Fachgruppe Informatik
Prof. Dr. A. Ittner, C. Menzer
Wintersemester 2020/21
IF18/MI18

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Entscheidungsbaumlernen I

Aufgabe 1

Lesen Sie die Seiten 92 bis 96 in Cleve und Lämmel (2014).

a) Welche prinzipiellen Möglichkeiten der Attributauswahl gibt es?

b) Wie können Sie mit metrischen Attributen umgehen?

c) Wie lautet der Basisalgorithmus zum Erzeugen eines Entscheidungsbaums? (vgl. Ittner 2019, 6. Vorlesung S. 12)

Aufgabe 2

Vergleich mit RapidMiner

Erzeugen Sie einen neuen RapidMiner-Prozess und lesen Sie die Datei Wetter.csv ein. Erzeugen Sie mit dem Operator *ID3* einen Entscheidungsbaum der Wetterdaten aus Tabelle 5.2 in Cleve und Lämmel (2014, S. 93), vgl. Abbildung 1.

a) Welche Methoden zur Attribute-Auswahl stellt der ID3-Operator bereit? (vgl. *ID3 - RapidMiner Documentation* 2019)

Aufgabe 3

Iris-Daten mit RapidMiner

Erstellen Sie einen RapidMiner-Prozess und laden Sie die Iris-Daten aus dem Repository.

a) Diskretisieren Sie die numerischen Werte mit dem Operator *Discretize*. Verwenden Sie die Operatoren *Work on Subset* und *Map* und weisen Sie den Attributwerten sinnvolle Bezeichnungen zu.

b) Teilen Sie die vorverarbeiteten Daten mit dem Operator *Split* in eine Trainings- und eine Testmenge. Erzeugen Sie ein Modell mit den Operatoren *ID3* und *Apply Model*. Überprüfen Sie Ihr Ergebnis mit dem *Performance*-Operator.

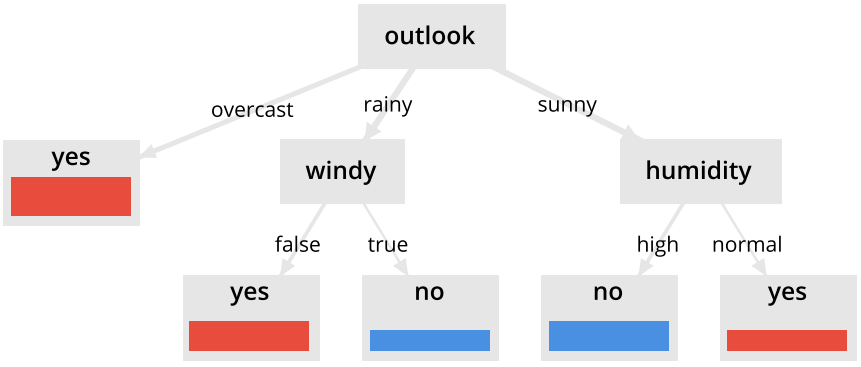

Abbildung 1: Visualisierung eines Entscheidungsbaums in RapidMiner

Lösungen

1. a) manuell, zufällig, berechnet

1. b) Gruppierung (Intervalle), Schwellwerte

1. c)

WENN alle Datensätze in der Beispielmenge E zur selben Klasse C gehören

DANN ist C das Ergebnis

SONST

- wähle ein Attribut a aus der Attributmenge A mit den Werten $\omega_1, \ldots, \omega_n$ $(\omega \in \omega_a)$
- partitioniere E in $E_1, \ldots, E_n$ $(E^\omega \subseteq E)$, abhängig von den Wertausprägungen $\omega \in \omega_a$ des Attributes a
- konstruiere Unterbäume $T_1, \ldots, T_n$ für $E_1, \ldots, E_n$
- das Ergebnis ist der Baum T mit der Wurzel a und den beschrifteten Kanten $\omega_1, \ldots, \omega_n$ zu den Unterbäumen $T_1, \ldots, T_n$

Der Algorithmus wird rekursiv wieder auf die jeweiligen Unterbäume T_i angewendet, solange bis jeder Knoten nur noch nicht weiter unterscheidbare Datensätze (einer Klasse) enthält.

3. Decision␣Tree␣Iris.rmp

Literatur

Cleve, Jürgen und Uwe Lämmel (2014). *Data mining*. Studium. München: De Gruyter Oldenbourg. 306 S. ISBN: 978-3-486-72034-1 978-3-486-71391-6.

ID3 - RapidMiner Documentation (2019). URL: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html (besucht am 01. 12. 2019).

Ittner, Prof. Dr.-Ing. Andreas (2019). „Data Mining“. Vorlesung, Wintersemester 2019/20 (Hochschule Mittweida).