Lernen probabilistischer und possibilistischer Netze aus Daten: Theorie und Anwendung

Christian Borgelt und Rudolf Kruse

Otto-von-Guericke-Universität Magdeburg
Institut für Informations- und Kommunikationssysteme
Universitätsplatz 2, 39106 Magdeburg
E-mail: borgelt@iik.cs.uni-magdeburg.de

Guido Lindner

Daimler-Benz AG
Forschung und Technologie FT3S/E

Wilhelm-Runge-Straße 11, 89081 Ulm
E-mail: lindner@str.Daimler-Benz.com

Kurzfassung. Sowohl die seit längerem bekannten probabilistischen als auch die in neuerer Zeit entwickelten possibilistischen Schlußfolgerungsnetze erfreuen sich großer Beliebtheit, wenn es darum geht, das Schließen in hochdimensionalen Räumen handhabbar zu machen. Da es jedoch für einen menschlichen Experten aufwendig und langwierig sein kann, ein Schlußfolgerungsnetz zu erstellen, sucht die aktuelle Forschung verstärkt nach Methoden zum automatischen Erlernen solcher Netze aus Daten. In diesem Artikel geben wir einen Überblick über probabilistische und possibilistische Netzwerke und über die grundlegenden Ideen, wie sie aus Datenbanken von Beispielen gelernt werden können. Anhand einer Anwendung in der Automobilindustrie zeigen wir, daß die vorgestellten Methoden nicht allein von theoretischer Bedeutung, sondern auch praktisch relevant sind.

1 Einleitung

Da das Schließen in hochdimensionalen Räumen — insbesondere bei Vorliegen von Unsicherheit und/oder Impräzision — meist undurchführbar ist, wenn es auf den Gesamtraum erfolgen muß, werden Zerlegungstechniken immer beliebter, durch die das Ziehen von Schlußfolgerungen auf Berechnungen in niedrigdimensionalen Unterräumen beschränkt werden kann. Vor allem im Bereich der graphischen Modellierung werden Zerlegungstechniken untersucht, die Abhängigkeiten und Unabhängigkeiten zwischen Variablen ausnutzen [19]. Zu den am besten bekannten Ansätzen dieser Art gehören Bayessche Netze [25], Markov Netze [22], sowie die allgemeineren bewertungsbasierten (valuation-based) Netze [33]. In neuerer Zeit haben außerdem possibilistische Netze durch ihre enge Verwandtschaft mit Fuzzy-Methoden einige Beachtung erlangt [20]. Alle genannten Ansätze führten zur Entwicklung effizienter Programmsysteme, z.B. HUGIN [1], PULCINELLA [30], PATHFINDER [13] und POSSINFER [9].

In diesem Artikel geben wir einen Überblick über die wesentlichen Ideen probabilistischer und possibilistischer Netze und die Methoden, mit denen sie aus Daten gelernt werden können, d.h. mit denen aus einer Datenbank von Beispielen eine für Schlußfolgerungen geeignete Zerlegung der zugrundeliegenden Wahrscheinlichkeits- oder Possibilitätsverteilung bestimmt werden kann [7, 14, 10, 11]. Solch automatisiertes Lernen ist wichtig, da die Konstruktion eines Netzwerkes durch einen menschlichen Experten aufwendig und langwierig sein kann. Wenn, wie es oft der Fall ist, eine Datenbank von Beispielen vorliegt, können Lernalgorithmen wenigstens einen Teil der Konstruktionsarbeit übernehmen.

Diese neuen Methoden können zu „Data Mining“ benutzt werden, d.h. zur Gewinnung nützlichen Wissens aus umfangreichen Datenbeständen. Wir zeigen die praktische Relevanz dieser Ansätze anhand einer Anwendung in der Automobilindustrie, in der die Induktion von probabilistischen Netzwerken benutzt wurde, um nach Schwachstellen in Mercedes-Benz Fahrzeugen zu suchen. So gewonnenes Wissen kann, indem die gefundenen Schwachstellen beseitigt werden, zur Erhöhung der Produktqualität beitragen.

