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

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Entscheidungsbaumlernen II

Berechnete Attributauswahl

Die bevorzugte Art, Attribute auszuwählen, ist die berechnete Auswahl. Für die Berechnung benötigen wir ein Kriterium, nach dem die Attribute miteinander verglichen werden können, und ein Verfahren, welches das optimale Attribut auswählt. Ein einfaches Verfahren ist es, ein Attribut anhand der jeweiligen lokalen Klassifikationsleistung zu messen. Das heißt, in jedem Schritt während der Entwicklung eines Entscheidungsbaumes wird das Attribut mit der im Ergebnis geringsten Fehlerrate gewählt (vgl. Cleve und Lämmel 2014, S. 95).

Ein weiteres Verfahren nach Quinlan (1986) benutzt den Begriff der Entropie von Shannon (1948). Das Attribut mit dem größten Informationsgewinn wird für jeden Teilbaum als Wurzelknoten gewählt.

Aufgabe 1

Lesen Sie die Seiten 95 bis 106 in Cleve und Lämmel (2014) und entwickeln Sie für die folgenden Datensätze einen Entscheidungsbaum. Klassifizieren Sie die letzten drei Datensätze.

Alter-native	Fr/Sa	Hungrig	Gäste	Reser-vierung	Typ	Zeit	Warten
ja	nein	ja	einige	ja	Franz.	0–10	ja
ja	nein	ja	voll	nein	Chin.	30–60	nein
nein	nein	nein	einige	nein	Burger	0–10	ja
ja	ja	ja	voll	nein	Chin.	10–30	ja
ja	ja	nein	voll	ja	Franz.	>60	nein
nein	nein	ja	einige	ja	Ital.	0–10	ja
nein	nein	nein	keine	nein	Burger	0–10	nein
nein	nein	ja	einige	ja	Chin.	0–10	ja
nein	ja	nein	voll	nein	Burger	>60	nein
ja	ja	ja	voll	ja	Ital.	10–30	nein
nein	nein	nein	keine	nein	Chin.	0–10	nein
ja	ja	ja	voll	nein	Burger	30–60	ja
ja	nein	ja	einige	nein	Franz.	30–60	
ja	ja	ja	voll	ja	Chin.	10–30	
nein	nein	nein	keine	nein	Burger	0–10	

Lösungen

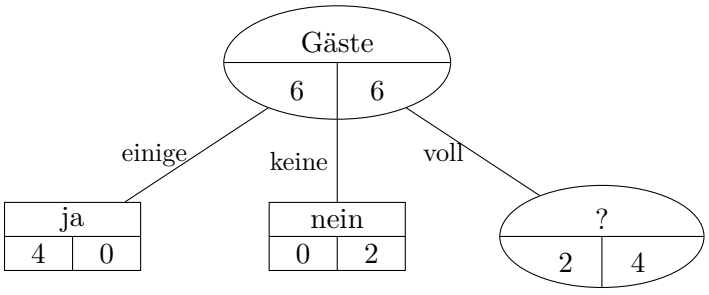
1.

