Bewertung von partiellen Wahrscheinlichkeitsinformationen bei entropieoptimaler Wissensverarbeitung

W. Rödder, E. Reucher

FernUniversität in Hagen

Zusammenfassung: Informationsbeschaffung zur Reduktion von Ungewißheit in einer Wissensdomäne ist das zentrale Thema dieses Aufsatzes. Hierzu werden alternative Informationspakete hinsichtlich ihres Gehalts für den Empfänger bewertet. Die Bewertung hängt von seinem Vorwissen und von seinem Ziel ab. Ziel kann die Informationsverdichtung auf der gesamten Wissensdomäne, Teilen davon oder bzgl. eines reellen Nutzens sein. Alle Bewertungsschritte vollziehen sich in einem Modell, das durch Verwendung des Entropiebegriffs auf der Basis Shannonscher Informationstheorie steht. Ist das optimale Informationspaket beschafft, können auch Entscheidungen unter Nutzeninformationsgesichtspunkten getroffen werden.

1. Einleitung

Eine wachsende Zahl von Wissenschaftlern und Praktikern verwendet probabilistische Inferenzmodelle in der Künstlichen Intelligenz. Das Wissen über eine Domäne wird in Form von Wahrscheinlichkeitsverteilungen erfaßt, und bei Evidentwerden gewisser Sachverhalte laufen Inferenzprozesse in diesen Verteilungen ab. Zum Aufbau und zur Manipulation der Verteilungen dienen Bayes- oder Markoff-Netze, vgl. Lauritzen [LAU] oder Whittaker [WHI]. Sind die probabilistischen Modelle nicht vollständig bekannt, versagen die klassischen Netze; als weiteres strukturierendes Element tritt dann neben die partielle Information die Entropie, vgl. [RKI], [RME], [RXU]. Diese Art der Wissensverarbeitung ist Grundlage der vorliegenden Schrift.

Unternehmen benötigen Information über Märkte, Käufer und Konkurrenten, um erfolgreich agieren zu können. Solche Information wird von Markt- oder Meinungsforschungsinstituten erhoben und dem Nachfrager angeboten. Den Wert von Information für das eigene Unternehmen richtig einzuschätzen, ist eine wichtige Management-aufgabe. Wert ist hier Reduktion von Ungewißheit über zukünftige Entwicklungen. Die Reduktion ist merkmal-bezogen, Informationsbeschaffung ist zielabhängig.

In Kapitel 2 liefern wir den modelltheoretischen Rahmen und formulieren einige grundlegende mathematische Aussagen. Kapitel 3 zeigt auf, wie mittels Informationsbewertung Wissensverdichtung erreicht wird. Ein kurzer Ausblick schließt die Arbeit ab.

2. Informationstheoretische Grundlagen

2.1 Information, Entropie und Unsicherheit

Shannon [SHA] untersuchte die Informationsrate, die von einer ergodischen Quelle über einen Kanal mittels einer Signalfolge gesendet werden kann. Er zeigte, daß diese Rate ($0\ ld\ 0 = 0$)

$$H(P) = -\sum_{v} p(v)\, ld\ p(v) \tag{1}$$

beträgt und nannte sie die Entropie der Quelle. Hier variieren die Signale *v* in einem endlichen Alphabet *V*, $v \in V$. *p* (*v*) ist die Signalwahrscheinlichkeit und $-ld\ p(v)$ die beim Empfang von *v* und bei optimaler Codierung gewonnene Information, *ld* ist der duale Logarithmus und damit hat *H* die Dimension [bit]. Bekanntlich wird *H* für die Gleichverteilung P^0 maximal und 0, falls das Eintreten eines *v* sicher ist.

Jaglom und Jaglom [JAJ] erkannten, daß das informationstheoretische Konzept der ergodischen Quelle auf eine Population mit Elementarereignissen $v \in V$ und Eintrittswahrscheinlichkeiten *p* (*v*) übertragbar ist. Auch hier ist $-ld\ p(v)$ Unsicherheitsreduktion bei Eintreten von *v* und damit Information, *H* ihr Erwartungswert. Ist die Verteilung der Eintrittwahrscheinlichkeiten der $v \in V$ von P zu Q übergegangen, so mißt $-ld\ p(v) - (-ld\ q(v))$ die Informationsdifferenz bei Eintritt von *v* und bei Kenntnis dieses Übergangs (gegenüber der Nichtkenntnis).

$$R(Q,P) = \sum_{v} q(v)\, ld\, {}^{q(v)}\!/_{p(v)} \tag{2}$$

ist mithin der informationstheoretische Vorteil, den diese Kenntnis beinhaltet, man vergleiche auch die Ausführungen von Shannon [SHW]. $R(Q,P)$ heißt Relative Entropie oder gerichtete Divergenz von Q bzgl. P [KUL].

