Lernen von Entscheidungsbäumen und Regelmengen

Wir gehen davon aus, daß eine Sammlung von Lernbeispielen vorliegt, die ein Konzepte repräsentieren. Die Aufgabe besteht darin, einen Entscheidungsbaum zu generieren, der für dieses Konzept **effizient** ist (Baumgröße, Entscheidungsgeschwindigkeit, etc.).

Die Lernbeispiele sind charakterisiert durch eine feste Anzahl von Attributen mit jeweils endlich vielen Ausprägungen und durch eine Klassifikation die die möglichen Entscheidungen angibt. Im Gegensatz zur Mustererkennung haben wir es hier also mit diskreten Attributwerten zu tun.

Da nur eine (relativ kleine) Menge von Fallbeispielen vorliegt, kann es sein, daß über die in diesen Fällen vorkommenden Ausprägungen der Attribute hinaus noch weitere Ausprägungen möglich sind. Beispielsweise kann das Attribut "Größe einer Person in cm´´ jeden Wert im Bereich von 150 bis 200 annehmen, auch wenn 174 nicht in einem Fallbeispiel auftritt. Hier können wir die weiteren Attibutausprägungen voraussehen. Aber auch Ausprägungen, die wir nicht vorausgesehen haben, sollten in einem Entscheidungsbaum verarbeitet werden können, hier etwa eine Größe von 225cm.

Die Fallbeispiele lassen sich einfach als Tabelle angeben:

A_1	A_2	$\ldots$	A_n	C
$val_{1,1}$	$val_{1,2}$	$\ldots$	$val_{1,n}$	$class_1$
$val_{2,1}$	$val_{2,2}$	$\ldots$	$val_{2,n}$	$class_2$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$val_{m,1}$	$val_{m,2}$	$\ldots$	$val_{m,n}$	$class_m$

Hier bezeichnen $A_1, \ldots, A_n$ die verschiedenen Attribute wie z.B. Kosten in DM, Autofarbe, Abiturnote, etc. $a_{i,1}, \ldots, a_{i,k_i}$ sind mögliche Ausprägungen eines Attributs, etwa {rot, gelb, grün, blau, schwarz} für das Attribut Autofarbe. Also gilt $val_{j,i} \in \{a_{i,1}, \ldots, a_{i,k_i}\}$.

Die Klassifikatio wird durch C gegeben, wobei $c_1, \ldots, c_k$ die möglichen Werte seien, d.h. $class_j \in \{c_1, \ldots, c_k\}$. Beispiele sind die Kreditwürdigkeit mit den Möglichkeiten *ja*, *eingeschränkt* und *nein* oder aber die Eßbarkeit von Pilzen mit den Möglichkeiten *eßbar* und *giftig*.

Wir schreiben kurz:

$$\begin{aligned} A_i &= \{a_{i,1}, \ldots, a_{i,k_i}\} \qquad 1 \leq i \leq n \\ C &= \{c_1, \ldots, c_k\} \end{aligned}$$

Die Fallsammlung oder *Objektsammlung* M wird also gebildet durch m Objekte (Fälle), die durch n Attribute mit jeweils k_i Ausprägungen beschrieben werden und die in k Klassen fallen.

Wir gehen weiter davon aus, daß M folgende Bedingung erfüllt:

> M ist nicht trivial, d.h. C enthält mindestens zwei verschiedene Klassifikationen

Als eine weitere Bedingung kann in speziellen Situationen die Konsistenz der Fallsammlung angenommen werden:

> M ist konsistent, d.h. M enthält keine Beispiele, die bei gleichen Attributausprägungen verschiedene Klassifikatione erhalten. Für zwei Fallbeispiele $(val_{u,1}, \ldots, val_{u,n}, class_u)$ und $(val_{v,1}, \ldots, val_{v,n}, class_v)$ aus M folgt also aus $val_{u,j} = val_{v,j}$ für $1 \leq j \leq n$ auch sofort $class_u = class_v$.

Aber schon bei verrauschten Daten kann diese Bedingung meist nicht mehr eingehalten werden. Daher sollte ein Klassifikations erfahren geeignet mit inkonsistenten Beispielsammlungen umgehen können.

