Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Rainer Höhne

Data Mining

Entscheidungsbäume

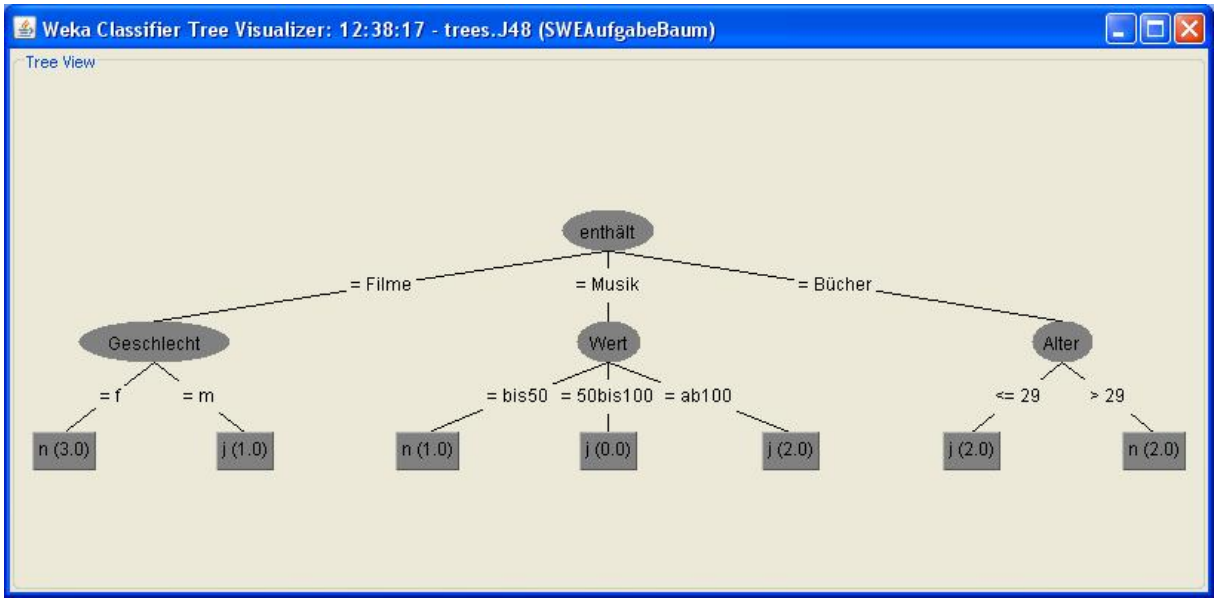
Ein Ziel beim Data Mining ist, aus vorhandenen Daten Regeln abzuleiten, mit denen Voraussagen getroffen werden können. Nehmen wir als Beispiel eine Internet-Buchhandlung, die mit Büchern, Filmen auf DVD und Musik auf DVD oder CD handelt. Folgekäufe sollen dadurch initiiert werden, das den Kunden 5 € Gutscheine zugesandt werden, die sie bei ihrem nächsten Einkauf einlösen können. Natürlich wäre es Geldverschwendung, denjenigen Kunden, die ohnehin wieder kaufen, einen Gutschein zukommen zu lassen. Deshalb sollen die Kunden identifiziert werden, die wahrscheinlich *nicht* innerhalb der nächsten drei Monate wieder bestellen.

Nehmen wir an, der Buchhandlung liegen die folgenden historischen Daten der Kunden vor:

Geschlecht	**Alter**	**Ort Kunde**	**Sendung innerhalb von 3 Tagen ausgeliefert**	**Sendung enthält hauptsächlich**	**Gesamtwert der Sendung (€)**	**Folgekauf innerhalb von 3 Monaten**
f	17	Dorf	n	Filme	bis50	n
f	29	Dorf	n	Filme	50bis100	n
m	42	Dorf	j	Musik	ab100	j
f	24	Stadt	n	Bücher	bis50	j
m	32	Dorf	j	Bücher	50bis100	n
m	47	Stadt	n	Musik	bis50	n
m	62	Stadt	n	Musik	ab100	j
f	26	Dorf	j	Bücher	50bis100	j
f	42	Dorf	j	Bücher	ab100	n
f	33	Stadt	j	Filme	bis50	n
m	18	Stadt	n	Filme	50bis100	j