2 Probabilistische und possibilistische Netze

Die wesentliche Voraussetzung, die jedem Schlußfolgerungsnetzwerk, sei es nun ein probabilistisches oder ein possibilistisches, zugrunde liegt, ist, daß eine hochdimensionale Verteilung ohne großen Informationsverlust zerlegt werden kann in eine Menge (überlappender) niedrigdimensionaler Verteilungen.[1] Diese Menge niedrigdimensionaler Verteilungen wird gewöhnlich durch einen Hypergraphen[2] dargestellt, in dem jeder Knoten für ein Attribut (bzw. eine Variable) und jede Hyperkante für eine Verteilung der Zerlegung steht. Jedem Knoten und jeder Hyperkante wird eine Projektion der hochdimensionalen Verteilung (eine Marginalverteilung) zugeordnet: dem Knoten eine Projektion auf das ihm zugeordnete Attribut, der Hyperkante eine Projektion auf den Unterraum, der durch die in ihr enthaltenen Attribute gebildet wird.

Die Hyperkanten stellen direkte Einflüße dar, die die durch sie verbundenen Attribute aufeinander haben, d.h. sie beschreiben, wie sich Einschränkungen der möglichen Werte eines Attributes auf die Wahrscheinlichkeiten oder Possibilitätsgrade der Werte der anderen Attribute der Hyperkante auswirken. Schlußfolgerungen werden in einem solchen Hypergraphen gezogen, indem Evidenz, d.h. beobachtete Einschränkungen der möglichen Werte einiger Attribute, entlang der Hyperkanten propagiert wird.

Die Idee der Propagation kann am besten an einem einfachen Beispiel erläutert werden. Gegeben seien drei Variablen, X, Y, und Z, und ein (Hyper-)Graph X—Y—Z. Wird Evidenz über den Wert der Variable X eingegeben, so wird diese folgendermaßen weitergeleitet: Die durch die Evidenz gegebenen Einschränkungen der möglichen Werte der Variable X werden auf den Unterraum $\{X, Y\}$ erweitert, um so Einschränkungen auf Tupeln (x_i, y_j) zu erhalten. Diese werden dann auf die Variable Y projiziert, um die Einschränkungen der möglichen Werte dieser Variable zu bestimmen. Die Einschränkungen der Werte der Variable Y werden anschließend auf den Unterraum $\{Y, Z\}$ erweitert und auf die Variable Z projiziert.

Damit dieses Verfahren ausführbar ist, müssen die Hauptoperationen, Erweiterung und Projektion, gewisse Bedingungen erfüllen, die sich durch Axiome beschreiben lassen [31]. In probabilistischen Netzen wird eine Produkt-Summe-Propagation benutzt, in der die Marginalverteilungen z.B. eines zweidimensionalen Unterraumes durch Summenbildung über eine Dimension berechnet wird, d.h. $P(x_i) = \sum_j P(x_i, y_j)$. Der Erweiterungsschritt besteht in der Multiplikation der A-priori-Wahrscheinlichkeiten auf dem Oberraum mit dem Quotienten aus A-posteriori- und A-priori-Wahrscheinlichkeit auf dem Unterraum.

Für unser Beispiel ist dies in den Abbildungen 1 und 2 dargestellt. Abbildung 1 zeigt eine dreidimensionale Wahrscheinlichkeitsverteilung auf dem gemeinsamen Wertebereich der Variablen X, $\mathrm{dom}(X) = \{x_1, x_2, x_3, x_4\}$, Y, $\mathrm{dom}(Y) = \{y_1, y_2, y_3\}$, und C, $\mathrm{dom}(Z) = \{z_1, z_2, z_3\}$, sowie die zugehörigen Marginalverteilungen (Zeilen-/Spaltensummen). Da in dieser Verteilung die Gleichungen