Ein Entscheidungsbaum *ist nun ein Baum, in dem jeder innere Knoten ein Attribut als Label hat, so daß auf keinem Pfad von der Wurzel zu einem Blatt ein Attribut mehrfach vorkommt, und in dem jede Kante, die von einem Knoten mit Label A_i zu einem Nachfolger des Knotens geht, eine Ausprägung $a_{i,j}$ dieses Attributes als Label erhält, und genau eine Kante das Label $*$ als "Sonst´´-Fall (für nicht vorkommende oder noch nicht bekannte Ausprägungen), so daß alle Kanten zu den Nachfolgern eines Knotens verschiedene Label aufweisen und die Kante mit dem Label $*$ zu einem Blattknoten führt. Die Blätter des Baumes erhalten als Label eine Klassifikation.*

Ein Entscheidungsbaum kann jedes Fallbeispiel mit beliebigen Ausprägungen für die Attribute klassifizie ren. Der Klassifikatio erfolgt mittels Durchlaufen des Entscheidungsbaumes entlang eines Pfades bis zu einem Blatt. Die Klassifikatio in dem Blattknoten wird als Klassifikatio des Fallbeispieles gewählt. Der zu durchlaufende Pfad ergibt sich aus den Attibutausprägungen des Falles: Die Ausprägung für das Attribut im Label des aktuellen Knotens (zu Anfang des Wurzelknotens) wird für das Fallbeispiel bestimmt und, falls vorhanden, eine Kante mit dieser Ausprägung vom aktuellen Knoten zum Nachfolgerknoten verfolgt, ansonsten wird die Kante mit dem Label $*$ benutzt.

Attribute, die nicht als Knotenlabel in dem Pfad auftreten, können eine beliebige Ausprägung haben; die Entscheidung ist von ihnen in diesem Entscheidungsbaum nicht abhängig. Die Blätter eines Entscheidungsbaumes müssen daher auch nicht unbedingt alle in gleicher Tiefe auftreten.

Wenn der Entscheidungsbaum in dieser Weise zur Klassifikatio benutzt wird, so wird jedes Fallbeispiel von genau einem Pfad entschieden. Dieser Pfad *repräsentiert* dieses Fallbeispiel. Für eine vorgegebene Sammlung von Fallbeispielen mit ihrer Klassifikatio soll natürlich ein Entscheidungsbaum gefunden werden, der diese Sammlung möglichst gut repräsentiert, d.h. im Falle einer konsistenten Beispielsammlung wird jeder Fall der Beispielsammlung korrekt klassifiziert

Wenn alle Attibutausprägungen bekannt sind und in der Beispielsammlung für jedes Tupel von Ausprägungen der Attribute ein Fallbeispiel vorliegt, d.h. $m = k_1 \cdot \ldots \cdot k_n$, können wir einen sehr einfachen Entscheidungsbaum generieren. Der Wurzelknoten (Knoten der Tiefe 0) erhält A_1 als Label und k_1 Nachfolgerknoten. Die Kanten der Knoten werden mit den Labeln $a_{1,1}, \ldots, a_{1,k_1}$ versehen. Die Knoten der Tiefe i mit $1 \leq i < n$ erhalten als Label alle das Attribut A_{i+1} und jeweils k_{i+1} Nachfolger mit den Kantenlabeln $a_{i+1,1}, \ldots, a_{i+1,k_{i+1}}$. Die Blätter erhalten als Label die Klassifikatio des Fallbeispiels, das durch die Label der Kanten auf dem Pfad von der Wurzel zu diesem Blatt beschrieben wird. Auf Kanten mit dem Label $*$ kann verzichtet werden.

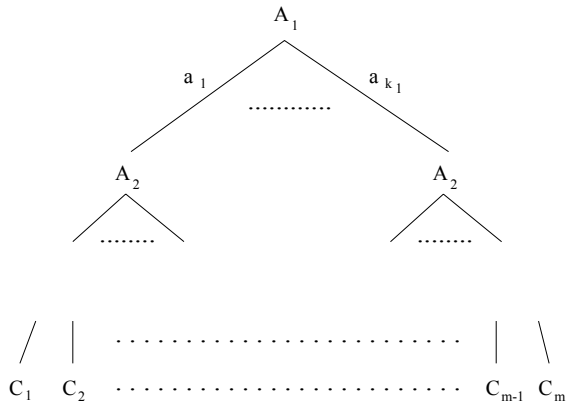

Abbildung 1: Entscheidungsbaum für vollständige Objektsammlung

Jeder Pfad beschreibt (hier genau) ein Beispiel der Objektsammlung; zu jedem Fallbeispiel gibt es einen Pfad, der das Beispiel darstellt. Der Entscheidungsbaum repräsentiert also die beschriebene Objektsammlung.

