

Kriterien der Prognosegüte für kategoriale Merkmale

Sei es die Kaufentscheidung für eine Marke oder die Zuordnung zu einem Segment – häufig ist die Prognose der Ausprägung einer kategorialen Variable gefragt. Die wichtigsten Gütemaße zur Beurteilung und zum Vergleich von Prognosemodellen für kategoriale Merkmale werden anhand der Vorhersage des Ausgangs von Fußballspielen veranschaulicht.

Klassifikations- oder Konfusionsmatrix

Wie gut sind Wettquoten geeignet, um Ergebnisse von Fußballspielen vorherzusagen? Dazu wurden von der Website football-data.co.uk die Quoten für Heimsieg, Unentschieden und Auswärtssieg für die 153 Spiele der Hinrunde 2018/19 der 1. Bundesliga eines bekannten Wettanbieters erfasst. Da sich die Wahrscheinlichkeit für den jeweiligen Ausgang eines Spiels im Kehrwert der Wettquote widerspiegelt, ist der Ausgang mit der kleinsten Quote eine naheliegende Prognose. Allerdings hatte ein Unentschieden niemals die geringste Wettquote und auch nur für fünf Spiele waren die Quoten für Heim- und Auswärtssieg gleich. Weil in der Vergangenheit aber ungefähr ein Viertel der Spiele Unentschieden endete, wird das Prognosemodell um eine zweite Regel ergänzt: Ein Unentschieden wird vorhergesagt, wenn die Quote für den Heimsieg um weniger als 20% höher ist als die für den Auswärtssieg oder umgekehrt. In der Klassifikationsmatrix in Abbildung 1 sind die durch dieses Modell resultierenden Prognosen den tatsächlichen Ergebnissen gegenübergestellt.

		Tatsächlicher Ausgang			Gesamt
		Heimsieg	Unentschieden	Auswärtssieg	
Vorhergesagter Ausgang	Heimsieg	46	21	14	81
	Unentschieden	12	7	12	31
	Auswärtssieg	11	10	20	41
Gesamt		69	38	46	153

Abbildung 1: Klassifikations- bzw. Konfusionsmatrix

Korrektklassifikationsrate

Der Anteil der Beobachtungen, für die die Kategorie richtig vorhergesagt wurde, bestimmt die **Korrektklassifikationsrate** = $(n_1 + \dots + n_k + \dots + n_K) / n$ mit n_k = Anzahl der richtig prognostizierten Beobachtungen in Kategorie k und n = Gesamtanzahl der Beobachtungen. In dem Beispiel beträgt sie $(46 + 7 + 20) / 153 = 47,7\%$. Mit einer dem Prognosemodell entsprechenden Tipp-Strategie hätte man fast die Hälfte der Spiele richtig vorhergesagt. Jedoch hätte sich mit einer einfachen „naiven“ Prognose – der Zuordnung aller zu der größten Gruppe, das heißt der Prognose Heimsieg für alle Spiele – bereits eine Korrektklassifikationsrate von $69 / 153 = 45,1\%$ ergeben. Ein Vergleich mit der naiven Prognose offenbart demnach vor allem bei ungleichen Verteilungen auf die Kategorien erst die wahre Qualität eines Prognosemodells.

Konfusionsmatrix und abgeleitete Gütemaße

Ein detaillierterer Blick ist mit Hilfe einer speziellen Konfusionsmatrix möglich. Für jede Kategorie wird unterschieden, ob sie auftritt (Positive) oder nicht (Negative), sodass sich jeweils eine 2×2 -Matrix wie in Abbildung 2 bestimmen lässt. Die Konfusionsmatrix für die Kategorie Heimsieg zeigt Abbildung 3.

		Beobachtung	
		Positive	Negative
Prognose	Positive	True Positive TP	False Positive FP
	Negative	False Negative FN	True Negative TN

Abbildung 2: 2×2 -Konfusionsmatrix

		Beobachtung	
		Heimsieg	Kein Heimsieg
Prognose	Heimsieg	46	35
	Kein Heimsieg	23	49

Abbildung 3: 2×2 -Konfusionsmatrix für die Kategorie Heimsieg

Die Formel der Korrektklassifikationsrate angewendet auf die 2×2 -Konfusionsmatrix ergibt die

$$\text{Accuracy} = (TP + TN) / n$$

Für Heimsiege ist sie gleich $(46 + 49) / 153 = 62,1\%$.

Stellt man analog eine Konfusionsmatrix für Unentschieden auf, so ergibt sich eine Accuracy von $(7 + 91) / 153 = 64,1\%$. Für Auswärtssiege beträgt die Accuracy $(20 + 86) / 153 = 69,3\%$.

Der Anteil der korrekt als „Positive“ prognostizierten Beobachtungen an allen tatsächlich „Positives“ ist der **Recall (oder Trefferquote)** = $TP / (TP + FN)$

Für Heimsiege ist Recall gleich $46 / (46 + 23) = 66,7\%$, für Unentschieden gleich $7 / (7 + 31) = 18,4\%$ und für Auswärtssiege gleich $20 / (20 + 26) = 43,5\%$. Accuracy deutet nicht auf große Unterschiede zwischen den Kategorien hin und zeigt somit ein verzerrtes Bild. Gemäß Recall werden Unentschieden jedoch erheblich schlechter vorhergesagt als Heim- oder Auswärtssiege.

Der Anteil der korrekt als „Positive“ prognostizierten an allen als „Positive“ vorhergesagten Beobachtungen ist die **Precision** = $TP / (TP + FP)$

Für Heimsiege ist Precision gleich $46 / (46 + 35) = 56,8\%$, für Unentschieden gleich $7 / (7 + 24) = 22,6\%$ und für Auswärtssiege gleich $20 / (20 + 21) = 48,8\%$. Das heißt die Prognose Heimsieg war in 56,8% der Spiele richtig, die Prognose Unentschieden jedoch nur in 22,6%.

Je nach Zielsetzung des Prognosemodells sind die Maße unterschiedlich relevant. Sollen beispielsweise auf einem Markt die Käufer verschiedener Marken prognostiziert werden, ist die Korrektklassifikationsrate zu betrachten. Ist nur von Interesse, die Käufer der eigenen Marke zu identifizieren, wäre der Recall heranzuziehen.

Ist nur von Interesse, die Käufer der eigenen Marke zu identifizieren, wäre der Recall heranzuziehen.

In Ausgabe 2/2019: Bestimmtheitsmaß der linearen Regression



Johannes Lükken, Diplom-Psychologe, ist Leiter des Bereichs Data Sciences bei IfaD.

jlueken@ifad.de



Prof. Dr. Heiko Schimmelpfennig, Diplom-Kaufmann, ist Projektleiter für Data Sciences bei IfaD.

hshimmelpfennig@ifad.de



Literatur

Runkler, T. A.: *Klassifikation*. In: Data Mining, Wiesbaden, 2015, S. 89-107.