$$\forall i, j, k : P(x_i, y_j, z_k) = \frac{P(x_i, y_j) P(y_j, z_k)}{P(y_j)}$$

gelten, kann sie in die Marginalverteilungen auf den Unterräumen $\{X, Y\}$ und $\{Y, Z\}$ zerlegt werden. Deshalb ist es auch möglich, Schlußfolgerungen aus der Beobachtung, daß die Variable X den Wert x_4 hat, mit Hilfe des in Abbildung 2 gezeigten Schemas zu ziehen.[3] In diesem Schema bezeichnet *alt* die A-priori-Wahrscheinlichkeiten, *neu* die sich nach der Einbeziehung der Evidenz durch Erweiterung und Projektion ergebenden A-posteriori-Wahrscheinlichkeiten. Man prüft leicht nach, daß die sich ergebenden A-posteriori-Marginalverteilungen die gleichen wie jene sind, die man bei einer Berechnung direkt im dreidimensionalen Raum erhalten hätte.

Wenden wir uns nun den possibilistischen Netzen zu. Ihre Entwicklung wurde ausgelöst durch die Tatsache, daß probabilistische Netze zwar hervorra-

[1] Diese Voraussetzung muß natürlich nicht erfüllt sein. Eine Verteilung kann unzerlegbar sein, selbst wenn man einen gewissen Informationsverlust in Kauf nimmt. In einem solchen Fall können Schlußfolgerungsnetze dann leider nicht verwendet werden.

[2] Der Unterschied zwischen einem normalen Graphen und einem Hypergraphen besteht lediglich darin, daß eine Kante in einem normalen Graphen nur genau zwei, in einem Hypergraphen dagegen im Prinzip beliebig viele Knoten verbinden kann. Jeder normale Graph ist folglich auch ein Hypergraph.

[3] Bei diesem Schema handelt es sich um eine Vereinfachung, die für eine Implemetierung nicht direkt brauchbar ist. Insbesondere zum Zusammenführen von Evidenz aus zwei (Hyper-)Kanten sind zusätzliche Berechnungen nötig, die hier vernachlässigt wurden.

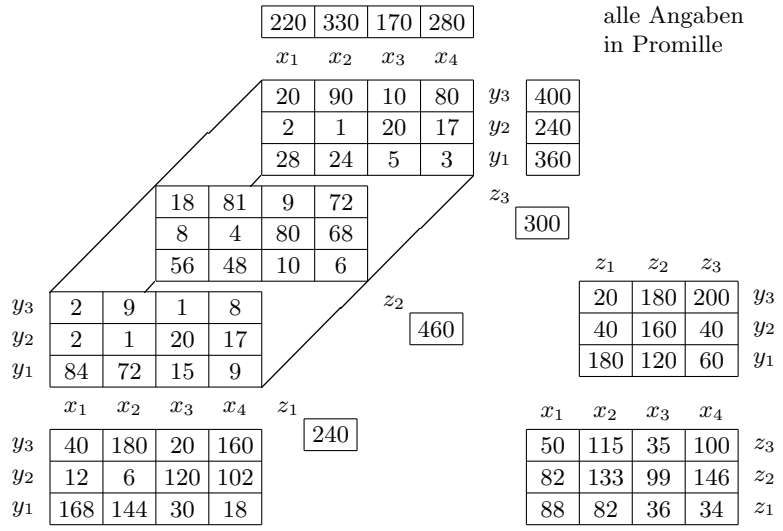

Abbildung 1: Eine dreidimensionale Wahrscheinlichkeitsverteilung mit Marginalverteilungen (Zeilen-/Spaltensummen). Da für diese Verteilung die Gleichungen $\forall i, j, k:$ $P(x_i, y_j, z_k) = \dfrac{P(x_i, y_j)P(y_j, z_k)}{P(y_j)}$ gelten, kann sie in die Marginalverteilungen auf den Unterräumen $\{X, Y\}$ und $\{Y, Z\}$ zerlegt werden.