Die Größe des Entscheidungsbaumes und damit auch die Geschwindigkeit einer Entscheidung hängt von einer geschickten Gruppierung der Fallbeispiele mit gleicher Klassifikatio ab.

Beispiel: Es sei die folgende Fallbeispielsammlung für Pilze gegeben.

Farbe	Größe	Punkte	Klasse
rot	klein	ja	giftig
braun	klein	nein	eßbar
braun	groß	ja	eßbar
grün	klein	nein	eßbar
rot	groß	nein	eßbar

Wir nehmen an, daß keine anderen Attributausprägungen möglich sind.

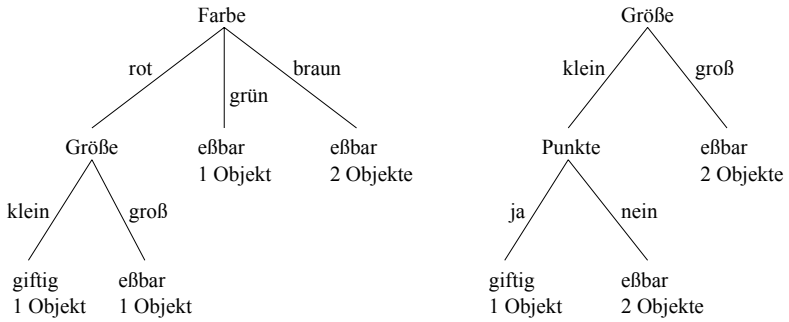

Abbildung 2: Entscheidungsbäume für Pilzbeispielsammlung

Wie man an den Entscheidungsbäumen des Beispiels erkennen kann, liegen im allgemeinen die Blätter nicht alle auf gleicher Ebene. Obwohl beide Bäume die gleiche Tiefe haben, werden im linken Baum die Beispiele im Schnitt früher entschieden als im rechten Baum. □

Um die Qualität eines Entscheidungsbaumes beurteilen zu können, benötigt man geeignete Maße. Die einfachsten Möglichkeiten wie Anzahl der Knoten des Baumes oder die Tiefe des Baumes reichen zur Beschreibung nicht aus, wie das vorstehende Beispiel zeigt.

Als Kriterien bieten sich an

- Anzahl der Blätter

  Die Anzahl der Blätter entspricht der Anzahl der Regeln, die man aus einem Entscheidungsbaum generiert.

- Höhe des Baumes

  Die Höhe des Baumes entspricht der maximalen Regellänge, d.h. der Anzahl der maximal zu verifi zierenden Prämissen für eine Entscheidung.

- externe Pfadlänge

  Die externe Pfadlänge ist definier als die Summe der Längen aller Pfade von der Wurzel zu einem Blatt. Dieses Maß gibt Auskunft über den Speicheraufwand für die aus dem Baum generierte Regelmenge.

- gewichtete externe Pfadlänge

  Die gewichtete externe Pfadlänge wird berechnet wie die externe Pfadlänge, außer das die Längen der Pfade jeweils multipliziert werden mit der Anzahl der Fallbeispiele, die dieser Pfad repräsentiert. Dieses Maß bewertet die Klassifikations osten, indem für jedes Beispiel die Länge der zu testenden Regel berücksichtigt wird.

Beispiel: Für die beiden Bäume aus dem Beispiel ergibt sich:

Kriterium	linker Baum	rechter Baum
Anz. Blätter	4	5
Höhe	2	2
ext. Pfadlänge	6	5
gew. ext. Pfadlänge	7	8

□

Zur Komplexität des Problems, einen Entscheidungsbaum zur effiziente Klassifikatio gemäß einer vorgegebenen Beispielsammlung zu finden läßt sich folgender Satz zeigen.

Satz 1 *Das Problem der Entscheidung, ob für eine beliebige Objektsammlung M ein Entscheidungsbaum $T(M)$ mit einer durch eine gegebene Konstante k beschränkten externen Pfadlänge existiert, ist NP-vollständig.*

ID3 - Ein heuristischer Algorithmus

Die Güte eines Entscheidungsbaumes hängt wesentlich von der geschickten Wahl der Attribute ab, nach denen verzweigt wird. Abgesehen davon läßt sich die Generierung eines Entscheidungsbaumes durch einen einfachen rekursiven Algorithmus bewerkstelligen:

1. Ist M trivial, d.h. enthält M nur Fallbeispiele mit einer Klassifikation so ist der Entscheidungsbaum ein Blattknoten mit der Klassifikatio der Beispiele in M als Label.

