

Unterstützung der Bildung stochastischer Modelle: Von qualitativen verbalen Relationsbeschreibungen zu quantitativen Beziehungen

Claus Möbus (Claus.Moebus@informatik.uni-oldenburg.de)

Olaf Schröder (Olaf.Schroeder@offis.uni-oldenburg.de)

Carl-von-Ossietzky-Universität Oldenburg, FB Informatik, Postfach 2503, D-26111 Oldenburg, und
Institut OFFIS, Escherweg 2, D-26121 Oldenburg

Fragestellung:

In vielen Gegenstandsbereichen wie etwa der Betriebswirtschaft oder der Medizin müssen Entscheidungen unter Unsicherheit getroffen werden, da das Wissen unvollständig ist und / oder die Beziehungen zwischen den relevanten Variablen stochastisch oder nicht genau bekannt sind. Auch ist relevantes Datenmaterial oft nicht verfügbar, die Information über stochastische Fakten und Relationen liegt häufig nur in verbaler Form vor, wie z.B.: "Es ist sehr gut möglich, daß eine Verlagerung intrakranieller Blutgefäße anhaltenden Kopfschmerz nach sich zieht." Auch Domänenexperten zögern oft, genauere (z.B. numerische) Angaben zu machen, etwa in der Medizin (NAKAO & AXELROD, 1983).

Andererseits sollten "Domänenmodelle", also Modelle über das Wirkungsgefüge der interessierenden Variablen wie möglicher Ursachen, ihrer Auswirkungen und deren Indikatoren, so präzise wie möglich formuliert werden, wenn sie für Entscheidungsprozesse wie etwa in der Medizin sinnvoll herangezogen werden können. In der Medizin ist dies eine Voraussetzung rationeller Diagnose, Beratung und Therapieplanung sowie der Dokumentation der getroffenen Entscheidungssequenzen. Auch zur Vorbereitung betriebswirtschaftlicher Entscheidungen können solche Modelle oft sinnvoll eingesetzt werden.

Explizite Modellvorstellungen sind wichtig im Rahmen von Qualitätssicherungsmaßnahmen, da sie getroffene Entscheidungssequenzen nachvollziehbar machen. In der Praxis werden vorhandene Modellvorstellungen allerdings häufig nicht explizit formuliert.

MEDIKUS (Modellierung, Erklärung und Diagnoseunterstützung bei komplexen, unsicheren Sachverhalten; SCHRÖDER, MÖBUS, THOLE, 1996) ist ein System, das die Modellbildung ebenso wie Modellanwendung für Diagnose und Beratung in ausgewählten medizinischen Bereichen unterstützt. Domänenmodelle werden als Bayes-Netze (PEARL, 1988) repräsentiert. Es soll hier die Frage untersucht werden, wie Personen bei der stochastischen Modellbildung unterstützt werden können. Dazu wird ein Ansatz vorgestellt, der die Gewinnung des für die Domänenmodelle erforderlichen quantitativen stochastischen Wissens anhand qualitativer verbaler Konzept- und Relationsbeschreibungen ermöglicht. Mit diesem Ansatz können qualitative Aussagen von Experten in nichtreaktiver Weise für das Domänenmodell quantifiziert werden. Nach Validierung der so gewonnenen numerischen Relationen kann das entstandene Modell für diagnostische oder Beratungszwecke bis auf weiteres eingesetzt werden.

Existierende Verfahren zur Gewinnung quantitativen Wissens (z.B. SPETZLER & STAEL von HOLSTEIN, 1975) sind für größere Gegenstandsbereiche zu zeitaufwendig. Experten sind nicht bereit, Fragen über Hunderte bedingter Wahrscheinlichkeiten zu beantworten. Ansätze zur Gewinnung von Hypothesen über die quantitative Bedeutung von Adverbien wie "wahrscheinlich", "vielleicht" usw. sowie von Modalverben wie "kann", "dürfte" etc. (KIPPER, 1995; TEIGEN, 1988; TEIGEN & BRUN, 1995; WALLSTEN et al., 1986; WALLSTEN & BUDESCU, 1995) sind hier u.E. ebenfalls nicht ausreichend, weil sie multivariate Wahrscheinlichkeitsverteilungen nicht in Betracht ziehen. Dies ist jedoch erforderlich, wenn beispielsweise Aussagen über Einflüsse, Korrelationen, Kovarianzen usw. von Interesse sind.