Ebene 1

$$
\begin{aligned}
I(\text{Tabelle}) &= -\frac{6}{12} \cdot \log_2(\frac{6}{12}) - \frac{6}{12} \cdot \log_2(\frac{6}{12}) = 1 \\
I(\text{Alternative = ja}) &= -\frac{3}{6} \cdot \log_2(\frac{3}{6}) - \frac{3}{6} \cdot \log_2(\frac{3}{6}) = 1 \\
I(\text{Alternative = nein}) &= -\frac{3}{6} \cdot \log_2(\frac{3}{6}) - \frac{3}{6} \cdot \log_2(\frac{3}{6}) = 1 \\
I(\text{Fr/Sa = ja}) &= -\frac{2}{5} \cdot \log_2(\frac{2}{5}) - \frac{3}{5} \cdot \log_2(\frac{3}{5}) = 0,970950594454668 \\
I(\text{Fr/Sa = nein}) &= -\frac{4}{7} \cdot \log_2(\frac{4}{7}) - \frac{3}{7} \cdot \log_2(\frac{3}{7}) = 0,985228136034252 \\
I(\text{Hungrig = ja}) &= -\frac{5}{7} \cdot \log_2(\frac{5}{7}) - \frac{2}{7} \cdot \log_2(\frac{2}{7}) = 0,863120568566631 \\
I(\text{Hungrig = nein}) &= -\frac{1}{5} \cdot \log_2(\frac{1}{5}) - \frac{4}{5} \cdot \log_2(\frac{4}{5}) = 0,721928094887362 \\
I(\text{Gäste = keine}) &= 0 \\
I(\text{Gäste = einige}) &= 0 \\
I(\text{Gäste = voll}) &= -\frac{2}{6} \cdot \log_2(\frac{2}{6}) - \frac{4}{6} \cdot \log_2(\frac{4}{6}) = 0,91829583405449 \\
I(\text{Reservierung = ja}) &= -\frac{3}{5} \cdot \log_2(\frac{3}{5}) - \frac{2}{5} \cdot \log_2(\frac{2}{5}) = 0,970950594454668 \\
I(\text{Reservierung = nein}) &= -\frac{3}{7} \cdot \log_2(\frac{3}{7}) - \frac{4}{7} \cdot \log_2(\frac{4}{7}) = 0,985228136034252 \\
I(\text{Typ = Franz.}) &= -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1 \\
I(\text{Typ = Chin.}) &= -\frac{2}{4} \cdot \log_2(\frac{2}{4}) - \frac{2}{4} \cdot \log_2(\frac{2}{4}) = 1 \\
I(\text{Typ = Burger}) &= -\frac{2}{4} \cdot \log_2(\frac{2}{4}) - \frac{2}{4} \cdot \log_2(\frac{2}{4}) = 1 \\
I(\text{Typ = Ital.}) &= -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1 \\
I(\text{Zeit = 0-10}) &= -\frac{4}{6} \cdot \log_2(\frac{4}{6}) - \frac{2}{6} \cdot \log_2(\frac{2}{6}) = 0,91829583405449 \\
I(\text{Zeit = 10-30}) &= -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1 \\
I(\text{Zeit = 30-60}) &= -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1 \\
I(\text{Zeit = >60}) &= 0
\end{aligned}
$$

$$G(\text{Alternative}) = \frac{6}{12} \cdot 1 + \frac{6}{12} \cdot 1 = 1$$
$$G(\text{Fr/Sa}) = \frac{5}{12} \cdot 0,970950594454668 + \frac{7}{12} \cdot 0,985228136034252 = 0,979279160376092$$
$$G(\text{Hungrig}) = \frac{7}{12} \cdot 0,863120568566631 + \frac{5}{12} \cdot 0,721928094887362 = 0,804290371200269$$
$$G(\text{Gäste}) = 0 + 0 + \frac{6}{12} \cdot 0,91829583405449 = 0,459147917027245$$
$$G(\text{Reservierung}) = \frac{5}{12} \cdot 0,970950594454668 + \frac{7}{12} \cdot 0,985228136034252 = 0,979279160376092$$
$$G(\text{Typ}) = \frac{2}{12} \cdot 1 + \frac{4}{12} \cdot 1 + \frac{4}{12} \cdot 1 + \frac{2}{12} \cdot 1 = 1$$
$$G(\text{Zeit}) = \frac{6}{12} \cdot 0,91829583405449 + \frac{2}{12} \cdot 1 + \frac{2}{12} \cdot 1 + 0 = 0,792481250360578$$

$$gewinn(\text{Alternative}) = 1 - 1 = 0$$
$$gewinn(\text{Fr/Sa}) = 1 - 0,979279160376092 = 0,020720839623908$$
$$gewinn(\text{Hungrig}) = 1 - 0,804290371200269 = 0,195709628799731$$
$$gewinn(\text{Gäste}) = 1 - 0,459147917027245 = \underline{\underline{0,540852082972755}}$$
$$gewinn(\text{Reservierung}) = 1 - 0,979279160376092 = 0,020720839623908$$
$$gewinn(\text{Typ}) = 0$$
$$gewinn(\text{Zeit}) = 1 - 0,792481250360578 = 0,207518749639422$$