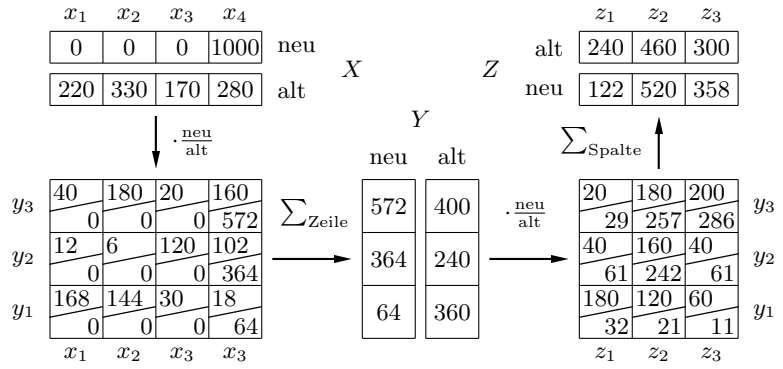

Abbildung 2: Propagation der Beobachtung, daß das Attribut X den Wert x_4 hat, in der dreidimensionalen Wahrscheinlichkeitsverteilung aus Abbildung 1 unter ausschließlicher Verwendung der Marginalverteilungen auf den Unterräumen $\{X, Y\}$ und $\{Y, Z\}$.

gend geeignet sind, um *unsichere* Information darzustellen und zu verarbeiten, die Einbeziehung *impräziser* Information jedoch Schwierigkeiten bereitet. Gerade die Einbeziehung impräziser Information wird jedoch für die praktische Verwertbarkeit von Lern- und Schlußfolgerungsverfahren für immer wichtiger gehalten. Unter impräziser Information verstehen wir dabei das Wissen, daß der Wert eines Attributes in einer bestimmten Menge von Werten liegt, zwischen diesen Werten aber nicht mehr — weder durch Angabe von Wahrscheinlichkeiten noch durch Festlegung von Präferenzen — unterschieden werden kann.

Der Unterschied zwischen unsicherer und impräziser Information läßt sich am besten anhand der Interpretation eines Possibilitätsgrades präzisieren. Der von uns bevorzugte Interpretationsansatz stützt sich auf das Kontextmodell [8, 20]. In diesem Modell werden Possibilitätsverteilungen gesehen als informationskomprimierte Darstellungen (nicht notwendigerweise geschachtelter) zufälliger Mengen (random sets); ein Possibilitätsgrad als die Ein-Punkt-Überdeckung (one-point coverage) einer zufälligen Menge [24].

Genauer bedeutet dies folgendes: Sei ω_0 der wahre, aber unbekannte Zustand des modellierten Weltausschnitts, der in einer Menge Ω möglicher Weltzustände liegt. Wir nehmen an, daß wir bei der Untersuchung des Weltausschnitts eine Menge von *Kontexten* $C = \{c_1, c_2, \ldots, c_m\}$ unterscheiden können, über denen eine Wahrscheinlichkeitsverteilung P angegeben werden kann. Diese Kontexte sollen z.B. physikalische oder beobachtungsabhängige Rahmenbedingungen widerspiegeln. Wir nehmen weiter an, daß wir für jeden Kontext c eine Menge $\gamma(c)$ von Zuständen auszeichnen können, von der wir sicher sind, daß ω_0 in ihr enthalten ist. Die Menge $\gamma(c)$ soll die *spezifischste* solche Menge