Nachfolgend wird ein Ansatz dargestellt, mit dem basierend auf der Bayes-Methodologie Hypothesen über quantitative Beziehungen zwischen Variablen auf der Basis qualitativer verbaler Aussagen gewonnen werden können. Ein wichtiger Vorteil des Ansatzes besteht darin, daß er mit einer begrenzten Anzahl von Beurteilungen auskommt.

Methode: Ein Bayes-Netz-basierter Ansatz mit Verteilungshypothesen

Der hier vorgestellte Ansatz zur Gewinnung quantitativer Beziehungen zwischen Variablen aus qualitativen verbalen Relationsbeschreibungen besteht aus vier Schritten:

1. *Repräsentation von Beurteilungssituationen als Bayes-Netz.* Es wird eine beschränkte Menge vorgegebener Häufigkeitsverteilungen festgelegt. Diese Häufigkeitsverteilungen werden als Ausprägungen eines Hypothesenknotens in einem Bayes-Netz betrachtet. Alle Häufigkeitsverteilungen (Hypothesen H_1, H_2, \dots) sind gleich wahrscheinlich. Ferner wird eine ausgewählte Menge verbaler Relationsbeschreibungen (wie z.B. "X übt einen schwachen Einfluß auf Y aus") festgelegt. Jede verbale Relationsbeschreibung wird als Nachfolgerknoten des Hypothesenknotens repräsentiert (Abb. 1).

Abb. 2 zeigt sieben Verteilungshypothesen H_1, H_2, \dots (hier als Wahrscheinlichkeiten dargestellt) über die stochastische Beziehung zwischen zwei binären Variablen. Wegen der gleichen Aprioriwahrscheinlichkeiten ist der Erwartungswert dieser Hypothesen eine nichtinformativ stochastische Relation (Abb. 2 rechts).

2. *Empirische Gewinnung der bedingten Wahrscheinlichkeiten $p(\text{Relationsbeschreibung } i / \text{Hypothese } j)$.* Probanden wird jeweils eine Häufigkeitsverteilung (= Hypothese) zur Zeit vorgelegt. Die Probanden beurteilen für jede der verbalen Relationen, ob sie die vorliegende Häufigkeitsverteilung angemessen beschreibt. Mit der empirischen Gewinnung dieser bedingten Wahrscheinlichkeiten ist das Bayes-Netz vollständig spezifiziert.

3. *Evidenzpropagierung.* Eine konkrete Äußerung (wie z.B. "X spielt eine erhebliche Rolle für Y") wird als Evidenz in dem Bayes-Netz betrachtet. Durch Rückpropagierung wird eine Aposterioriverteilung über den Hypothesen erzeugt.

4. *Erzeugung der erwarteten Aposterioriverteilung.* Anhand der Aposterioriverteilung wird für die Hypothesen eine erwartete Verteilung berechnet. Damit wird eine maßgeschneiderte Verteilung erzeugt, von der wir annehmen, daß die betreffende verbale Äußerung für sie am angemessensten ist. (Dies ist eine empirisch zu prüfende Hypothese.)