2. Nur bei inkonsistenten Beispielsammlungen:
   Ist die Attributmenge $\{A_1, \ldots, A_n\}$ leer, also alle Attribute bereits früher für Entscheidungen benutzt, so ist der Entscheidungsbaum ein Blattknoten mit der Klassifikatio der Mehrheit der Fallbeispiele in M als Label.

3. Ansonsten wähle (nach einer geeigneten Heuristik) ein Attribut $A_i \in \{A_1, \ldots, A_n\}$ und zerlege M in r disjunkte Teilmenge $M_1, \ldots, M_r$, wobei $a_{i,j_1}, \ldots, a_{i,j_r}$ die verschiedenen für das Attribut A_i vorkommenden Werte sind und M_j genau die Beispiele enthält, für die das Attribut A_i den Wert $a_{i,j}$ annimmt.

4. Entferne das Attribut A_i aus $\{A_1, \ldots, A_n\}$ und die Attributwerte aus den Beispielen in $M_1, \ldots, M_r$.

   (a) Für jedes nicht-leere M_j erzeuge einen Entscheidungsbaum durch rekursive Anwendung des Algorithmus auf M_j.

   (b) Für die Kante mit dem Label $*$ als "Sonst´´-Fall für nicht auftretende und noch unbekannte Attributausprägungen erzeuge als Entscheidungsbaum einen Blattknoten mit der Klassifikatio c_M der Mehrheit der Beispiele in M als Label.

Sind $T(M_1), \ldots, T(M_r)$ und $T(M_*)$ die generierten Entscheidungsbäume für $M_1, \ldots, M_r$, so bilde $T(M)$ wie in folgender Skizze:

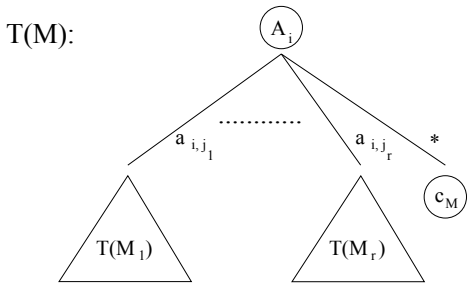

Abbildung 3: Rekursives Generieren des Entscheidungsbaumes

Die Anzahl der möglichen Entscheidungsbäume ist zu groß, als daß eine systematische Suche nach einem optimalen Entscheidungsbaum möglich wäre. Bei n Attributen mit je 2 Ausprägungen gibt es bereits mehr als $n!$ verschiedene Entscheidungbäume, die pro Blatt genau ein Beispiel repräsentieren (n Möglichkeiten für das erste Attribut und für jeden Teilbaum alle Möglichkeiten mit $n-1$ Attributen).

Daher muß man versuchen, mit geeigneten Heuristiken die Auswahl des nächsten Attributes so vorzunehmen, daß insgesamt ein möglichst optimaler Entscheidungsbaum generiert wird.

Die Attributauswahlheuristik ID3

Grundlage dieser Heuristik ID3 (Iterative Dichotomizer Version 3; Dichotomizer, da ursprünglich $C = \{c_1, c_2\}$) sind informationstheoretische Überlegungen. Dabei gehen wir von folgenden Voraussetzungen aus:

1. Ein Entscheidungsbaum für M soll Fallbeispiele entscheiden, deren relative Häufig eiten genauso verteilt sind wie für die Beispiele in M.
2. Ein Entscheidungsbaum kann als eine Informationsquelle aufgefaßt werden, dessen Äste jeweils die Information der Klassifikatio im Blatt des Astes darstellen.

Wenn wir nun eine große Gruppe von Beispielen in der Objektsammlung haben, für die die Klassifikatio gleich ist, so ist der Informationsgehalt eines solchen Beispiels geringer als bei einer Klassifikation für die nur ein Beispiel in der Sammlung existiert.

Definitio 1
Tritt ein Ereignis $\mathcal{E}$ mit Wahrscheinlichkeit $p(\mathcal{E}) > 0$ ein, dann ist der Informationsgehalt $I(\mathcal{E})$ *dieses Eintretens*

$$I(\mathcal{E}) = -\log(p(\mathcal{E}))$$

Wir betrachten im folgenden immer den Logarithmus zur Basis 2, d.h. $\log = \log_2$.

