 3.1) gerechtfertigt werden kann [Klir und Mariano 1987]. Wenn man eine Gleichverteilung über den möglichen Konfidenzgraden $\alpha \in [0, \sup(\pi)]$ annimmt, beschreibt $\mathrm{nsp}(\pi)$ die durchschnittliche Menge an Information (in Bit), die man auf jedem α-Schnitt $[\pi]_\alpha$ noch hinzufügen muß, um den wahren Wert zu bestimmen [Gebhardt und Kruse 1996].

Indem man das U-Unsicherheitsmaß der Nichtspezifizität in der gleichen Weise verwendet wie die Shannonsche Entropie, läßt sich der *Spezifizitätsgewinn* (specificity gain) definieren [Borgelt et al. 1996]

$$\begin{aligned} S_{\mathrm{gain}}(C, A) &= \mathrm{nsp}(\pi_C) + \mathrm{nsp}(\pi_A) - \mathrm{nsp}(\pi_{CA}) \\ &= \int_0^{\sup(\pi_A)} \log_2 |[\pi_A]_\alpha| \, \mathrm{d}\alpha + \int_0^{\sup(\pi_C)} \log_2 |[\pi_C]_\alpha| \, \mathrm{d}\alpha \\ &\quad - \int_0^{\sup(\pi_{CA})} \log_2 |[\pi_{CA}]_\alpha| \, \mathrm{d}\alpha \end{aligned}$$

Dieses Maß ist äquivalent zu dem von [Gebhardt und Kruse 1996] zum Lernen possibilistischer Netzwerke verwendeten Maß. Bei der Berechnung dieses Maßes hat man zu beachten, daß die Randverteilungen in Übereinstimmung mit der Possibilitätstheorie gebildet werden, d.h. man darf die Fallzahlen der gemeinsamen Verteilung nicht (wie in der Wahrscheinlichkeitstheorie) summieren, sondern muß das Maximum bilden:

$$\begin{aligned} \forall a \in A: \ \pi_A(a) &= \max_{c\in C} (\pi_{CA}(c, a)) \\ \forall c \in C: \ \pi_C(c) &= \max_{a\in A} (\pi_{CA}(c, a)) \end{aligned}$$

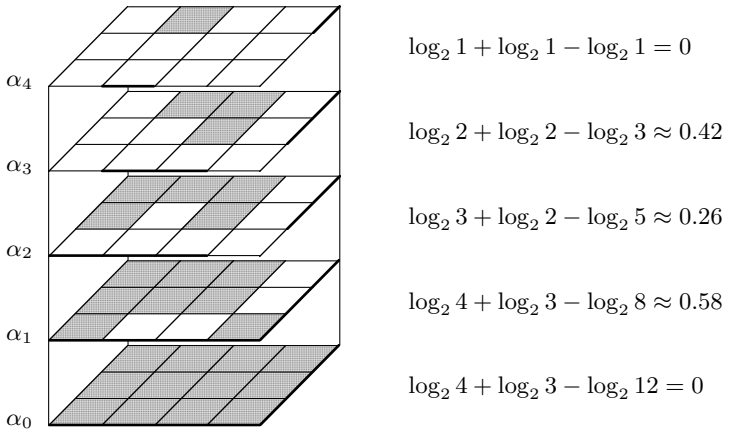

Abbildung 4. Eine zweidimensionale Possibilitätsverteilung wird wie eine Menge von Relationen gesehen: eine Relation für jeden α-Schnitt. Indem man für jeden α-Wert den Hartley-Informationsgewinn berechnet und diesen über alle möglichen α-Werte integriert, erhält man den Spezifizitätsgewinn.

Analog zum Informationsgewinn gibt es verschiedene Möglichkeiten der Normalisierung, um eine eventuelle Verzerrung zugunsten von Attributen mit vielen Werten zu beseitigen oder wenigstens zu verringern. Man gelangt so zum *Spezifizitätsgewinnverhältnis* (specificity gain ratio)

$$S_{\mathrm{gr}}(C,A) = \frac{S_{\mathrm{gain}}(A)}{\mathrm{nsp}(\pi_A)} = \frac{\mathrm{nsp}(\pi_C) + \mathrm{nsp}(\pi_A) - \mathrm{nsp}(\pi_{CA})}{\mathrm{nsp}(\pi_A)}$$