Alle Überlegungen dieses Abschnitts erhalten eine neue Qualität, wenn man auf folgende modelltheoretische Interpretation erweitert. Die Elementarereignisse *v* seien nun durch Ausprägungen v_e endlichwertiger Variabler V_e beschreibbar: $v = v_1,...,v_n$. *H* und *R* erlauben nach Faktorisierung $(0 \mid 0 = 0)$

$p(v_1,...,v_n) = p(v_1) \cdot ... \cdot p(v_n \mid v_1,...,v_{n-1})$ eine wichtige Umformung. Nach einfacher Umordnung und nach Aufsummieren ergibt sich für *H* (*P*) bzw. *R* (*Q*, *P*)

$$H(P) = -\sum_{v_1} p(v_1)\, ld\ p(v_1) - ... - \sum_{v_1,...,v_{n-1}} p(v_1,...,v_{n-1}) \sum_{v_n} p(v_n \mid v_1,...,v_{n-1})\, ld\ p(v_n \mid v_1,...,v_{n-1}) \qquad \text{und} \tag{3}$$

$$R(Q,P) = \sum_{v_1} q(v_1)\, ld\, \frac{q(v_1)}{p(v_1)} + ... + \sum_{v_1,...,v_{n-1}} q(v_1,...,v_{n-1}) \sum_{v_n} q(v_n \mid v_1 ... v_{n-1})\, ld\, \frac{q(v_n \mid v_1,...,v_{n-1})}{p(v_n \mid v_1,...,v_{n-1})}. \tag{4}$$

H mißt die wechselseitige Unbestimmtheit der Zufallsvariablen V_e in P. Je weniger Information die Realisierungen von Variablen über die anderer Variabler enthalten, umso größer ist *H*. Die Information „in" P ist klein, die mögliche Unsicherheitsreduktion noch groß. *R* mißt den bereits erfolgten Informationsgewinn über bedingte Ereignisse bei Kenntnis des Übergangs von P nach Q. *H* und *R* machen Aussagen über Abhängigkeitsstrukturen in P bzw. zwischen Q und P. Über die hier intuitiv dargestellten Zusammenhänge hinausgehende axiomatische Zugänge zu Entropie und Relativer Entropie findet der Leser in [SHO], [SJO], [KIS] und [PAV]. Insbesondere Kern-Isberners Arbeit [KIS] fokussiert die Abhängigkeitsstruktur zwischen Variablen.

2.2 Konditionale und Information

Zum Aufbau von Wahrscheinlichkeitsverteilungen führen wir nun Konditionale ein. Wie schon in Abschnitt 2.1 sei $V = \{V_1,...,V_n\}$ eine endliche Menge endlichwertiger Variabler mit dem Ereignisfeld über allen Elementar-

ereignissen $v = v_1,...,v_n$. Jedes Ereignis kann in natürlicher Weise mit einer Aussage identifiziert werden, die aus Literalen $V_e = v_e$ und den Konnektiven Negation, Konjunktion sowie Disjunktion gebildet wird. Solche Ereignisse notieren wir mittels großer Buchstaben *A*, *B*. Für Ereignisse *B* und *A* heißt $B|A$ ein Konditional und für positives $P(A)$ ist die Wahrscheinlichkeit eines Konditionals seine bedingte Wahrscheinlichkeit $P(B|A) = P(AB) \,/\, P(A)$.

Hier dienen die Konditionale als Sprache zum Aufbau einer Wahrscheinlichkeitsverteilung über dem Ereignisfeld der Wissensdomäne. Für die Algebra solcher Konditionale in ihrer inhärenten dreiwertigen Logik vgl. Calabrese [CAL].

R_0 sei nun eine Menge solcher Konditionale $B_i|A_i$ mit vom Experten geschätzten Wahrscheinlichkeiten x_i ihres Erfülltseins in der zugrundeliegenden Population: $R_0 = \{B_1|A_1[x_1],..., B_I|A_I[x_I]\}$. I. a. gibt es überabzählbar viele Verteilungen, die allen Konditionalen gehorchen; die folgende Aufgabe wählt davon eine, informationstheoretisch ausgezeichnete, aus:

$$P^* = arg\ Max\ H(Q) \quad \text{s. d.} \quad Q \text{ erfüllt } R_0. \tag{5}$$

Man rechnet sofort nach, daß der Imperativ *Max H* (*Q*) durch $Min\ R(Q, P^0)$ ersetzt werden kann; Maximierung der Entropie und Minimierung der Relativen Entropie zur Gleichverteilung P^0 führen zum gleichen Ergebnis. Diese formal triviale Tatsache ist informationstheoretisch tiefsinnig. P^* erlaubt in R_0 noch maximale Unsicherheitsreduktion und enthält gleichzeitig minimalen Informationsgewinn gegen P^0.