Entscheidungsbäume sind eine Möglichkeit, eine Regel zu gewinnen und darzustellen, die dann auf neue Datensätze angewendet wird, um eine Voraussage zu treffen - im Beispiel würde bestimmt werden, ob ein Kunde mit einem bestimmten Profil innerhalb von drei Monaten wieder einkaufen würde.

Ein Entscheidungsbaum für dieses Problem könnte so aussehen (hier mit dem Tool weka erzeugt):



Um den besten Entscheidungsbaum zu bestimmen kann das Konzept der *Entropie* verwendet werden.

Wenn n Werte mit den Wahrscheinlichkeiten $p_1, p_2, ..., p_n$ auftreten, ist die Entropie definiert durch $\sum_{i=1}^{n}(-p_i log(p_i))$, wobei *log* der Logarithmus zur Basis 2 ist. Schritt für Schritt wird jeweils das Attribut ausgewählt, das die minimale gewichtete Entropie erzeugt. Auf diese Weise ergibt sich ein Baum, dessen Knoten so weit wie möglich nur noch Elemente der selben Klasse (also Objekte, die im Wert des Zielattributs übereinstimmen) enthalten.

Würde man zum Beispiel im ersten Schritt das Attribut "Geschlecht" wählen, ergäbe sich ein Baum mit zwei Knoten, der eine mit den weiblichen Kunden (insgesamt 6, 2 Folgekäufer, 4 keine Folgekäufer) der andere mit den männlichen (insgesamt 5, 3 Folgekäufer, 2 keine Folgekäufer), so daß $\frac{6}{11}(-\frac{2}{6}log(\frac{2}{6}) - \frac{4}{6}log(\frac{4}{6})) + \frac{5}{11}(-\frac{3}{5}log(\frac{3}{5}) - \frac{2}{5}log(\frac{2}{5}))$ die gewichtete Entropie ist.

Bei stetigen Attributen wird prinzipiell nur binär verzweigt. Die minimale gewichtete Entropie bezüglich aller möglichen Klasseneinteilungen entsprechend dem Mittelwert zweier Attributwerte wird verwendet. Im Beispiel des Attributs Alter wären das die Klassen

Alter	**Klasse**
17	
	$\leq 17,5, > 17,5$
18	
	≤ 21,> 21
24	
	≤ 25,> 25
26	
	$\leq 27,5, > 27,5$
29	
	$\leq 30,5, > 30,5$
32	
	$\leq 32,5, > 32,5$
33	
	$\leq 37,5, > 37,5$
42	
	$\leq 44,5, > 44,5$
47	
	$\leq 54,5, > 54,5$
62	

Aufgabe

Erstellung eines Programms, das den Algoruthmus, einen "optimalen" Entscheidungsbaum zu finden, demonstriert. Weil das Programm insbesondere für Demonstrationszwecke eingesetzt werden soll muss insbesondere auch auf eine entsprechende Gestaltung der Oberfläche (Verwendug von aussagekräftigen Farben, genügend großer Schriftgrad, etc.) geachtet werden. Das Programm soll unter allen gängigen Windows-Betriebssystemen ohne Installation lauffähig sein (natürlich darf Java vorausgesetzt werden, sonst aber nichts spezielles).

Programm

Im Programm kann zwischen drei verschiedenen Ansichten hin- und hergewechselt werden, von denen stets genau eine sichtbar ist:

- Tabellenansicht
- Baum interaktiv
- Baum automatisch

Aus den beiden Baumdarstellungen kann die zusätzliche

- Regeldarstellung

aufgerufen werden.