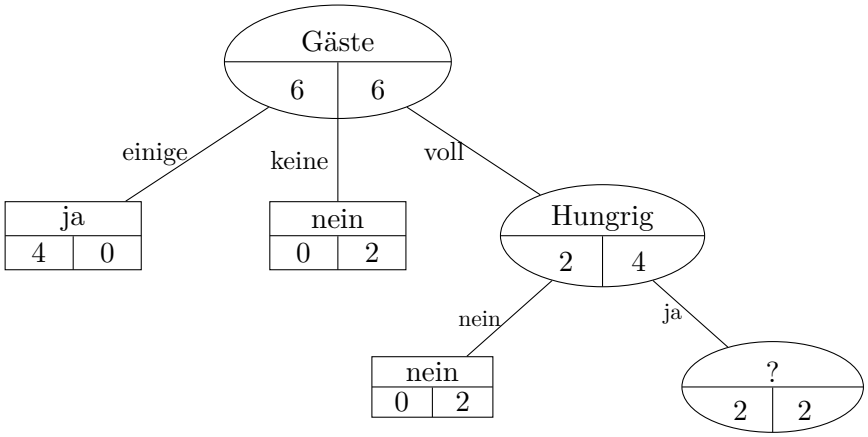
Ebebe 2

Alternative	Fr/Sa	Hungrig	Gäste	Reservierung	Typ	Zeit	Warten
ja	nein	ja	voll	nein	Chin.	30–60	nein
ja	ja	ja	voll	nein	Chin.	10–30	ja
ja	ja	nein	voll	ja	Franz.	>60	nein
nein	ja	nein	voll	nein	Burger	>60	nein
ja	ja	ja	voll	ja	Ital.	10–30	nein
ja	ja	ja	voll	nein	Burger	30–60	ja

$$I(\text{Tabelle}) = -\frac{2}{6} \cdot \log_2(\frac{2}{6}) - \frac{4}{6} \cdot \log_2(\frac{4}{6}) = 0,91829583405449$$
$$I(\text{Alternative} = \text{ja}) = -\frac{2}{5} \cdot \log_2(\frac{2}{5}) - \frac{3}{5} \cdot \log_2(\frac{3}{5}) = 0,970950594454668$$
$$I(\text{Alternative} = \text{nein}) = 0$$
$$I(\text{Fr/Sa} = \text{ja}) = -\frac{2}{5} \cdot \log_2(\frac{2}{5}) - \frac{3}{5} \cdot \log_2(\frac{3}{5}) = 0,970950594454668$$
$$I(\text{Fr/Sa} = \text{nein}) = 0$$
$$I(\text{Hungrig} = \text{ja}) = -\frac{2}{4} \cdot \log_2(\frac{2}{4}) - \frac{2}{4} \cdot \log_2(\frac{2}{4}) = 1$$
$$I(\text{Hungrig} = \text{nein}) = 0$$
$$I(\text{Reservierung} = \text{ja}) = 0$$
$$I(\text{Reservierung} = \text{nein}) = -\frac{2}{4} \cdot \log_2(\frac{2}{4}) - \frac{2}{4} \cdot \log_2(\frac{2}{4}) = 1$$
$$I(\text{Typ} = \text{Franz.}) = 0$$
$$I(\text{Typ} = \text{Chin.}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Typ} = \text{Burger}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Typ} = \text{Ital.}) = 0$$
$$I(\text{Zeit} = \text{0-10}) = 0$$
$$I(\text{Zeit} = \text{10-30}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Zeit} = \text{30-60}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Zeit} = >60) = 0$$