und zu zwei *symmetrischen Spezifizitätsgewinnverhältnissen*

$$S^{(1)}_{\mathrm{sgr}}(C,A) = \frac{S_{\mathrm{gain}}(A)}{\mathrm{nsp}(\pi_{CA})} = \frac{\mathrm{nsp}(\pi_C) + \mathrm{nsp}(\pi_A) - \mathrm{nsp}(\pi_{CA})}{\mathrm{nsp}(\pi_{CA})},$$
$$S^{(2)}_{\mathrm{sgr}}(C,A) = \frac{S_{\mathrm{gain}}(A)}{\mathrm{nsp}(\pi_A) + \mathrm{nsp}(\pi_C)} = \frac{\mathrm{nsp}(\pi_C) + \mathrm{nsp}(\pi_A) - \mathrm{nsp}(\pi_{CA})}{\mathrm{nsp}(\pi_A) + \mathrm{nsp}(\pi_C)}.$$

Die Idee des Spezifizitätsgewinns ist in Abbildung 4 illustriert. Eine zweidimensionale Possibilitätsverteilung wird als Menge von Relationen gesehen: eine Relation für jeden α-Schnitt. Indem man für jeden α-Wert den Hartley-Informationsgewinn berechnet und diesen über alle möglichen α-Werte integriert, erhält man den Spezifizitätsgewinn.

4 Experimentelle Ergebnisse

Zwar ist eine theoretische Begründung eines Attributauswahlmaßes, wie sie im vorangehenden Abschnitt gegeben wurde, wichtig, am Ende zählt jedoch der

Maß	Baum ungestutzt	gestutzt
Informationsgewinn	41.6	39.8
Info.gew.verhältnis	40.6	39.8
sym. Info.gew.verh. 1	40.5	40.6
sym. Info.gew.verh. 2	40.5	38.6
Gini-Index	42.3	40.8
norm. Gini-Index	40.9	40.1
Relevanz	56.6	49.2
χ^2-Maß	43.4	41.6
$log_2(g)/N$	40.7	39.6
Kod. rel. Häufigkeit	42.3	40.5
Kod. abs. Häufigkeit	40.0	39.2
Stoch. Komplexität	42.9	41.0
Spezifizitätsgewinn	42.5	39.3
Spez.gew.verhältnis	44.8	41.7
sym. Spez.gew.verh. 1	42.6	39.7
sym. Spez.gew.verh. 2	43.1	40.0

Tabelle 4. Über 27 Datensätze aus dem UC Irvine Machine Learning Repository gemittelte Ergebnisse der verschiedenen Bewertungsmaße. Als Vergleichsgröße wurde die naive Vorhersage der Mehrheitsklasse gewählt. Die Zahlenwerte geben an, wieviel Prozent der Fehlerrate der naiven Vorhersage mit dem entsprechenden Maß erzielt wurden. (Es gilt also: Je kleiner der Wert, desto besser das Maß.) Offenbar ist die Qualität der Maße recht einheitlich, nur das Relevanzmaß schneidet schlechter ab.

Erfolg in der Praxis. Daher vergleichen wir in diesem Abschnitt die beschriebenen Maße anhand experimenteller Ergebnisse. Zur Erzielung dieser Ergebnisse gingen wir von dem bekannten Entscheidungsbaum-Lernprogramm C4.5 (Release 7) [Quinlan 1993] aus, das in der Originalversion über die Auswahlmaße Informationsgewinn und Informationsgewinnverhältnis verfügt. Es wurde mit einer Schnittstelle versehen, an der beliebige andere Auswahlmaße eingebunden werden können. Allerdings mußten dazu einige Besonderheiten dieses Programms entfernt werden. Die mit dem Informationsgewinnverhältnis erzielten Ergebnisse unterscheiden sich daher von denen, die die Originalversion liefert.

Für unsere Tests wählten wir 27 Datensätze aus dem UC Irvine Machine Learning Repository (abalone, anneal, audiology, autos, balance, breast-wisconsin, credit-australia, credit-germany, glass, heart-cleveland, heart-hungary, hepatitis, horse-colic, hypo, iris, labor, new-thyroid, pima-indians-diabetes, satimage, segment, sick, sonar, soybean, vehicle, vote, vowel, waveform). Die über diese Datensätze gemittelten, jeweils mit zehnfacher Cross-Validierung und anschließender Mittelung erzielten Ergebnisse sind in Tabelle 4 dargestellt. Wie man sieht, sind die Ergebnisse relativ einheitlich, nur das Relevanzmaß schneidet deutlich schlechter ab.