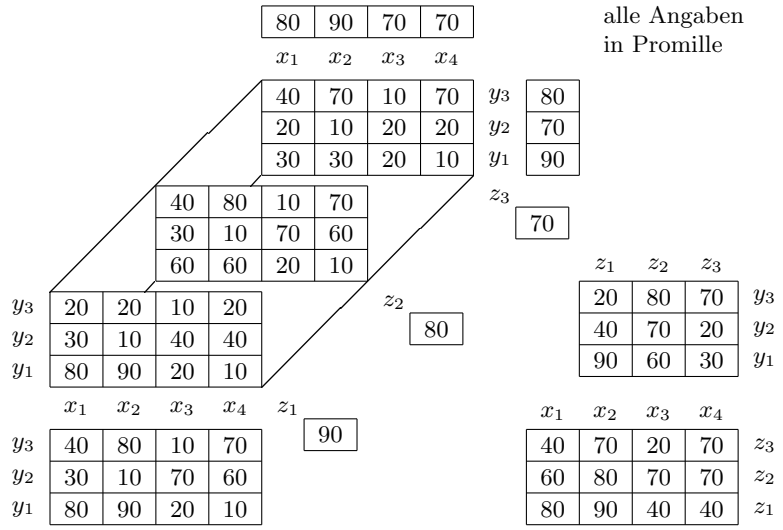

Abbildung 3: Eine dreidimensionale Possibilitätsverteilung mit zugehörigen Maximumprojektionen. Da für diese Verteilung die Gleichungen
$\forall i,j,k : \pi(x_i,y_j,z_k) = \min_j(\max_i \pi(x_i,y_j,z_k), \max_k \pi(x_i,y_j,z_k))$
gelten, kann sie in die Maximumprojektionen auf die Unterräume $\{X,Y\}$ und $\{Y,Z\}$ zerlegt werden.

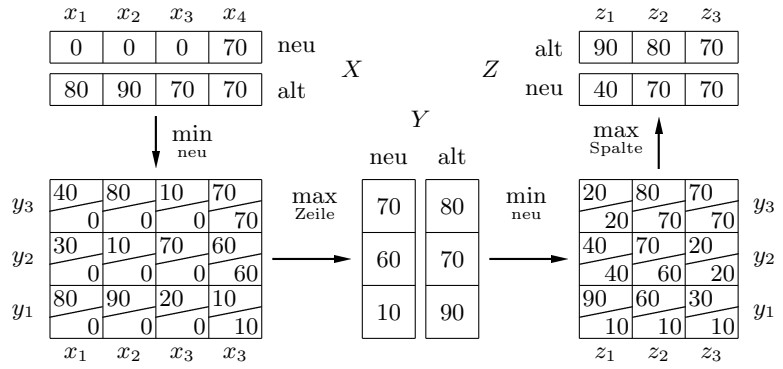

Abbildung 4: Propagation der Beobachtung, daß das Attribut X den Wert x_4 hat, in der dreidimensionalen Possibilitätsverteilung aus Abbildung 3 unter ausschließlicher Verwendung der Maximumprojektionen auf die Unterräume $\{X,Y\}$ und $\{Y,Z\}$.

sein, d.h. für jede echte Teilmenge von $\gamma(c)$ soll es nicht sicher sein, ob ω_0 in ihr enthalten ist. Man nennt $\gamma(c)$ daher die *spezifischste korrekte mengenwertige Beschreibung* von ω_0.

Offenbar erhält man mit diesem Modell eine rein wahrscheinlichkeitsbasierte Beschreibung, wenn alle Mengen $\gamma(c)$ genau ein Element enthalten. In diesem Sinne ist die wahrscheinlichkeitsbasierte Beschreibung präzise: innerhalb eines Kontextes ist der Zustand eindeutig festgelegt. Läßt man dagegen zu, daß $\gamma(c)$ mehr als ein Element enthält, kann impräzise Information berücksichtigt werden. Die sich ergebende *zufällige Menge* (random set) $\Gamma = (\gamma, P)$ ist eine imperfekte (d.h. impräzise *und* unsichere) Beschreibung von ω_0.

Sei nun π_Γ die *Ein-Punkt-Überdeckung* (one-point coverage) von Γ (die durch Γ induzierte *Possibilitätsverteilung*), die durch

$$\pi_\Gamma : \Omega \to [0,1], \quad \pi_\Gamma(\omega) = P\left(\{c \in C \mid \omega \in \gamma(c)\}\right)$$

definiert ist. In einer vollständigen Modellierung müssen zwar die Kontexte im Detail angegeben werden, um ihre Beziehungen aufzudecken, doch wenn die Kontexte unbekannt sind oder vernachlässigt werden, wird durch $\pi_\Gamma(\omega)$ wenigstens die Wahrscheinlichkeitsmasse der Kontexte c angegeben, in denen $\omega = \omega_0$ möglich ist. Diese Masse quantifiziert die Möglichkeit der Wahrheit der Aussage „$\omega = \omega_0$" [8].

