Duale Hochschule Baden - Württemberg Mannheim

Seminararbeit

Entscheidungsbäume

Studiengang Angewandte Informatik

Studienrichtung Informatik

Autor:	Martin Pretz
Matrikelnummern:	7060026
Kurs:	TINF18AI1
Bearbeitungszeitraum:	19.05.2021 - 10.06.2021

1 Abstract

2 Einführung

3 Was sind Entscheidungsbäume?

Bei Entscheidungsbäumen handelt es sich um eine bestimmte Form von Klassifikationsalgorithmen.

3.1 Motivation und Ziel

3.2 Generischer Aufbau

Im wesentlichen bestehen Entscheidungsbäume aus Knoten und Kanten. Bei einem Knoten handelt es sich um ein zu prüfendes Attribut während es sich bei einer Kante um das Ergebnis dieser Überprüfung handelt. [1] Darüber hinaus können Knoten wiederrum in Entscheidungsknoten, Wahrscheinlichkeitsknoten und Endknoten unterteilt werden.

Entscheidungsbäume bestehen im wesentlichen aus den vier Bestandteilen Wurzel, Knoten, Kante und Blatt. Bei der Wurzel handelt es sich im Grunde um einen Knoten. Bei einem Blatt handelt es sich um eine

4 Der ID3 Algorithmus

Bei ID3 (Iterative Dichotomiser 3) handelt es sich um einen Algorithmus zur Erstellung eines Entscheidungsbaumes welcher von Ross Quinlan entwickelt wurde. [2]

4.1 Funktionsweise

Der ID3-Algorithmus macht sich zwei Konzepte der Informationstheorie zu nutze. Es handelt sich zum einen um die Entropie und zum anderen um den Informationsgewinn. Beide Konzepte werden im nachfolgenden erläutert. Im Anschhluss daran wird der eigentliche ID3 Algorithmus erläutert.

4.1.1 Entropie

In der Informationstheorie wird mit der Entropie H die Sicherheit bzw. Unsicherheit einer Variablen X angegeben. Dementsprechend ist x_i eine mögliche Ausprägung der Variablen X und $P(x_i)$ die Wahrscheinlichkeit mit der die Variable X die Ausprägung x_i hat. [3]

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

Abbildung 4.1: Definition der Entropie nach Shennon[3]

Beispiel: Sei D ein Datensatz in dem das Attribut X mit den möglichen Ausprägungen x_1, x_2 und x_3 vorkommt. Weiterhin gelte, dass x_1 neun mal, x_2 drei mal und x_3 fünf mal in D vorhanden ist. Zur Bestimmung der Entropie von X egribt sich die nachfolgende Berechnung. Für den ID3-Algorithmus wird üblicherweise der Logarithmus zur Basis b=*2* verwendet. [4]

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{3} P(x_i) \log_2 P(x_i) \\
&= -(P(x_1) \log_2 P(x_1) + P(x_2) \log_2 P(x_2) + P(x_3) \log_2 P(x_3)) \\
&= -\left(\frac{9}{17} \cdot \log_2 \left(\frac{9}{17}\right) + \frac{3}{17} \cdot \log_2 \left(\frac{3}{17}\right) + \frac{5}{17} \cdot \log_2 \left(\frac{5}{17}\right)\right) \\
&\approx 0,485755 + 0,441618 + 0,519275 \\
&\approx \underline{\underline{1,446648}}
\end{aligned}
$$

Abbildung 4.2: Exmeplarische Berechung der Entropie

4.1.2 Informationsgewinn

In der Informationstheorie beschreibt der Informationsgewinn IG das Maß an Informationen das über eine Zufallsvariablen X durch Beobachtung einer anderen Zufallsvariablen Y gewonnen werden kann. [1] Konkret ergibt sich der Informationsgewinn aus der Differenz der Entropie $H(X)$ und der bedingten Entropie $H(X|Y)$. [5]

$$
IG(X,Y) = H(X) - H(X|Y) = H(X) - \sum_{y \in Y} P(Y = y) H(X|Y = y)
$$

Abbildung 4.3: Allgemeine Definition des Informationsgewinns [1, 6, 7]

Beispiel: Sei S ein Datensatz mit den in Tabelle 4.1 dargestellten Werten. Außerdem seien A, B, C und T Attribute von S mit den möglichen Ausprägungen *True* und *False*. Sei weiterhin das Attribut T das Zielattribut gegen das der Informationsgewinn der übrigen Attribute ermittelt werden soll.