Start ist stets in der Tabellenansicht, erst wenn eine Tabelle geladen oder eingegeben worden ist kann in eine Baumansicht gewechselt werden. Ansonsten kann stets zwischen den drei Ansichten gewechselt werden.

Tabellenansicht

- Die Tabelle (in der Form des obigen Beispiels) besteht aus maximal etwa 16 Attributen und 1000 Objekten
- Das Programm erkennt automatisch, ob die Werte eines Attributes stetig (also Zahlen) oder diskret sind
- Dezimaltrenner ist der Punkt
- Einlesen / Speichern von Tabellen im .csv-Format (Trennzeichen ist das Komma)
- Eingeben / Editieren der Tabelle
- (Farbliche) Markierung des Zielattributs. Zielattribut ist standardmäßig das letzte, das kann aber geändert werden
- Import und Export von Excel-Dateien

Baum interaktiv

- Zu Beginn wird nur der oberste Knoten (also der, der allen Objekten in der Tabelle entspricht) dargestellt
- Das gilt ebenso, wenn in der Tabellenansicht ein Attribut hinzugefügt oder gelöscht oder das Zielattribut geändert wurde
- In jedem Knoten ist die Zahl der enthaltenen Objekte, die Zahl der Objekte jeder Klasse und die Entropie angegeben

- Die Kanten sind bei diskreten Attributen mit dem jeweiligen Attributwert, bei stetigen Attributen mit der entsprechenden Fallunterscheidung markiert
- Bei Identifikation eines Knotens des Baums wird eine Tabelle mit den Objekten, die von diesem Knoten repräsentiert werden, dargestellt
- Wird in dieser Tabelle ein Attribut identifiziert und im Falle eines stetigen Attributs zusätzlich ein Split-Wert eingegeben, wird die gewichtete Entropie, die sich bei Aufteilung nach diesem Attribut ergeben würde, ausgegeben
- In dieser Tabelle kann ein Attribut markiert (und ggf. ein Split-Wert eingegeben) werden so, dass entsprechend den Werten dieses Attributs bzw. des Split-Wertes Unterknoten entstehen. Die Tabelle verschwindet. Hatte der Knoten schon Unterknoten, verschwinden diese natürlich ebenfalls

Baum automatisch

- Der optimale Entscheidungsbaum wird mit dem Algorithmus, der sukzessive jeweils das Attribut auswählt, das die minimale gewichtete Entropie ergibt, erzeugt und dargestellt
- Wenn ein Knoten nur noch eine (einstellbare, Voreinstellung ist 1) Anzahl von Objekten repräsentiert wird nicht weiter aufgeteilt
- Wenn ein Knoten nur noch Objekte einer Klasse enthält wird ebenfalls nicht weiter aufgeteilt
- In jedem Knoten ist die Zahl der enthaltenen Objekte, die Zahl der Objekte jeder Klasse und die Entropie angegeben
- Die Kanten sind bei diskreten Attributen mit dem jeweiligen Attributwert, bei stetigen Attributen mit der entsprechenden Fallunterscheidung markiert
- Bei Identifikation eines Knotens des Baums wird eine Tabelle mit den Objekten, die von diesem Knoten repräsentiert werden, dargestellt

Regeldarstellung

- Visualisierung der Regel, die dem aktuellen Baum entspricht
- In den Knoten wird nur der Wert des Zielattributs (wenn der Knoten nicht rein ist, der häufigste Wert (wenn dieser nicht eindeutig bestimmt ist, ein aus den häufigsten Werten zufällig bestimmter)) dargestellt
- Die Kanten sind bei diskreten Attributen mit dem jeweiligen Attributwert, bei stetigen Attributen mit der entsprechenden Fallunterscheidung markiert
- Ein Objekt kann eingegeben werden (ohne Zielattributwert). Visualisiert wird, wie die Klassenzuordnung des Objekts entsprechend den Regeln, die der Baum repräsentiert, von der Wurzel bis zum Blatt abläuft