Da die Wahrscheinlichkeiten immer Werte im Intervall $[0, 1]$ sind, ist der Informationsgehalt für Ereignisse mit positiver Wahrscheinlichkeit definier und hat ebenfalls einen positiven Wert. Der Informationsgehalt eines Ereignisses ist also um so größer, je seltener das Ereignis ist. Ein absolut sicheres Ereignis hat den Informationswert 0.

Da die Häufig eit von Ereignissen variiert, betrachtet man den mittleren Informationsgehalt.

Definitio 2 (Entropie)
Hat ein Versuch $\mathcal{E}$ die r möglichen Ausgänge $\mathcal{E}_1, \ldots, \mathcal{E}_r$, so heißt der mittlere Informationsgehalt

$$H(\mathcal{E}) = -\sum_{i=1}^{r} p(\mathcal{E}_i) \log(p(\mathcal{E}_i))$$

die Entropie *des Versuchs.*

Die Entropie ist also ein Maß für die *Verunreinigung* einer Grundmenge, also für ihre Nicht-Uniformität.

Definitio 3 (Informationsgewinn)
Hat ein zweiter auf der gleichen Objektmenge durchgeführter Versuch $\mathcal{F}$ die s Ausgänge $\mathcal{F}_1, \ldots, \mathcal{F}_s$, dann ist die bedingte Entropie *des kombinierten Versuchs $(\mathcal{F}|\mathcal{E})$ gegeben durch*

$$H(\mathcal{F}|\mathcal{E}) = -\sum_{i=1}^{r} p(\mathcal{E}_i) \sum_{j=1}^{s} p(\mathcal{F}_j|\mathcal{E}_i) \log(p(\mathcal{F}_j|\mathcal{E}_i))$$

Der mittlere Informationsgewinn*, der durch den Versuch $\mathcal{E}$ entsteht, ist dann definiert als*

$$H(\mathcal{F}) - H(\mathcal{F}|\mathcal{E})$$

Der mittlere Informationsgewinn gibt die erwartete Reduktion der Entropie an, die aufgrund der Kenntnis der Ausgänge für $\mathcal{E}$ entsteht. Der Versuch $\mathcal{F}$ wird dabei auf den Gruppen von Objekten mit gleichem Ausgang bei Versuch $\mathcal{E}$ durchgeführt. Sind die Versuche $\mathcal{E}$ und $\mathcal{F}$ unabhängig in dem Sinne, daß der Versuch $\mathcal{F}$ auf den Gruppen von Objekten mit gleichem Ausgang bei Versuch $\mathcal{E}$ jeweils die gleiche Verteilung von Ergebnissen liefert wie auf der gesamten Objektmenge, so ergibt sich aus der Kenntnis der Ausgänge für $\mathcal{E}$ kein Vorteil, der mittlere Informationsgewinn hat den Wert 0.

Wenn dagegen die Gruppen von Objekten mit gleichem Ausgang bei Versuch $\mathcal{E}$ jeweils einen festen Ausgang bezüglich des Versuchs $\mathcal{F}$ liefern, so hat die bedingte Entropie $H(\mathcal{F}|\mathcal{E})$ den Wert 0, also ist der mittlere Informationsgewinn maximal.

In unserem Kontext der Entscheidungsbäume ist das Ereignis $\mathcal{E}$ das ausgewählte Attribut A_t, die Ausgänge des Versuchs sind die (zufälligen) Werte des Attributs $a_{t,1}, \ldots, a_{t,k_t}$, d.h. wir haben $r = k_t$ Ausgänge.

Der Versuch $\mathcal{F}$ ist die Klassifikatio der Objekte, wir haben daher die Ausgänge $c_1, \ldots, c_k$ und $r = k$.

> Attributauswahlheuristik ID3:
> Maximiere den mittleren Informationsgewinn, d.h. wähle als nächstes Attribut dasjenige mit höchsten mittleren Informationsgewinn.

Unter der Voraussetzung, daß die Beispielsammlung die Wahrscheinlichkeiten der Attributausprägungen und Klassifikatione korrekt widerspiegelt, können wir alle vorkommenden Wahrscheinlichkeitswerte über die relativen Häufig eiten berechnen. Wir nehmen weiter an, das jede Ausprägung eines Attributes auch tatsächlich in mindestens einem Beispiel auftritt. (# bezeichnet die Anzahl der Elemente einer Menge.)

$$\begin{array}{lcl} x_i & := & \#\{\text{ Objekte in } M \text{ mit Wert } a_{t,i} \text{ für } A_t\} \\ p_i & := & P(E_i) = \frac{x_i}{m} \\ x_{i,j} & := & \#\{\text{ Objekte in } M \text{ mit Wert } a_{t,i} \text{ für } A_t \text{ und Klassifikatio } c_j\} \end{array}$$