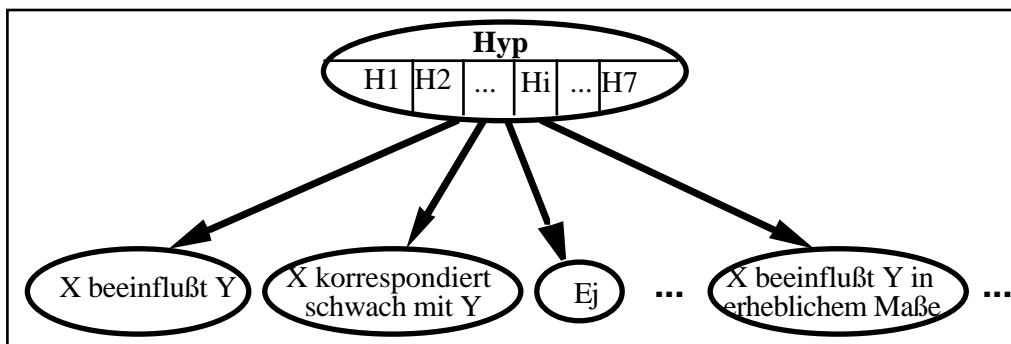


Abb. 1: Repräsentation von Beurteilungssituationen als Bayes-Netz

Ein Problem mit diesem Ansatz besteht jedoch darin, daß durch Linearkombination über den Hypothesen nicht alle Aposterioriverteilungen erzeugt werden können. Um dies zu erreichen, müssen wir von einer minimalen Menge unabhängiger Hypothesen ausgehen (Abb. 3) und das Netz in Abb. 1 durch einen weiteren Hypothesenknoten Hyp' (mit den Verteilungen in Abb. 3 als Hypothesen) erweitern. Hyp' bildet in dem erweiterten Netz den Elternknoten des Hypothesenknotens Hyp in Abb. 1. Die bedingten Verteilungen $p(Hyp | Hyp')$ werden wieder empirisch durch Ähnlichkeitsurteile von Probanden gewonnen.

Vorläufige Ergebnisse:

Der Ansatz wurde zunächst in Voruntersuchungen für bivariate Verteilungen binärer Variablen empirisch erprobt. Probanden schätzten die Angemessenheit verbaler Relationsbeschreibungen für die Verteilungen in Abb. 2 sowie die Ähnlichkeit der Verteilungen in Abb. 2 und 3 ein. Die Ergebnisse lieferten für eine Reihe verbaler Relationsbeschreibungen sinnvolle "maßgeschneiderte" Verteilungen.