Ausgehend von dieser Interpretation läßt sich die Theorie possibilistischer Netze weitgehend analog zu der der probabilistischen Netze aufbauen. Der einzige Unterschied besteht darin, daß statt einer Produkt-Summe-Propagation eine Minimum-Maximum-Propagation verwendet wird. Das heißt, die Projektion z.B. einer zweidimensionalen Verteilung wird durch Maximumbildung über eine Dimension bestimmt, der Erweiterungschritt besteht in der Berechnung des Minimums der A-priori-

Possibilitätsverteilung auf dem Oberraum und der A-posteriori-Verteilung auf dem Unterraum. Dies ist notwendig, um die impräzise Information angemessen zu behandeln, da eine Summation von Possibilitätsgraden wegen der dadurch möglichen mehrfachen Berücksichtigung von $P(c)$ problematisch ist. Damit ändert sich allerdings auch die Interpretation der Verteilungen. Sie beziehen sich, auch wenn sie Unterräumen zugeordnet sind, stets auf vollständige Vektoren über alle zur Beschreibung des Weltausschnitts verwendeten Attribute und nicht mehr, wie bei Wahrscheinlichkeitsverteilungen, nur auf die Attribute des Unterraums.

Für unser Beispiel ist eine possibilistische Beschreibung in den Abbildungen 3 und 4 dargestellt. Abbildung 3 zeigt eine dreidimensionale Possibilitätsverteilung auf dem gemeinsamen Wertebereich der drei Variablen X, Y und Z sowie die zugehörigen, durch Maximumprojektion bestimmten Marginalverteilungen. Da in dieser Verteilung die Gleichungen

$$\forall i, j, k : \pi(x_i, y_j, z_k) = \min_j(\max_i \pi(x_i, y_j, z_k), \max_k \pi(x_i, y_j, z_k))$$

gelten, kann sie in die Marginalverteilungen auf den Unterräumen $\{X, Y\}$ und $\{Y, Z\}$ zerlegt werden. Deshalb ist es auch möglich, Schlußfolgerungen aus der Beobachtung, daß die Variable X den Wert x_4 hat, mit Hilfe des in Abbildung 4 gezeigten Schemas zu ziehen. Wieder sind die so erhaltenen Marginalverteilungen die gleichen wie jene, die man aus einer direkten Schlußfolgerung im dreidimensionalen Raum erhalten hätte.

3 Lernen aus Daten

Das Lernen eines probabilistischen oder possibilistischen Schlußfolgerungsnetzes besteht darin, eine gegebene mehrdimensionale Wahrscheinlichkeits- oder Possibilitätsverteilung in Verteilungen auf Unterräume zu zerlegen. Die zu zerlegende Verteilung ist dabei jedoch nicht direkt gegeben, sondern es steht nur eine Datenbank von Beispielen zur Verfügung. Diese wird benutzt, um (bedingte) relative Häufigkeiten auszuzählen, aus denen die (bedingten oder marginalen) Wahrscheinlichkeiten und Possibilitätsgrade geschätzt werden. (Gewöhnlich wird dabei jeder Datensatz als ein Kontext angesehen.)

Ein Algorithmus zum Lernen von Schlußfolgerungsnetzen aus Daten besteht immer aus zwei Teilen: einem Bewertungsmaß und einer Suchmethode. Mit Hilfe des Bewertungsmaßes wird die Güte einer gegebenen Zerlegung (eines gegebenen Hypergraphen) eingeschätzt, während die Suchmethode bestimmt, welche Zerlegungen (welche Hypergraphen) überhaupt betrachtet werden. Oft kann das Bewertungsmaß auch benutzt werden, um die Suche zu steuern, da es gewöhnlich das Ziel ist, seinen Wert zu maximieren (oder zu minimieren).

Es gibt eine Vielzahl von Bewertungsmaßen, sowohl für das Lernen probabilistischer, als auch für das Lernen possibilistischer Netzwerke. Natürlich können wir hier nicht alle im Detail besprechen (es sei auf [3, 4] verwiesen) und führen daher nur eine (unvollständige) Liste an. Alle aufgeführten Maße haben die wünschenswerte Eigenschaft, daß sie sich lokal, d.h. auf Teilnetzen bzw. einzelnen Hyperkanten, berechnen lassen. Die Gesamtbewertung wird aus diesen Einzelbewertungen zusammengesetzt.