ID	Attribut A	Attribut B	Attribut C	Attribut T
1	True	True	True	False
2	True	False	True	True
3	False	False	True	True
4	False	True	True	False
5	False	True	False	True

Tabelle 4.1: Beispiel Datensatz S [5]

Im nachfolgenden wird der Informationsgewinn für das Attribut A berechnet. Dabei wird zu erst die Entropie des Datensatz S bestimmt, welche durch die Entropie des Zielattributes T charakterisiert wird. Es gilt also $H(S) = H(T)$.

$$\begin{aligned} H(S) = H(T) &= -\sum_{i=1}^{2} P(x_i) \log_2 P(x_i) \\ &= -(P(\text{True}) \log_2 P(\text{True}) + P(\text{False}) \log_2 P(\text{False})) \\ &= -\left(\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) + \frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right)\right) \\ &\approx 0,970951 \end{aligned}$$

Als nächstes müssen die bedingte Entropie für das Attribut A berechnet werden. [7, 6]

$$\begin{aligned} H(T|A) &= \sum_{a \in A} P(A = a) \cdot H(T|A = a) \\ &= P(A = True) \cdot H(T|A = True) + P(A = False) \cdot H(T|A = False) \\ &= \frac{2}{5} \cdot \left(-\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right) + \frac{3}{5} \cdot \left(-\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right)\right) \\ &= \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot 0,918296 \\ &\approx 0.950978 \end{aligned}$$

Nachdem nun sowohl die Entropie als auch die bedingte Entropie berechnet sind kann final der Informationsgewinn bestimmt werden.

$$\begin{aligned} IG(T, A) &= H(T) - H(T|A) \\ &= 0,970951 - 0.950978 \\ &= \underline{\underline{0,019973}} \end{aligned}$$

Analog können die Informationsgewinne für die Attribute B und C berechnet werden. Diese liegen bei $IG(T, B) = 0,419973$ und bei $IG(T, C) = 0,170951$.

4.1.3 Eigentlicher Algorithmus

Der ID3 Algorithmus startet mit einem Datensatz S mit n Objekten welche verschiedenen Attribute A und eine Klassifizierung C besitzen. [4, 2] Die Werte der Attribute $A_1, A_2, ..., A_k$ sind dabei normalerweise endlich und diskret. [8]
Zu Beginn muss zunächst ein Wurzelknoten bestimmt werden. Dazu wird der Informationsgewinn $IG(S)$ für jedes Attribut $A_1, A_2, ..., A_k$ von S berechnet. Das Attribut mit dem höchsten Informationsgewinnungswert wird dann als Wurzelknoten gewählt. Basierend auf dem ausgewählten Attribut wird S aufgeteilt.

4.1.4 Eigenschaften von ID3

Der ID3-Algorithmus führt eine Best-First-Search für lokale Optima durch. Außerdem ist ID3 ein gieriger Algorithmus, da er stets das Attribut auswählt welches lokal den besten Informationsgewinn aufweist.

4.2 Datensatz

Für diese Arbeit wurde ein Datensatz verwendet welcher auf "RiskSample.csv" basiert. [9] In diesem Datensatz werden verschiedene Attribute im Zusammenhang mit einer Kreditvergabe erfasst. Das Ziel ist es anhand von bestimmten Attributen das Risiko zu klassifizieren welches in dem Attribut *RISK* erfasst wird. Dabei wird zwischen hohem Risiko *good risk*, schlechtem Profit *bad profit* und schwerem Verlust *bad loss* unterschieden. Der Datensatz ist exemplarisch in Auszügen in Tabelle 4.3 dargestellt.

AGE	INCOME	NUMKIDS	MORTGAGE	LOANS	RISK
45	58381	1.0	Yes	0.0	good risk
38	55752	0.0	Yes	1.0	good risk
34	55497	1.0	Yes	1.0	bad profit
42	55140	1.0	Yes	0.0	good risk
38	52887	0.0	Yes	0.0	good risk
37	52545	1.0	Yes	0.0	good risk
45	50552	0.0	Yes	0.0	good risk
40	50199	0.0	Yes	0.0	bad profit
42	49485	1.0	Yes	1.0	good risk
39	49415	0.0	Yes	0.0	good risk

Tabelle 4.2: Auszug aus dem originalen Datensatz mit ausgewählten Attributen

Der originale Datensatz erfasst insgesamt 11 Attribute. Diese sind *AGE*, *INCOME*, *GENDER*, *MARTIAL*, *NUMKIDS*, *NUMCARDS*, *HOWPAID*, *MORTGAGE*, *STORECAR*, *LOANS* und *ID*. Der Beispiel-Datensatz für den ID3 Algorithmus berücksichtigt davon nur noch fünf Attribute, nämlich *AGE*, *INCOME*, *NUMKIDS*, *MORTGAGE* und *LOANS*.