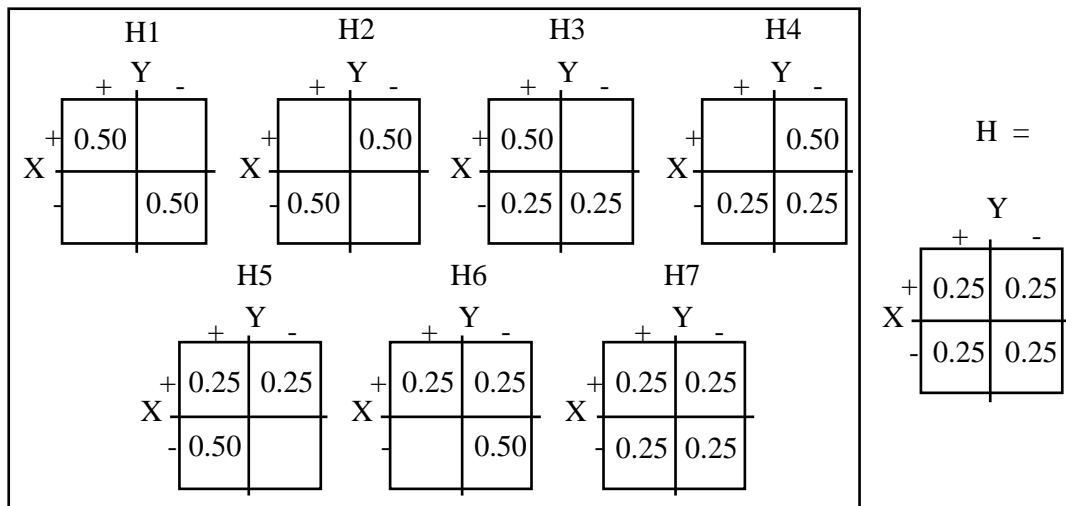


Abb. 2: Einige bivariate Verteilungshypothesen

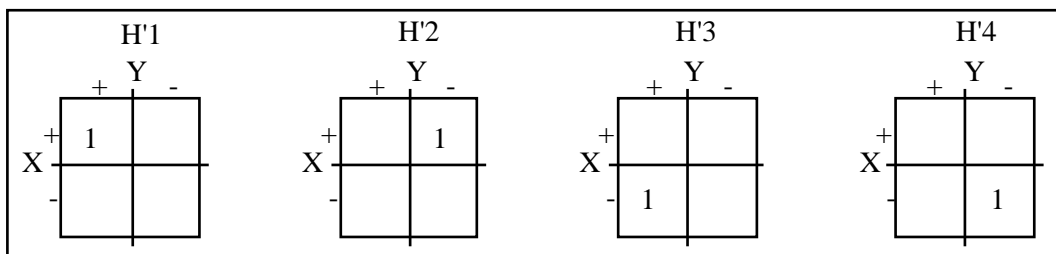


Abb. 3: Für die Generierung aller Aposterioriverteilungen erforderliche Verteilungen

Schlußfolgerung:

Das dargestellte Verfahren scheint uns für die Generierung quantitativer Beziehungen als verbalen Relationsbeschreibungen vielversprechend zu sein. Besonders wichtig ist hierbei, daß nur eine begrenzte Anzahl von Verteilungen zur Beurteilung vorgelegt werden muß.

Für die nähere Zukunft ist eine größere empirische Untersuchung geplant, die auch eine empirische Validierung der gewonnenen Verteilungen vorsieht, beispielsweise:

- Probanden werden aufgefordert, für vorgegebene verbale Relationsbezeichnungen die "zugehörigen" Häufigkeitsverteilungen selbst zu zeichnen oder auszuwählen. Diese sollten mit der "maßgeschneiderten" Verteilung übereinstimmen.
- Probanden werden gebeten, für jede "maßgeschneiderte" Verteilung die angemessensten verbalen Beschreibungen auszuwählen oder selbst zu formulieren. Diese sollten mit der ursprünglichen Formulierung, für die die Verteilung generiert wurde, übereinstimmen.

Literatur:

- KIPPER, B. (1995). Repräsentation und Verarbeitung propositionaler Einstellungen in natürlichsprachlichen Systemen. Dissertation, Universität des Saarlandes, Saarbrücken.
- NAKAO, M.N., AXELROD, S. (1983). Numbers are Better Than Words, *Am. J. of Med.*, 74, 1061-1065.
- PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufman.
- SPETZLER, C.S., STAEL von HOLSTEIN, C.-A.S. (1975). Probability Encoding in Decision Analysis. *Management Science*, 22 (3), 340-358.
- SCHRÖDER, O., MÖBUS, C., THOLE, H.-J. (1996). Acquiring Knowledge from Linguistic Models in Complex, Probabilistic Domains. In P. BRNA, A. PAIVA, J. SELF (eds): *Proceedings of the European Conference on Artificial Intelligence in Education EuroAIED*, Lisbon, Portugal: Colibri, 206-212.
- TEIGEN, K.H. (1988). The Language of Uncertainty, *Acta Psychologica*, 68, 27-38.
- TEIGEN, K.H., BRUN, W. (1995). Yes, But it is Uncertain: Direction and Communicative Intention of Verbal Probab. Terms, *Acta Psych.*, 88, 233-258.
- WALLSTEN, T.S., BUDESCU, D.V., RAPOPORT, A., ZWICK, R., FORSYTH, B. (1986). Measuring the Vague Meanings of Probability Terms, *Journal of Exper. Psychology: General*, 115, 4, 348-365.
- WALLSTEN, T.S., BUDESCU, D.V. (1995). A Review of Human Linguistic Probability Processing: General Principles and Empirical Evidence, *The Knowledge Engineering Review*, 10, 1, 43-62.