Probabilistische Maße

- χ^2-Maß
- Informationsgewinn/wechselseitige Information (information gain/mutual inform.) [21, 26, 27]
- (symmetrisches) Informationsgewinnverhältnis [26, 27, 23]
- Gini-Index [5]
- symmetrischer Gini-Index [34]
- Minimale Beschreibungslänge mit relativer oder absoluter Häufigkeitscodierung [28, 17]
- Stochastische Komplexität [18, 29]
- g-Funktion (ein Bayessches Maß) [7]

Possibilistische Maße

- d_{χ^2}, abgeleitet vom χ^2-Maß [3, 4]
- d_{mi}, abgeleitet von wechselseitiger Information [3, 4]
- Spezifizitätsgewinn (specificity gain) [10, 2]
- (symmetr.) Spezifizitätsgewinnverhältnis [2]

Viele dieser Maße stammen ursprünglich aus dem Entscheidungsbaumlernen. Die den meisten dieser Maße zugrundeliegende Idee erläutern wir, indem wir zwei von ihnen als Beispiele herausgreifen, und zwar die eng verwandten Maße Informationsgewinn und Spezifizitätsgewinn.

Der Informationsgewinn ist für zwei Variablen X und Y definiert als

$$\begin{aligned} I_{\text{gain}} &= H_X - H_{X|Y} = H_X + H_Y - H_{XY} \\ &= \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)}, \end{aligned}$$

wobei H die Shannonsche Entropie ist [32]. Betrachtet man die Entropieschreibweise, so vergleicht er die Entropie der Verteilung über den Werten von X (also den „Grad der Unkenntnis“ des Wertes von X) mit und ohne Kenntnis des Wertes von Y, und mißt so die Information (Verringerung der Unkenntnis, gemessen in Bit) die man im Durchschnitt durch die Kenntnis des Wertes von Y über den Wert von X gewinnt. In der anderen Schreibweise kann er als Maß für die Differenz der gemeinsamen Verteilung $P(x_i, y_j)$ und der unabhängigen Verteilung $P(x_i)P(y_j)$ gedeutet werden [21]. In beiden Fällen mißt er, anschaulich gesprochen, die Stärke der Abhängigkeit von Variablen, und legt daher nahe, solche (Hyper-)Kanten in den (Hyper-)Graphen aufzunehmen, für die er besonders groß ist. Eine Erweiterung der obigen Definition auf mehr als zwei Variablen ist leicht zu finden [3, 4].

Der Spezifizitätsgewinn stützt sich auf das U-Unsicherheitsmaß der *Nichtspezifizität* einer Possibilitätsverteilung [16], welches als

$$\text{nsp}(\pi) = \int_0^{\sup(\pi)} \log_2 |[\pi]_\alpha| d\alpha$$

definiert ist und als Verallgemeinerung der Hartley-Information [12] auf den possibilistischen Fall gerechtfertigt werden kann [15]. $\text{nsp}(\pi)$ beschreibt die zu erwartende Menge an Information (gemessen in Bit), die noch hinzugefügt werden muß, um den tatsächlichen Wert innerhalb der Menge $[\pi]_\alpha$ von Alternativen zu bestimmen, wobei eine Gleichverteilung auf der Menge $[0, \sup(\pi)]$ der möglichen possibilistischen Vertrauensgrade α angenommen wird [11].

Die Rolle, die die Nichtspezifität in der Possibilitätstheorie spielt, ist derjenigen der Entropie in der Wahrscheinlichkeitstheorie vergleichbar. Es liegt daher nahe, aus der Nichtspezifizität ein Bewertungsmaß in der gleichen Weise zu konstruieren wie der Informationsgewinn aus der Entropie konstruiert werden kann, d.h. durch die Berechnung des Gewinnes an Spezifizität, die sich aus der Verwendung der gemeinsamen anstelle der marginalen Verteilungen ergibt. Wir definieren daher für zwei Variablen X und Y den *Spezifizitätsgewinn* als

$$S_{\text{gain}} = \text{nsp}(\pi_{\max X}) + \text{nsp}(\pi_{\max Y}) - \text{nsp}(\pi_{XY}).$$

Auch diese Definition läßt sich leicht auf mehr als zwei Variablen erweitern [3, 4]. Das sich ergebende Maß ist equivalent zu dem in [11] definierten.