Im nächsten Abschnitt beschäftigt uns die Frage, welches der alternativen Informationsangebote $R_{11},...,R_{1K}$ das in R_0 gelieferte Wissen am besten ergänzt. Jedes dieser Angebote besteht wiederum aus einer endlichen Menge von Konditionalen mit deren Wahrscheinlichkeiten. Alle Informationspakete seien miteinander und mit R_0 verträglich, also durch Wahrscheinlichkeitsverteilungen erfüllbar.

2.3 Nutzeninformation

In Verallgemeinerung der bisherigen Überlegungen lassen wir nun zu, daß die Elementarereignisse *v* des Merkmalsraums reelle Nutzen *u*(*v*) tragen. Gesucht ist das Informationspaket aus $R_{11},...,R_{1K}$, das die Nutzenungewißheit des Empfängers am stärksten reduziert. Um diese sprachliche Formel mit mathematischem Inhalt zu füllen, bedarf es eines gegenüber (2) verallgemeinerten Divergenzbegriffs.

Ist *u* eine beliebige Nutzenbelegung auf *V* und sind P, Q beliebige Verteilungen, so ist

$$R(u, Q, P) = \sum_v u(v)q(v)\, ld\, \frac{u(v)q(v)}{u(v)p(v)} + u(v)p(v) - u(v)q(v) \tag{6}$$

die *u*-Nutzendivergenz von P und Q. Sie hat folgende Eigenschaften:

- für alle $u > 0$ und für alle $Q, P: R(u, Q, P) \geqq 0$
- für alle $u > 0$ und für alle $Q, P: R(u, Q, P) = 0$ g.d.w. $Q = P$
- Ist $u(v) = 1$ für alle *v*, so reduziert sich $R(u, Q, P)$ zu $R(Q, P)$

- $R(u,Q,P)$ ist linear in *u*.

$R(u,Q,P)$ mißt die Divergenz zwischen allen Komponenten des Vektors $u(v)\cdot q(v)$, alle *v*, und des Vektors $u(v)\cdot p(v)$, alle *v*. Mit $R(u,Q,P)$ läßt sich ein wichtiger Satz formulieren.

Satz (Vorausschauende Nutzeninformationsgewinnung VN)

Es sei $\overline{P}$ eine beliebige a priori Verteilung auf *V*.

i) Ist $\overline{P}_u^* = arg\ Min\ R(u,Q,\overline{P})$ s.d. Q erfüllt R_0 und
ist $\overline{P}_{uk}^* = arg\ Min\ R(u,Q,\overline{P})$ s.d. Q erfüllt $R_0 \cup R_{1k}$, so gilt
$\overline{P}_{uk}^* = arg\ Min\ R(u,Q,\overline{P}_u^*)$ s.d. Q erfüllt $R_0 \cup R_{1k}$.

ii) Für jedes Q, das $R_0 \cup R_{1k}$ erfüllt, gilt $R(u,Q,\overline{P}_{uk}^*)+R(u,\overline{P}_{uk}^*,\overline{P})=R(u,Q,\overline{P})$.

iii) Für $P \neq \overline{P}_{uk}^*$ gibt es stets ein Q, das $R_0 \cup R_{1k}$ erfüllt, mit $R(u,Q,P)+R(u,P,\overline{P})>R(u,Q,\overline{P})$.

Die Beweise zum Satz lehnen sich an die Ausführungen von Csiszár [CSI] und Meyer [MEY] für den Spezialfall an, daß für alle *u* (*v*) = 1 sind. Sie sprengen den Rahmen dieser Schrift.

Laut i) ist stufenweise Informationsgewinnung $\overline{P} \to \overline{P}_u^* \to \overline{P}_{uk}^*$ der direkten $\overline{P} \to P_{uk}^*$ gleichwertig. ii) und iii) erweitern diese Aussage dahingehend, daß der Übergang $\overline{P} \to P_{uk}^* \to Q$ für jedes in $R_0 \cup R_{1k}$ zulässige Q gleichmäßig besser ist als $\overline{P} \to P \to Q$. $\overline{P}_{uk}^*$ ist eine vorsichtige Vorwegnahme zukünftiger Nutzeninformationsentwicklungen und nach iii) die einzige Verteilung mit dieser Eigenschaft.