Bezeichnen wir $\mathcal{F}$ mit der Klassifikatio C und $\mathcal{E}$ mit dem Attribut A_t, läßt sich die bedingte Entropie berechnen als

$$H(C|A_t) = -\sum_{i=1}^{k_t} p_i \sum_{j=1}^{k} \frac{x_{i,j}}{x_i} \log(\frac{x_{i,j}}{x_i})$$

Mit der Festlegung $0 \log 0 := 0$ können den mittleren Informationsgewinn $H(C) - H(C|A_t)$ also einfach ausrechnen. Wollen wir das Attribut bestimmen, für den er maximal ist, so müssen wir das t mit minimalem Wert $H(C|A_t)$ bestimmen, da $H(C)$ konstant ist. Dieses Attribut liefert den höchsten mittleren Informationsgewinn.

Berechnungsvorschrift für Attributauswahlheuristik ID3:
Bestimme t mit $1 \leq t \leq n$ mit $H(C) - H(C|A_t)$ maximal, also $H(C|A_t)$ minimal, da $H(C)$ fest.

Beispiel: (Fortsetzung)
Wir betrachten die Beispielsammlung zur Pilzklassifikation Um das erste Attribut auszuwählen, wird zunächst für jedes Attribut die Matrix der bedingten Vorkommen $X = (x_{i,j})$ aufgestellt. Die Summen der Zeilen ergeben die Werte x_i. Anschließend werden die Werte für die bedingte Entropien bestimmt und das Attribut A_t mit minimalem Wert gewählt. Auf die Teil-Beispielsammlungen M_j mit festem Attributwert $a_{t,j}$ für dieses Attribut wird das Verfahren rekursiv angewandt.

Rekursionsebene I:

i. Farbe:

$$X = \begin{array}{c|c|c} & \text{giftig} & \text{eßbar} \\ \hline \text{rot} & 1 & 1 \\ \text{braun} & 0 & 2 \\ \text{grün} & 0 & 1 \end{array} \quad \Longrightarrow \quad \begin{array}{ll} x_{\text{rot}} & = 2 \\ x_{\text{braun}} & = 2 \\ x_{\text{grün}} & = 1 \end{array}$$

Es folgt wegen $m = 5$

$$p_{\text{rot}} = \frac{2}{5} = 0.4, \;\; p_{\text{braun}} = \frac{2}{5} = 0.4, \;\; p_{\text{grün}} = \frac{1}{5} = 0.2$$

und damit

$$\begin{array}{rcl} H(C|A_{\text{Farbe}}) & = & -[0.4(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) + 0.4(\frac{0}{2}\log\frac{0}{2} + \frac{2}{2}\log\frac{2}{2}) \\ & & \ldots + 0.2(\frac{0}{1}\log\frac{0}{1} + \frac{1}{1}\log\frac{1}{1})] \\ & = & -[0.4(-1) + 0 + 0] \\ & = & 0.4 \end{array}$$

(Beachte $2^{-1} = \frac{1}{2}$.)

ii. Größe:

$$X = \begin{array}{c|c|c} & \text{giftig} & \text{eßbar} \\ \hline \text{klein} & 1 & 2 \\ \text{groß} & 0 & 2 \end{array} \quad \Longrightarrow \quad \begin{array}{ll} x_{\text{klein}} & = 3 \\ x_{\text{groß}} & = 2 \end{array}$$

Also

$$p_{\text{klein}} = \frac{3}{5} = 0.6, \;\; p_{\text{groß}} = \frac{2}{5} = 0.4$$

und damit

$$\begin{aligned} H(C|A_{\text{Größe}}) &= -[0.6(\tfrac{1}{3}\log\tfrac{1}{3} + \tfrac{2}{3}\log\tfrac{2}{3}) + 0.4(\tfrac{0}{2}\log\tfrac{0}{2} + \tfrac{2}{2}\log\tfrac{2}{2})] \\ &= -[0.4(-\log 3 + \tfrac{2}{3}) + 0] \\ &\approx 0.4562 \end{aligned}$$

(Beachte $\log\frac{x}{y} = \log x - \log y$ und $\log 2 = 1$ sowie $\log 1 = 0$.)