$$G(\text{Alternative}) = \frac{5}{6} \cdot 0,970950594454668 + 0 = 0,809125495378891$$
$$G(\text{Fr/Sa}) = \frac{5}{6} \cdot 0,970950594454668 + 0 = 0,809125495378891$$
$$G(\text{Hungrig}) = \frac{4}{6} \cdot 1 + 0 = 0,666666666666667$$
$$G(\text{Reservierung}) = 0 + \frac{4}{6} \cdot 1 = 0,666666666666667$$
$$G(\text{Typ}) = 0 + \frac{2}{6} \cdot 1 + \frac{2}{6} \cdot 1 + 0 = 0,666666666666667$$
$$G(\text{Zeit}) = 0 + \frac{2}{6} \cdot 1 + \frac{2}{6} \cdot 1 + 0 = 0,666666666666667$$

$$gewinn(\text{Alternative}) = 0,91829583405449 - 0,809125495378891 = 0,109170338675599$$
$$gewinn(\text{Fr/Sa}) = 0,91829583405449 - 0,809125495378891 = 0,109170338675599$$
$$gewinn(\text{Hungrig}) = 0,91829583405449 - 0,666666666666667 = \underline{0,251629167387823}$$
$$gewinn(\text{Reservierung}) = 0,91829583405449 - 0,666666666666667 = 0,251629167387823$$
$$gewinn(\text{Typ}) = 0,91829583405449 - 0,666666666666667 = 0,251629167387823$$
$$gewinn(\text{Zeit}) = 0,91829583405449 - 0,666666666666667 = 0,251629167387823$$



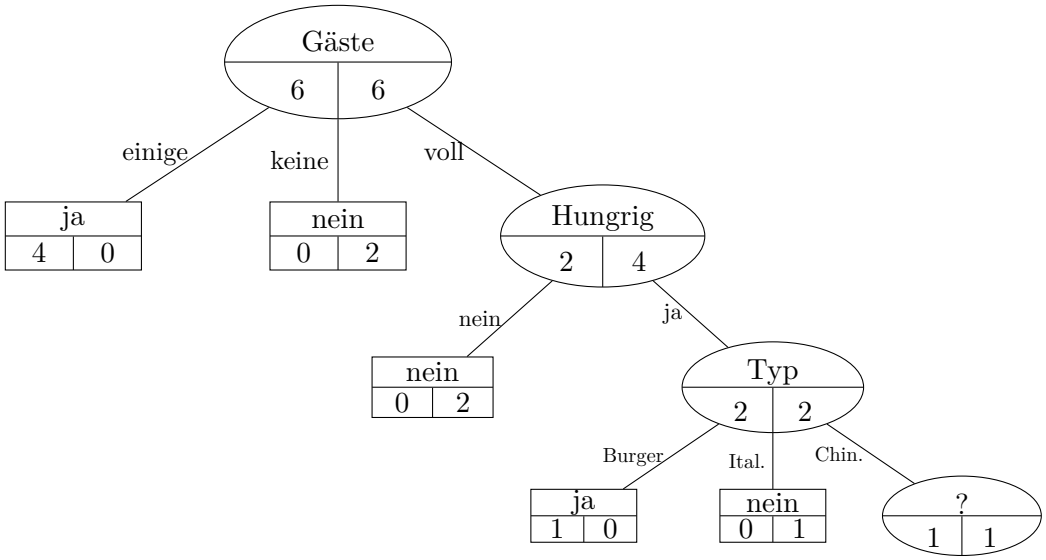
Ebene 3

Alternative	Fr/Sa	Hungrig	Gäste	Reservierung	Typ	Zeit	Warten
ja	nein	ja	voll	nein	Chin.	30–60	nein
ja	ja	ja	voll	nein	Chin.	10–30	ja
ja	ja	ja	voll	ja	Ital.	10–30	nein
ja	ja	ja	voll	nein	Burger	30–60	ja