$R(u,Q,\overline{P}_{uk}^*)=R(u,Q,\overline{P})-R(u,\overline{P}_{uk}^*,\overline{P})$ ist ein Maß dafür, um wieviel die Nutzendivergenz $R(u,Q,\overline{P})$ das Minimum übersteigt. Das Maximum $Max\ R(u,Q,\overline{P}_{uk}^*)$ für Q, die $R_0 \cup R_{1k}$ erfüllen, mißt mithin Nutzenungewißheit in $R_0 \cup R_{1k}$. Mit dem obigen Satz ist es gelungen, ausgehend von einer a priori Verteilung $\overline{P}$ ein solches $\overline{P}_{uk}^*$ zu berechnen, das die hervorragende Eigenschaft der gleichmäßig geringsten Nutzenungewißheit besitzt. Der Leser ist aufgefordert, alle Aussagen für den Fall zu wiederholen, daß für alle *v* *u*(*v*) = 1 gilt. Sind die Nutzen aller Elementarereignisse gleich 1, reduziert sich $R(u,Q,\overline{P}_{uk}^*)$ zur relativen Entropie $R(Q,\overline{P}_k^*)$ und Nutzenungewißheit zur informationstheoretischen Ungewißheit. Es gilt $\sqrt{2R(Q,\overline{P}_k^*)} \geq \sum_v |q(v)-\overline{p}_k^*(v)|$. Eine Abschätzung für $\sum_v u(v)|q(v)-\overline{p}_{uk}^*(v)|$ und damit für die betragliche Nutzendifferenz durch $R(u,Q,\overline{P}_{uk}^*)$ ist noch unbewiesen.

Für $\overline{P}=P^0$, d. h. für die Gleichverteilung, liegt ein wichtiger Spezialfall des obigen Satzes vor, den wir in folgendem Korollar behandeln. Hierzu sei zunächst die absolute Nutzenentropie $H(u,Q)=-R(u,Q,P^0)+ld\ \mathrm{M}$, mit M = Mächtigkeit des Elementarereignisraums aller *v* definiert.

Korollar (VN bei a priori-Gleichverteilung)

Unter den Voraussetzungen wie im Satz und für $\overline{P} = P^0$ gelten alle Aussagen analog für ein P_u^* statt $\overline{P}_u^*$. Die Differenz $R(u,Q,P_{uk}^*) = R(u,Q,P^0) - R(u,P_{uk}^*,P^0)$ wird dann zu $H(u,P_{uk}^*) - H(u,Q)$.

3. Informationsbewertung

Mit den Vorbereitungen im vorigen Kapitel ist die Wahl eines Informationspaketes R_{1k} aus K möglich und, bei Berücksichtigung vorhandenen Wissens $\overline{P}$, mittels folgendem Algorithmus durchführbar.

Algorithmus

1. Schritt: Zu jedem R_{1k} errechne $\overline{P}_{uk}^*$ gemäß Satz

2. Schritt: $\underset{Q}{Max}\, R(u,Q,\overline{P}_{uk}^*)$ s.d. Q erfüllt $R_0 \cup R_{1k}$ mißt die maximale Nutzenungewißheit in R_{1k}.

3. Schritt: $\underset{k}{Min}\, \underset{Q}{Max}\, R(u,Q,\overline{P}_{uk}^*)$ s.d. Q erfüllt $R_0 \cup R_{1k}$ liefert minimale Nutzenungewißheit über *k*.

Ein einfaches Beispiel bei leerer Regelmenge R_0 soll die Zusammenhänge verdeutlichen.
Eine Forschungsabteilung steht vor der Durchführung zweier sich ergänzender Forschungsprojekte *A* und *B*, deren Ergebnisse gut (*g*) oder schlecht (*s*) sein können. Sie wertet die Ergebnisse mit den Nutzen und vergibt aufgrund ihrer Ungewißheitseinschätzung die a priori-Wahrscheinlichkeiten der Ergebnisse wie in Tabelle 1:

A	*B*	Nutzen *U*	$\overline{P}$
g	*g*	100.000	0.01
g	*s*	10.000	0.09
s	*g*	100	0.50
s	*s*	1	0.40

Tabelle 1

$\overline{P}_{u1}^*$	Q_1	$\overline{P}_{u2}^*$	Q_2
0.01	0.60	0.01	0.70
0.09	0.40	0.09	0.00
0.59	0.00	0.30	0.30
0.31	0.00	0.60	0.00

Tabelle 2

Nun werden die beiden Informationspakete angeboten: $R_{11} = \{B = g\ [0.6]\}$, $R_{12} = \{A = \text{s} \wedge B = \text{g}\ [0.3]\}$.
Gemäß Schritt 1 und 2 des Algorithmus erhält man für das jeweilige Paket die Verteilungen aus Tabelle 2. Nach Schritt 3 gilt $Min\{R(u,Q_1,\overline{P}_{u1}^*), R(u,Q_2,\overline{P}_{u2}^*)\} = R(u,Q_2,\overline{P}_{u1}^*) = Min\{109.455, 214.573\} = 109.455$, womit das Informationspaket R_{11} minimale Nutzenungewißheit liefert.