Eine einfache Rangfolge der Maße läßt sich jedoch nicht angeben. Vielmehr hängt es vom einzelnen Datensatz ab, welches Maß zu dem besten Klassifikationsergebnis führt. Dies wird durch die beiden in Tabelle 5 dargestellten Ergebnisse belegt. Insbesondere das Relevanzmaß schneidet für den Datensatz credit-australia (wie im Durchschnitt) sehr schlecht ab, liefert aber für den Datensatz soybean eines der besten Ergebnisse. Gerade umgekehrt verhält es sich mit den Spezifizitätsmaßen. Wenn man an einem möglichst guten Klassifikator interessiert ist, kann es sich daher lohnen, mehrere oder gar alle der vorgestellten Maße auszuprobieren.

credit-australia (621/69)	vor Stutzen des Baumes Größe	Lerndaten	Testdaten	nach Stutzen des Baumes Größe	Lerndaten	Testdaten
default		276.3 (44.5)	30.7 (44.5)		276.3 (44.5)	30.7 (44.5)
Informationsgewinn	156.6	28.7 (4.6)	13.7 (19.9)	45.0	52.8 (8.5)	10.1 (14.6)
Info.gew.verhältnis	140.6	22.2 (3.6)	13.2 (19.1)	55.2	40.5 (6.5)	11.3 (16.4)
sym. Info.verh. 1	138.5	24.0 (3.9)	12.2 (17.7)	39.9	53.2 (8.6)	10.7 (15.5)
sym. Info.verh. 2	153.7	24.8 (4.0)	12.9 (18.7)	34.7	56.6 (9.1)	11.3 (16.4)
Gini-Index	130.5	33.1 (5.3)	13.3 (19.3)	47.0	51.8 (8.4)	10.4 (15.1)
norm. Gini-Index	130.0	25.3 (4.0)	11.8 (17.1)	46.5	48.3 (7.8)	9.8 (14.2)
Relevanz	345.3	46.8 (7.5)	20.3 (29.4)	26.7	83.7 (13.5)	12.6 (18.2)
χ^2-Maß	122.7	34.6 (5.6)	12.4 (18.0)	48.8	51.5 (8.3)	10.4 (15.1)
$\log_2(g)/N$	104.2	26.8 (4.3)	11.0 (15.9)	56.2	38.8 (6.3)	10.1 (14.6)
Kod. rel. Häufigkeit	145.1	21.5 (3.5)	12.6 (18.2)	48.7	44.8 (7.2)	10.3 (14.9)
Kod. abs. Häufigkeit	109.8	23.8 (3.8)	10.6 (15.3)	56.8	39.3 (6.3)	9.9 (14.3)
Stoch. Komplexität	261.3	43.0 (6.9)	15.2 (22.0)	3.0	90.0 (14.5)	10.0 (14.5)
Spezifizitätsgewinn	221.7	25.3 (4.1)	12.9 (18.7)	42.4	51.7 (8.3)	10.4 (15.1)
Spez.gew.verhältnis	198.1	29.5 (4.7)	12.0 (17.4)	53.5	44.3 (7.1)	10.6 (15.3)
sym. Spez.verh. 1	217.2	31.2 (5.0)	13.1 (19.0)	36.0	57.1 (9.2)	9.4 (13.6)
sym. Spez.verh. 2	230.4	31.9 (5.1)	14.2 (20.6)	34.9	58.8 (9.5)	10.7 (15.5)