iii. Punkte:

$$X = \begin{array}{|c|c|c|} \hline & \text{giftig} & \text{eßbar} \\ \hline \text{ja} & 1 & 1 \\ \text{nein} & 0 & 3 \\ \hline \end{array} \quad \Longrightarrow \quad \begin{aligned} x_{\text{ja}} &= 2 \\ x_{\text{nein}} &= 3 \end{aligned}$$

Also

$$p_{\text{ja}} = \frac{2}{5} = 0.4, \;\; p_{\text{nein}} = \frac{3}{5} = 0.6$$

und damit

$$\begin{aligned} H(C|A_{\text{Punkte}}) &= -[0.4(\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{2}\log\tfrac{1}{2}) + 0.6(\tfrac{0}{3}\log\tfrac{0}{3} + \tfrac{3}{3}\log\tfrac{3}{3})] \\ &= -[0.4(-1) + 0] \\ &= 0.4 \end{aligned}$$

Damit ist die bedingte Entropie minimal für das Attribut Farbe und auch für das Attribut Punkte. Wir wählen das Attribut Punkte aus, da es weniger Teilprobleme erzeugt. Außerdem können wir auf die Kante mit dem Label $*$ als "Sonst´´-Fall verzichten, da beide Attributsausprägungen auftreten und keine zusätzliche Ausprägung zu erwarten ist.

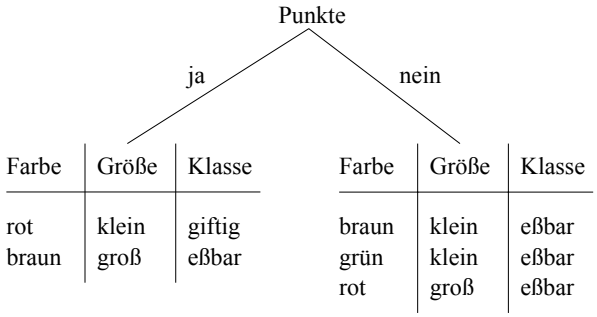

Abbildung 4: Entscheidungsbaum nach Rekursionsebene I

Rekursionsstufe II:

(a) Objektsammlung M_{nein} besteht aus 3 Beispielen, die alle in die Klasse der eßbaren Pilze gehören. Also ist M_{nein} trivial und der Entscheidungsbaum $T(M_{\text{nein}})$ besteht nur aus einem Blattknoten mit Label „eßbar".

(b) Objektsammlung M_{ja} besteht aus 2 Beispielen mit verschiedenen Klassifikationen

i. Farbe:
Das Attribut Farbe unterscheidet die Beispiele, daher könnte dieses Attribut sofort gewählt werden.

ii. Größe:
Das Attribut Größe unterscheidet die Beispiele ebenfalls, daher könnte auch dieses Attribut sofort gewählt werden.

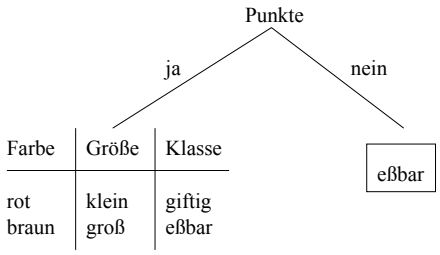

Abbildung 5: Entscheidungsbaum nach Rekursionsebene II, Teil 1

Die Werte für die bedingte Entropie sind offensichtlich gleich, da in beiden Fällen analoge Situationen vorliegen. Wir wählen als Attribut die Farbe.

Die Farbe grün tritt al Beispiel gar nicht auf. Die Menge der Fallbeispiele enthält gleichviele Beispiele der Klassen „eßbar" und "giftig´´. Wir entscheiden uns für den „Sonst"-Fall daher für die Klassifikatio "giftig´´.

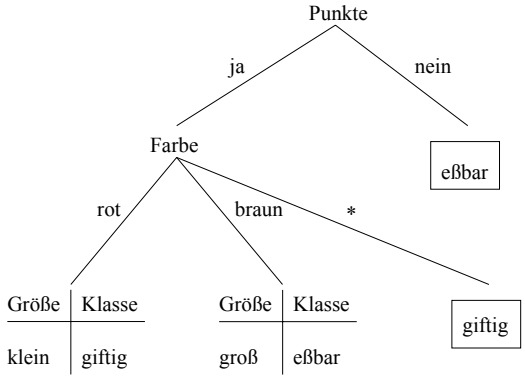

Abbildung 6: Entscheidungsbaum nach Rekursionsebene II

Rekursionsstufe III:

(a) Objektsammlung $M_{\text{ja,rot}}$

(b) Objektsammlung $M_{\text{ja,braun}}$

Da jeweils nur noch ein Beispiel übrig ist, sind die Objektsammlungen trivial, die Entscheidungsbäume bestehen nur aus den Blättern mit Label giftig bzw. eßbar.

