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

RAINER HÖHNE

Data Mining

Entscheidungsbäume

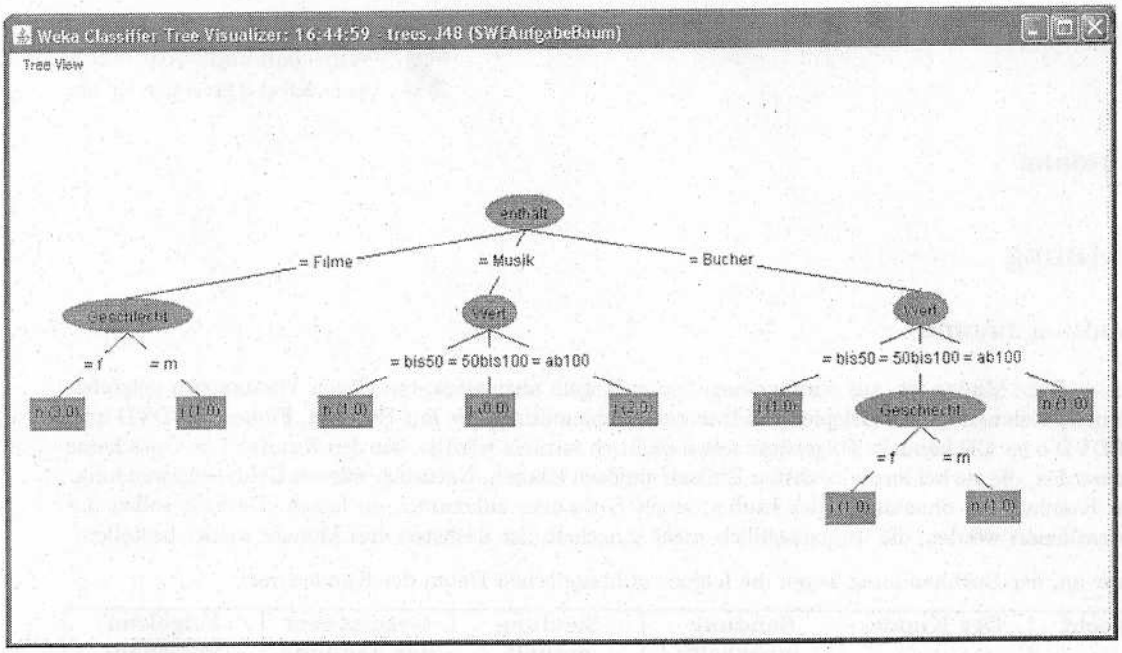
Ein Ziel beim Data Mining ist, aus vorhandenen Daten Regeln abzuleiten, mit denen Voraussagen getroffen werden können. Nehmen wir als Beispiel eine Internet-Buchhandlung, die mit Büchern, Filmen auf DVD und Musik auf DVD oder CD handelt. Folgekäufe sollen dadurch initiiert werden, das den Kunden 5 € 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	**Ort Kunde**	**Sendung innerhalb von 3 Tagen ausgeliefert**	**Sendung enthält hauptsächlich**	**Gesamtwert der Sendung (€)**	**Folgekauf innerhalb von 3 Monaten**
f	Dorf	n	Filme	bis50	n
f	Dorf	n	Filme	50bis100	n
m	Dorf	j	Musik	ab100	j
f	Stadt	n	Bücher	bis50	j
m	Dorf	j	Bücher	50bis100	n
m	Stadt	n	Musik	bis50	n
m	Stadt	n	Musik	ab100	j
f	Dorf	j	Bücher	50bis100	j
f	Dorf	j	Bücher	ab100	n
f	Stadt	j	Filme	bis50	n
m	Stadt	n	Filme	50bis100	j

ENTSCHEIDUNGSBÄUME sind nun 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 verwendet man das Konzept der *Entropie*: 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 Attribut mit der minimalen gewichteten Entropie ausgewählt. Auf diese Weise wird ein Baum erzeugt, in dessen Knoten so weit wie möglich nur noch Elemente der selben Klasse (also Objekte, die im Wert des Zielattributs übereinstimmen) enthalten sind. 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.

Aufgabe

Erstellung eines Programms, dass den Algorithmus, einen "optimalen" Entscheidungsbaum zu finden, demonstriert. Weil das Programm insbesondere für Präsentationszwecke eingesetzt werden soll, ist insbesondere auch auf eine entsprechende Gestaltung der Oberfläche (aussagekräftige Verwendung von Farben, genügend großer Schriftgrad, etc.) zu achten. Das Programm soll unter allen gängigen Windows-Betriebssystemen ohne Installation lauffähig sein (Java darf vorausgesetzt werden, sonst nichts spezielles).

Programm

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

- Tabellenansicht
- Baum interaktiv
- Baum automatisch

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

Tabellenansicht

- Eine Tabelle (in der Form des obigen Beispiels) besteht aus maximal etwa 10 Attributen und 100 Objekten.
- Einlesen / Speichern einer Tabelle im .csv-Format (wobei das Trennzeichen wirklich ein Komma ist, das soll hier schließlich kein Microsoft-Produkt werden ...).
- Alle Attributwerte sind Zeichenketten.
- Eingeben / Editieren einer 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 mit dem entsprechenden Attributwert 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, wird die gewichtete Entropie, die sich bei Aufteilung nach diesem Attribut ergeben würde, ausgegeben.
- In dieser Tabelle kann ein Attribut selektiert werden so, dass entsprechend den Werten dieses Attributs 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 Objekte einer Klasse enthält wird nicht weiter aufgeteilt.
- Wenn ein Knoten nur noch eine (einstellbare, Voreinstellung ist 1) Anzahl von Elementen enthält wird 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 mit dem entsprechenden Attributwert markiert.
- Bei Identifikation eines Knotens des Baums wird eine Tabelle mit den Objekten, die von diesem Knoten repräsentiert werden, dargestellt.