soybean (615/68)	vor Stutzen des Baumes Größe	Lerndaten	Testdaten	nach Stutzen des Baumes Größe	Lerndaten	Testdaten
default		532.2 (86.5)	58.8 (86.5)		532.2 (86.5)	58.8 (86.5)
Informationsgewinn	263.1	17.3 (2.8)	9.6 (14.1)	110.6	29.8 (4.8)	6.9 (10.1)
Info.gew.verhältnis	161.1	16.5 (2.7)	6.7 (9.9)	93.8	22.4 (3.6)	5.7 (8.4)
sym. Info.verh. 1	266.4	14.6 (2.4)	7.4 (10.9)	92.7	31.0 (5.1)	6.3 (9.3)
sym. Info.verh. 2	334.7	13.3 (2.2)	9.2 (13.5)	109.5	27.8 (4.5)	6.8 (10.0)
Gini-Index	328.2	23.8 (3.9)	11.4 (16.8)	128.1	39.9 (6.5)	8.9 (13.1)
norm. Gini-Index	212.2	17.5 (2.8)	6.0 (8.8)	96.7	24.4 (4.0)	5.1 (7.5)
Relevanz	209.8	15.8 (2.6)	5.6 (8.2)	102.2	25.5 (4.1)	5.8 (8.5)
χ^2-Maß	232.6	21.7 (3.5)	8.0 (11.8)	99.8	33.5 (5.5)	6.7 (9.8)
$\log_2(g)/N$	409.4	25.1 (4.1)	7.7 (11.3)	90.1	39.9 (6.5)	5.7 (8.4)
Kod. rel. Häufigkeit	224.8	17.2 (2.8)	6.3 (9.3)	91.8	20.8 (3.4)	4.9 (7.2)
Kod. abs. Häufigkeit	242.5	16.3 (2.7)	6.3 (9.3)	89.2	24.9 (4.0)	5.3 (7.8)
Stoch. Komplexität	441.3	52.0 (8.5)	13.4 (19.7)	81.8	69.9 (11.4)	10.2 (15.0)
Spezifizitätsgewinn	271.6	14.0 (2.3)	8.2 (12.1)	105.9	28.7 (4.7)	6.8 (10.0)
Spez.gew.verhältnis	242.0	19.3 (3.1)	7.6 (11.2)	98.7	28.8 (4.7)	7.0 (10.3)
sym. Spez.verh. 1	290.0	25.9 (4.2)	9.7 (14.3)	135.7	39.4 (6.4)	8.6 (12.7)
sym. Spez.verh. 2	296.5	27.1 (4.4)	10.4 (15.3)	136.7	41.4 (6.7)	9.5 (14.0)

Tabelle 5. Experimentelle Ergebnisse, die mit einer angepaßten Version des Entscheidungsbaum-Lernprogramms C4.5 [Quinlan 1993] unter Verwendung zehnfacher Cross-Validierung auf zwei Datensätzen aus dem UC Irvine Machine Learning Repository erzielt wurden. Die Zahlen geben die Anzahl Knoten des Baumes und die absolute und (in Klammern) relative Fehlerhäufigkeit an.

5 Zusammenfassung

In diesem Aufsatz haben wir eine ganze Reihe von Attributauswahlmaßen für die Induktion von Entscheidungsbäumen besprochen, die auf zum Teil sehr verschiedenen Ideen beruhen. Der Informationsgewinn und der Gini-Index stützen sich auf die Informationstheorie, und zwar ersterer auf die Shannonsche und letzterer auf die quadratische Entropie einer Wahrscheinlichkeitsverteilung. Mit dem Relevanzmaß sucht man nach möglichst eindeutigen Anzeichen für eine Klassenzugehörigkeit, und das χ^2-Maß vergleicht die gemeinsame Verteilung von Attributwerten und Klassen mit einer hypothetischen unabhängigen Verteilung. Die g-Funktion wird über Bayessches Schließen begründet, und die mit ihr formal eng verwandten, auf dem Prinzip der minimalen Beschreibungslänge basierenden Maße lassen sich aus einem Datenübertragungsmodell für die Klassenzuordnungen ableiten. Der Spezifizitätsgewinn schließlich wird analog zum Informationsgewinn aus der Nichtspezifizität einer Possibilitätsverteilung abgeleitet.

Die experimentellen Ergebnisse zeigen, daß alle Maße für die Induktion von Entscheidungsbäumen geeignet sind, wenn man auch beim Relevanzmaß leichte Abstriche machen muß. Zwar scheinen im Durchschnitt die Informationsgewinnverhältnisse und die auf dem Prinzip der minimalen Beschreibungslänge basierenden Maße zu leicht besseren Ergebnisse zu führen als die anderen Maße, eine klare Überlegenheit eines einzelnen Maßes läßt sich jedoch nicht feststellen. Vielmehr ist die Qualität des gelernten Klassifikators auch vom Datensatz abhängig. Je nach seinen besonderen Eigenschaften wird einmal dieses, einmal jenes Maß das beste Ergebnis liefern. Es kann sich daher lohnen, mehrere oder gar alle Maße auszuprobieren, um einem möglichst guten Entscheidungsbaum zu erhalten.