□

Bemerkungen:

1. Der Entscheidungsbaum generalisiert die Klassifikatio von der Menge der Fallbeispiele auf den allgemeinen Fall dadurch, daß alle Attribute eine Kante mit Label $*$ als "Sonst´´-Fall für nicht auftretende und noch unbekannte Attributausprägungen und ein zugehöriger Blattknoten erzeugt werden, der nach der Mehrheit der Klassifikatione in der aktuellen Beispielmenge entscheidet.

2. Die Heuristik kann mit inkonsistenten Beispielmengen umgehen, da im Falle uneinheitlicher Klassifikatio bei ansonsten gleichen Attributausprägungen nach der Mehrheit der Fälle entschieden wird.

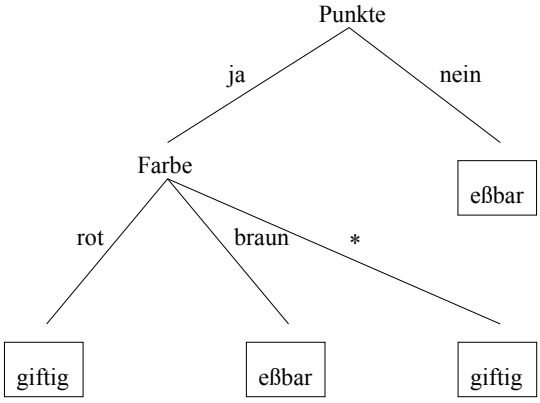

Abbildung 7: Entscheidungsbaum nach Rekursionsebene III

Daher sollte die Beispielsammlung auch nicht als Menge im mathematischen Sinne gesehen werden, sondern als Liste mit eventuellen Mehrfachvorkommen.

3. Die Heuristik braucht Attribute, für die keine Beispiele mit verschiedenen Ausprägungen in der Objektsammlung vorhanden sind, bei der Auswahl nicht zu berücksichtigen. Wie man leicht nachrechnet ergibt die bedingte Entropie den Wert 0 für diesen Fall. Wenn nur noch ein Attribut übrig ist, braucht natürlich ebenfalls keine Berechnung zu erfolgen, wir wählen einfach dieses Attribut.

4. Der Entscheidungsbaum wird nur dann *von hoher Güte* sein, wenn die Beispielsammlung repräsentativ ist für die anschließend anhand des Entscheidungsbaumes vorgenommenen Klassifikationen

Komplexität

Es ist leicht zu sehen, daß für alle benötigten Hilfsoperationen wie Test auf triviale Objektsammlung, Berechnung der vorkommenden Attributausprägungen, Berechnung der reduzierten Objektsammlungen, etc. ein Zeitaufwand von $O(mn)$ pro Berechnung ausreicht.

Offensichtlich ist die Anzahl der rekursiven Aufrufe des Verfahrens maximal, wenn der Entscheidungsbaum die Tiefe n hat und auch alle Blätter in der Ebene n liegen. Es zeigt sich, daß in diesem Fall die Rechenzeit für die Konstruktion des gesamten Baumes (ohne Berücksichtigung der Attributauswahlheuristik) in $O(mn^2)$ liegt. Unter Berücksichtigung der Heuristik ergibt sich eine Gesamtkomplexität von $O(mn^2 + nf_h(m,n))$ unter der Annahme, daß f_h (Zeitschranke für die Berechnung des nächsten Attributes bei m Beispielen und n Attributen) für die Heuristik h ein Polynom in m und n ist.

Für die ID3 Heuristik zeigte Quinlan die Komplexität $O(mn)$ und eine Gesamtkomplexität von $O(mn\alpha)$, wobei α die Anzahl der inneren Knoten während der Konstruktion des Baumes ist. Im beschriebenen schlechtesten Fall liegt α in der Größenordnung von m, so daß sich wegen $m > n$ eine Gesamtkomplexität von $O(m^2n)$ ergibt. Da aber üblicherweise $m \gg n$ gilt, erscheint diese Schranke als zu hoch. Setzen wir die Komplexität $O(mn)$ für ID3 in die Gesamtkomplexität ein so erhalten wir $O(mn^2)$ als Komplexität für das Verfahren mit der Heuristik ID3.