

Kapitel 4

Entropiecodierung

Dieses Kapitel befasst sich mit verschiedenen Methoden zur Kompression von Daten unter Ausnutzung der statistischen Verteilung von Symbolen. Die Symbole werden hierbei als unabhängig voneinander betrachtet. Zunächst wird auf einige Aspekte der Codierungstheorie eingegangen und anschließend werden die Verfahren anhand von Beispielen erläutert. Ausführungen zur adaptiven Codierung und zu Problemen bei sehr großen Symbolalphabeten bilden den Abschluss

4.1 Codierungstheorie

Die Entropiecodierung, auch statistische Codierung genannt, umfasst Methoden der Datenkompression, welche in der Lage sind, die Codierungsredundanz ΔR_{cod} durch Ausnutzen der Symbolverteilung zu reduzieren. Ziel ist die Zuordnung von Codewörtern (Bitfolgen) derart, dass die mittlere Bitrate minimiert und an die Signalentropie H_{src} angenähert wird. Es wird versucht, jedem Symbol nur so viele Bits zuzuordnen, wie es aufgrund des Informationsgehalts des Symbols erforderlich ist. Symbolen mit hoher Auftretenswahrscheinlichkeit werden kurze Codewörter zugewiesen, während seltene Symbole längere Codewörter erhalten.

Die Theorie der Codierung, wie wir sie heute verwenden, geht auf Claude E. Shannon zurück [Sha48, Sha76]. l_i sei die Länge jenes Codewortes c_i , das dem Symbol s_i mit der Auftretenswahrscheinlichkeit p_i zugeordnet wird. Die mittlere Codelänge einer Symbolfolge kann dann mit

$$\bar{l}_i = \sum_{i=1}^K p_i \cdot l_i \quad [\text{Bits/Symbol}] \quad (4.1)$$

angegeben werden, wenn die Signalquelle K verschiedene Zeichen produziert. Die niedrigste Bitrate wird erreicht, wenn ein Code (eine geeignete Auswahl von Codewörtern) den kleinsten Wert für \bar{l}_i liefert. Die entscheidende Frage ist nun, ob es eine untere Grenze für die mittlere Codelänge gibt und wenn ja, wie groß sie ist. Shannon hat 1948 bewiesen, dass \bar{l}_i stets größer oder mindestens gleich der Quellenentropie H_{src} ist. Darüber hinaus hat er gezeigt, dass immer ein Code gefunden werden kann, der eine Übertragung mit weniger als $H_{\text{src}} + 1$ bit pro Abtastwert ermöglicht

$$H_{\text{src}} \leq \bar{l}_i < H_{\text{src}} + 1. \quad (4.2)$$

In Tabelle 4.1 ist ein Beispiel für ein Symbolalphabet mit $K = 4$ Zeichen angegeben. Auf Basis der Wahrscheinlichkeiten ergibt sich für jedes Symbol nach Gleichung (2.1)

i	1	2	3	4
s_i	a	b	c	d
p_i	0.4	0.2	0.1	0.3
$I_i[\text{bit}]$	1.32	2.32	3.32	1.74
c_i	00	01	10	11
l_i	2	2	2	2

Tabelle 4.1: Beispielalphabet mit 4 Symbolen

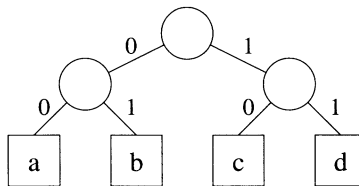


Abbildung 4.1: Codewortbaum mit Codewörtern gleicher Länge

ein bestimmter Informationsgehalt. Die Entropie beträgt somit laut Gleichung (2.2) $H_{\text{src}} \approx 0.4 \cdot 1.32 + 0.2 \cdot 2.32 + 0.1 \cdot 3.32 + 0.3 \cdot 1.74 = 1.846$ bit/Symbol. Den Symbolen wurden Codewörter mit einer festen Länge von $l_i = l = \lceil \log_2(4) \rceil = 2$ Bits zugeordnet. Die durchschnittliche Codelänge beträgt entsprechend Gl. (4.1)

$$\bar{l}_i = 0.4 \cdot 2 + 0.2 \cdot 2 + 0.1 \cdot 2 + 0.3 \cdot 2 = 2 \text{ Bits/Symbol}.$$

Es ist zu erkennen, dass dieser Code die durch die Entropie vorgegebene untere Grenze nicht unterschreitet ($\bar{l}_i > H_{\text{src}}$). Der Code ist aber auch schon so gut, dass er innerhalb der in Gleichung (4.2) angegebenen Grenzen liegt.

Eine übliche Darstellungsform für Codes sind sogenannte Codebäume. Abbildung 4.1 zeigt den Codebaum für das Beispiel aus Tabelle 4.1. Die Symbole bilden die Blätter des Baumes und die Beschriftung der Zweige von der Wurzel bis zum Blatt entspricht dem jeweiligen Codewort.

Vergleicht man nun den Informationsgehalt eines jeden Zeichens mit der Anzahl der zugewiesenen Codebits, so ist leicht zu erkennen, dass sie in keiner Weise miteinander korrespondieren, da die Codewörter eine feste (fixierte) Codelänge haben (FLC ... *Fixed Length Code*). Daraus resultiert die relativ hohe Redundanz von $\Delta R_{\text{cod}} = \bar{l}_i - H_{\text{src}} = 0.154$ bit/Symbol. In den folgenden Abschnitten werden verschiedene Verfahren vorgestellt und diskutiert, bei denen mit Hilfe von variablen Codelängen (VLC ... *Variable Length Code*) versucht wird, eine bessere Übereinstimmung von Informationsgehalt und Codewort und damit eine Verminderung der Codierungsredundanz zu erreichen.

4.2 Morse-Code

Einer der ältesten Codes, der noch heute (wenn auch nur sehr selten) Anwendung findet, ist der Morse-Code (nach Samuel Morse 1843). Basis für diesen Code sind kurze