Im vorliegenden Aufsatz wird gezeigt, wie man bei gegebener Nutzenbewertung auf einem Merkmalsraum und bei einer –vorsichtigen– a priori Einschätzung über die Wahrscheinlichkeiten der zukünftigen Realisierung solcher Nutzen, diese Vorsicht auch bei Informationszugang wahren kann. Nutzenorientierte Ungewißheitsreduktion ist das Ziel der Informationsgewinnung. Die Überlegungen in dieser Arbeit sind leicht auf den Fall übertragbar, in dem ein Rand $W \subset V$ Ziel solcher Informationsgewinnung ist. In Entscheidungssituationen wird sich der Entscheidende in einer Ungewißheitssituation nicht mehr am ungünstigsten Erwartungsnutzen (*Max-Min-Prinzip*), sondern an einem Erwartungsnutzen gleichmäßig geringster Nutzenungewißheit orientieren. Die Arbeiten hierzu sind in Vorbereitung.

Literatur

[CAL] P. G. Calabrese: *Deduction and Inference Using Conditional Logic and Probability*, in: Conditional Logic in Expert Systems, I. R. Goodman, M. M. Gupta, H. T. Nguyen and G. S. Rogers (editors). Elsevier Science Publishers B. V., 71 – 100, (1991).

[CSI] I. Csiszár: *I-Divergence Geometry of Probability Distributions and Minimisation Problems.* The Annals of Probability 3, (1), 146 – 158, (1975).

[JAJ] A. M. Jaglom and I. M. Jaglom: *Wahrscheinlichkeit und Information*, Harry Deutsch, (1984).

[KIS] G. Kern-Isberner: *Characterising the principle of minimum cross-entropy within a conditional-logical framework*. Artificial Intelligence, Vol. 98, 169 – 208, (1998).

[KUL] S. Kullback: *Information Theory and Statistics*, John Wiley & Suns, New York, (1959).

[LSP] S. L. Lauritzen and D. J. Spiegelhalter: *Local computations with probabilities in graphical structures and their applications to expert systems*, Journal of the Royal Statistical Society 13 (2), 415 – 448, (1988).

[MEY] C.-H. Meyer: *Korrektes Schließen bei unvollständiger Information*. Dissertation an der FernUniversität in Hagen, Peter Lang, Frankfurt (1998).

[PAV] J. B. Paris and A. Vencovská: *A note on the inevitability of maximum entropy*, in: International Journal of Approximate Reasoning, 14, 183 – 223, (1990).

[RKI] W. Rödder and G. Kern-Isberner: *Representation and extraction of information by probabilistic logic*. Information Systems 21 (8), 637 – 652, (1996).

[RME] W. Rödder and C.-H. Meyer: *Coherent knowledge processing at maximum entropy by SPIRIT*, in: Proceedings 12th Conference on Uncertainty in Artificial Intelligence, E. Horitz and F. Jensen (editors), Morgan Kaufmann, San Francisco, California, 470 – 476, (1996).

[RXU] W. Rödder, L. Xu: *Entropy-driven Inference and Inconsistency*, Proc. Artificial Intelligence and Statistics, Fort Lauderdale, Florida, 272 – 277, (1999).

[SHA] C. E. Shannon: *A mathematical theory of communication*, Bell System Tech. J. 27, 379 – 423 (I), 623 – 656 (II), (1948).

[SHO] J. E. Shore: *Relative Entropy, Probabilistic Inference, and AI*, in: Uncertainty in Artificial Intelligence. North Holland, Amsterdam, 211 - 215, (1986).

[SHW] C. E. Shannon and W. Weaver: *Mathematische Grundlagen der Informationstheorie*, Oldenbourg, München, (1976).

[SJO] J. E. Shore and R. W. Johnson: *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross Entropy*. IEEE Trans. Information Theory 26 (1), 26 – 37, (1980).

[WHI] J. Whittaker: *Graphical Models in Applied Mathematical Multivariate Statistics*, John Wiley & Sons, (1990).