Literatur

[Baim 1988] P.W. Baim. A Method for Attribute Selection in Inductive Learning Systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-10:888-896, 1988

[Borgelt et al. 1996] C. Borgelt, J. Gebhardt und R. Kruse. Concepts for Probabilistic and Possibilistic Induction of Decision Trees on Real World Data. *Proc. of the EUFIT'96*, Vol. 3:1556–1560, 1996

[Breiman et al. 1984] L. Breiman, J.H. Friedman, R.A. Olshen und C.J. Stone. *Classification and Regression Trees*, Wadsworth International, Belmont, CA, 1984

[Chow und Liu 1968] C.K. Chow und C.N. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Trans. on Information Theory* 14(3):462–467, IEEE 1968

[Cooper und Herskovits 1992] G.F. Cooper und E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9:309–347, Kluwer 1992

[Daróczy 1970] Z. Daróczy. Generalized Information Functions. *Information and Control* 16:36–51, 1970

[Gebhardt und Kruse 1992] J. Gebhardt und R. Kruse. A Possibilistic Interpretation of Fuzzy Sets in the Context Model. *Proc. IEEE Int. Conf. on Fuzzy Systems*, 1089-1096, San Diego 1992.

[Gebhardt und Kruse 1995] J. Gebhardt und R. Kruse. Learning Possibilistic Networks from Data. *Proc. 5th Int. Workshop on AI and Statistics*, 233–244, Fort Lauderdale, 1995

[Gebhardt und Kruse 1996] J. Gebhardt und R. Kruse. Tightest Hypertree Decompositions of Multivariate Possibility Distributions. *Proc. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems*, 1996

[Hartley 1928] R.V.L. Hartley. Transmission of Information. *The Bell Systems Technical Journal* 7:535–563, 1928

[Heckerman et al. 1995] D. Heckerman, D. Geiger und D.M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243, Kluwer 1995

[Higashi und Klir 1982] M. Higashi und G.J. Klir. Measures of Uncertainty and Information based on Possibility Distributions. *Int. Journal of General Systems* 9:43–58, 1982

[Huffman 1952] D.A. Huffman. A Method for the Construction of Minimum Redundancy Codes. *Proc. IRE 40*, No. 10, 1098–1101, 1952

[Klir und Mariano 1987] G.J. Klir und M. Mariano. On the Uniqueness of a Possibility Measure of Uncertainty and Information. *Fuzzy Sets and Systems* 24:141–160, 1987

[Kononenko 1995] I. Kononenko. On Biases in Estimating Multi-Valued Attributes. *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining*, 1034–1040, Montreal, 1995

[Krichevsky und Trofimov 1983] R.E. Krichevsky und V.K. Trofimov. The Performance of Universal Coding. *IEEE Trans. on Information Theory*, 27(2):199–207, 1983

[Kruse et al. 1991] R. Kruse, E. Schwecke und J. Heinsohn. *Uncertainty and Vagueness in Knowledge-based Systems: Numerical Methods.* Springer, Berlin 1991

[Kruse et al. 1994] R. Kruse, J. Gebhardt und F. Klawonn. *Foundations of Fuzzy Systems*, John Wiley & Sons, Chichester, England 1994

[Kullback und Leibler 1951] S. Kullback und R.A. Leibler. On Information and Sufficiency. *Ann. Math. Statistics* 22:79–86, 1951

[Lopez de Mantaras 1991] R. Lopez de Mantaras. A Distance-based Attribute Selection Measure for Decision Tree Induction. *Machine Learning* 6:81–92, Kluwer 1991

[Murphy und Aha 1994] P.M. Murphy und D. Aha, UCI Repository of Machine Learning Databases, ftp://ics.uci.edu/pub/machine-learning-databases, 1994

[Nguyen 1984] H.T. Nguyen. Using Random Sets. *Information Science* 34:265–274, 1984

[Quinlan 1986] J.R. Quinlan. Induction of Decision Trees. *Machine Learning* 1:81–106, 1986

[Quinlan 1993] J.R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993

[Rissanen 1983] J. Rissanen. A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics* 11:416–431, 1983

[Rissanen 1995] J. Rissanen. Stochastic Complexity and Its Applications. *Proc. Workshop on Model Uncertainty and Model Robustness*, Bath, England, 1995

[Shannon 1948] C.E. Shannon. The Mathematical Theory of Communication. *The Bell Systems Technical Journal* 27:379–423, 1948

[Wehenkel 1996] L. Wehenkel. On Uncertainty Measures Used for Decision Tree Induction. *Proc. IPMU*, 1996

[Zhou und Dillon 1991] X. Zhou und T.S. Dillon. A statistical-heuristic Feature Selection Criterion for Decision Tree Induction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-13:834–841, 1991