4.2.1 Transformation

Bevor die Daten verwendet werden, müssen sie zunächst eine Transformation durchlaufen, wobei diese "bereiningt" werden. Im nachfolgenden werden daher die Transformationen der betroffenen Attribute dargelegt.

Bei diesem Attribut *AGE* handelt es sich um das Alter einer Person welches im originalen Datensatz als Integer vorliegt. Im Zuge der Diskretisierung dieses Attributes wird das Alter in drei Kategorien eingeteilt. Dies sind *Young* (unter 30 Jahren), *Middle* (zwischen 30 und 50 Jahren) und *Old* (über 50 Jahre). Hierbei ist zu beachten dass das Alter im Datensatz lediglich zwischen minimal 18 und maximal 60 Jahren liegt.

Das Attribut *INCOME* liegt in originalen Datensatz als Integer vor und beziffert das jährliche Einkommen einer Person. Auch dieses Attribut wird diskretisiert und in vier Kategorien eingeteilt. Diese sind *Low* (unter 20.000 Euro), *Middle* (zwischen 20.000 und 30.000 Euro), *High* (zwischen 30.000 und 50.000 Euro) und *Very High* (über 50.000 Euro).

NUMKIDS erfasst im originalen Datensatz die Anzahl der Kinder einer Person. Allerdings wird dies im Ziel-Datensatz nicht länger berücksichtigt. Stattdessen gibt es nur eine Unterscheidung ob eine Person ein Kind hat oder nicht, also zwischen den beiden Zuständen *Yes* (Person hat Kinder) und *No* (Person hat keine Kinder).

Im originalen Datensatz wird mit dem Attribut *LOANS* die Anzahl der Darlehen erfasst während in dem transformierten Datensatz nur das Vorhandensein eventueller Darlehen, also nur die Zustände *Yes* (Person hat bereits Darlehen) oder *No* (Person hat aktuell kein Darlehen) erfasst werden.

4.2.2 Finaler Datensatz

Nachdem alle Transformation duchgeführt wurden, ergibt sich für die Tabelle 4.3 nun folgende Struktur.

AGE	INCOME	NUMKIDS	MORTGAGE	LOANS	RISK
Old	Very High	Yes	Yes	No	good risk
Middle	Very High	No	Yes	Yes	good risk
Middle	Very High	Yes	Yes	Yes	bad profit
Old	Very High	Yes	Yes	No	good risk
Middle	Very High	No	Yes	No	good risk
Middle	Very High	Yes	Yes	No	good risk
Old	Very High	No	Yes	No	good risk
Middle	Very High	No	Yes	No	bad profit
Old	Very High	Yes	Yes	Yes	good risk
Middle	Very High	No	Yes	No	good risk

Tabelle 4.3: Auszug aus dem transformierten Datensatz

4.3 Persönliche Implementation

Die Implementierung besteht im wesentlichen aus drei Bestandteilen. Zum einen aus der Berechnung des Informationsgewinns, welcher sich wiederrum aus der Berechnung der Entropie zusammensetzt. Zum anderen besteht die Implementation aus dem eigentlichen ID3 Algorithmus sowie anderen

```
def calculate_entropy(attribute):
    entropy = 0
    _, count = np.unique(attribute, return_counts=True)
    for index in range(len(count)):
        probablility = count[index] / sum(count)
        entropy += (-probablility * np.log2(probablility))
    return entropy
```

5 Zusammenfassung

Literatur

[1] M. Schinck, *Data Mining Vorlesung 2: Classification 1, Einführung, Validierung und Decision Trees*, Mannheim, 2021.

[2] J. Quinlan, *Induction of Decision Trees*, 01.08.1986.

[3] "Entropy (information theory) - Wikipedia." (), Adresse: https://en.wikipedia.org/wiki/Entropy_(information_theory)#cite_note-shannonPaper1-1.

[4] S. V. Rupali Bhardwaj, *Implementation of ID3 Algorithm*, CSE, Bahra Univerity India, 2013.

[5] *Information gain in decision trees - Wikipedia*, https://en.wikipedia.org/wiki/Information_gain_in_decision_trees, (Accessed on 05/27/2021).

[6] *Conditional entropy - Wikipedia*, https://en.wikipedia.org/wiki/Conditional_entropy, (Accessed on 05/27/2021).

[7] "Bedingte Entropie – Wikipedia." (), Adresse: https://de.wikipedia.org/wiki/Bedingte_Entropie.

[8] K. JEARANAITANAKIJ, *Classifying Continuous Data Set by ID3 Algorithm*.

[9] M. Schinck, *Datamining Vorlesung, RiskSample.csv*, Mannheim, 2021.