$$
\begin{aligned}
I(\text{Tabelle}) &= -\frac{2}{4}\cdot\log_2(\frac{2}{4}) - \frac{2}{4}\cdot\log_2(\frac{2}{4}) = 1\\
I(\text{Alternative} = \text{ja}) &= -\frac{2}{4}\cdot\log_2(\frac{2}{4}) - \frac{2}{4}\cdot\log_2(\frac{2}{4}) = 1\\
I(\text{Alternative} = \text{nein}) &= 0\\
I(\text{Fr/Sa} = \text{ja}) &= -\frac{2}{3}\cdot\log_2(\frac{2}{3}) - \frac{1}{3}\cdot\log_2(\frac{1}{3}) = 0,91829583405449\\
I(\text{Fr/Sa} = \text{nein}) &= 0\\
I(\text{Reservierung} = \text{ja}) &= 0\\
I(\text{Reservierung} = \text{nein}) &= -\frac{2}{3}\cdot\log_2(\frac{2}{3}) - \frac{1}{3}\cdot\log_2(\frac{1}{3}) = 0,91829583405449\\
I(\text{Typ} = \text{Franz.}) &= 0\\
I(\text{Typ} = \text{Chin.}) &= -\frac{1}{2}\cdot\log_2(\frac{1}{2}) - \frac{1}{2}\cdot\log_2(\frac{1}{2}) = 1\\
I(\text{Typ} = \text{Burger}) &= 0\\
I(\text{Typ} = \text{Ital.}) &= 0\\
I(\text{Zeit} = \text{0-10}) &= 0\\
I(\text{Zeit} = \text{10-30}) &= -\frac{1}{2}\cdot\log_2(\frac{1}{2}) - \frac{1}{2}\cdot\log_2(\frac{1}{2}) = 1\\
I(\text{Zeit} = \text{30-60}) &= -\frac{1}{2}\cdot\log_2(\frac{1}{2}) - \frac{1}{2}\cdot\log_2(\frac{1}{2}) = 1\\
I(\text{Zeit} = >\text{60}) &= 0
\end{aligned}
$$

$$
\begin{aligned}
G(\text{Alternative}) &= \frac{4}{4}\cdot 1 + 0 = 1\\
G(\text{Fr/Sa}) &= \frac{3}{4}\cdot 0,91829583405449 + 0 = 0,688721875540867\\
G(\text{Reservierung}) &= 0 + \frac{3}{4}\cdot 0,91829583405449 = 0,688721875540867\\
G(\text{Typ}) &= 0 + \frac{2}{4}\cdot 1 + 0 + 0 = 0,5\\
G(\text{Zeit}) &= 0 + \frac{2}{4}\cdot 1 + \frac{2}{4}\cdot 1 + 0 = 1
\end{aligned}
$$

$$gewinn(\text{Alternative}) = 0$$
$$gewinn(\text{Fr/Sa}) = 1 - 0,688721875540867 = 0,311278124459133$$
$$gewinn(\text{Reservierung}) = 1 - 0,688721875540867 = 0,311278124459133$$
$$gewinn(\text{Typ}) = 1 - 0,5 = \underline{\underline{0,5}}$$
$$gewinn(\text{Zeit}) = 0$$



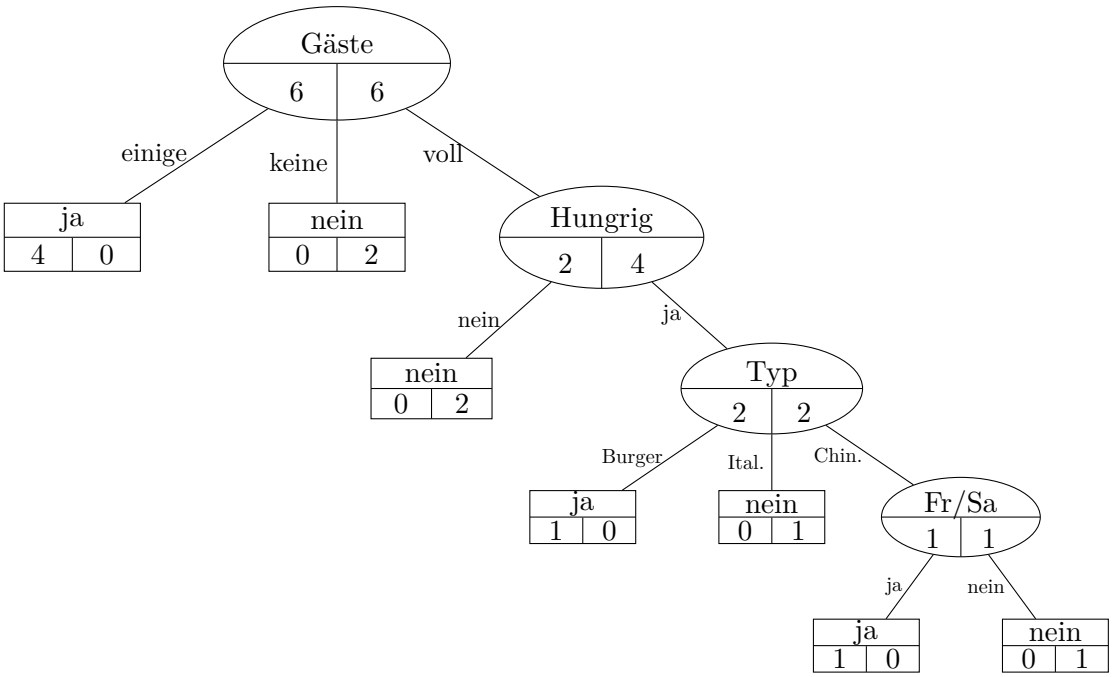
Ebene 4

Alternative	Fr/Sa	Hungrig	Gäste	Reservierung	Typ	Zeit	Warten
ja	nein	ja	voll	nein	Chin.	30–60	nein
ja	ja	ja	voll	nein	Chin.	10–30	ja

$$I(\text{Tabelle}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Alternative} = \text{ja}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Alternative} = \text{nein}) = 0$$
$$I(\text{Fr/Sa} = \text{ja}) = 0$$
$$I(\text{Fr/Sa} = \text{nein}) = 0$$
$$I(\text{Reservierung} = \text{ja}) = 0$$
$$I(\text{Reservierung} = \text{nein}) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$$
$$I(\text{Zeit} = \text{10-30}) = 0$$
$$I(\text{Zeit} = \text{30-60}) = 0$$

$$G(\text{Alternative}) = \frac{2}{2} \cdot 1 + 0 = 1$$
$$G(\text{Fr/Sa}) = 0$$
$$G(\text{Reservierung}) = 0 + \frac{2}{2} \cdot 1 = 1$$
$$G(\text{Zeit}) = 0$$

$$\mathit{gewinn}(\text{Alternative}) = 0$$
$$\mathit{gewinn}(\text{Fr/Sa}) = \underline{1}$$
$$\mathit{gewinn}(\text{Reservierung}) = 0$$
$$\mathit{gewinn}(\text{Zeit}) = 1$$



Klassifikation

Alternative	Fr/Sa	Hungrig	Gäste	Reservierung	Typ	Zeit	Warten
ja	nein	ja	einige	nein	Franz.	30–60	ja
ja	ja	ja	voll	ja	Chin.	10–30	ja
nein	nein	nein	keine	nein	Burger	0–10	nein

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.

Quinlan, J. R. (März 1986). „Induction of Decision Trees“. In: *Machine Learning* 1.1, S. 81–106. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00116251. URL: http://link.springer.com/10.1007/BF00116251 (besucht am 16. 12. 2019).

Shannon, Claude Elwood (Juli 1948). „A Mathematical Theory of Communication“. In: *Bell System Technical Journal* 27.3, S. 379–423. ISSN: 00058580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6773024 (besucht am 28. 04. 2019).