Alle der oben genannten Maße lassen sich in Verbindung mit einer Vielzahl von Suchmethoden verwenden. Die beiden am häufigsten verwendeten Methoden sind die Bestimmung eines optimalen spannenden Baumes [6], die gleichzeitig auch die älteste ist, sowie die gierige (greedy) Elternauswahl [7] (K2-Algorithmus). Im Prinzip lassen sich beliebige heuristische Suchverfahren, wie z.B. simuliertes Ausglühen (simulated annealing), genetische Algorithmen etc., nutzen.

4 Anwendung in der Automobilindustrie

Selbst so qualitativ hochwertige Produkte wie Mercedes-Benz-Fahrzeuge zeigen hin und wieder unerwünschtes Verhalten. Da es eines der Hauptziele der Mercedes-Benz AG ist, die Qualität ihrer Fahrzeuge noch weiter zu verbessern, wird erheblicher Aufwand getrieben, um die Ursachen eines festgestellten Fehlverhaltens ausfindig zu machen und so ein Wiederauftreten zu verhindern. Zu diesem Zweck pflegt Mercedes-Benz eine Datenbank, in die für jedes produzierte Fahrzeug sein Bauzustand (Baureihe, Motorbaureihe, Sonderausstattungen etc.) sowie jegliche Fehler, Schäden und Beanstandungen, die während der Produktion oder der Gewährleistungsfrist aufgetreten sind, eingetragen werden.

In einer Kooperation zwischen der Otto-von-Guericke-Universität Magdeburg und der Data Mining und Machine Learning Gruppe des Forschungszentrums Ulm der Daimler-Benz AG wurde das vom ersten Author dieses Artikels entwickelte Programm INES (Induktion von NEtzwerkStrukturen), eine prototypische Implementierung der oben angesprochenen Verfahren, auf Ausschnitte dieser Datenbank angewendet. Dieses Programm enthält alle aufgeführten Bewertungsmaße und die beiden genannten Suchmethoden: Bestimmung optimaler spannender Bäume und gierige (greedy) Elternauswahl.

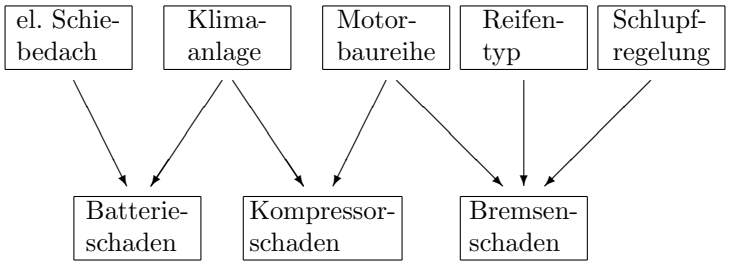

Abbildung 5: Ein Ausschnitt eines fiktiven zweischichtigen Netzes, das die Abhängigkeiten zwischen Schäden/Fehlern (untere Schicht) und Bauzustandsmerkmalen (obere Schicht) beschreibt. Übereinstimmungen mit tatsächlichen Abhängigkeiten sind rein zufällig.

(fiktive) Häufigkeit von Batterieschäden		Klimaanlage mit	Klimaanlage ohne
elektrisches Schiebedach	mit	9 %	3 %
	ohne	3 %	2 %

Abbildung 6: Ein fiktives Teilnetz, das die Abhängigkeit eines Batterieschadens vom Vorhandensein eines elektrischen Schiebedaches und einer Klimaanlage beschreibt.

Die Idee, von der wir in dieser Anwendung ausgingen, ist sehr einfach. Da man an Ursachen von Fehlern interessiert ist, wird ein zweischichtiges probabilistisches Netzwerk gelernt, deren obere Schicht diejenigen Attribute enthält, die den Bauzustand eines Fahrzeugs beschreiben, während die Attribute in der unteren Schicht mögliche Schäden oder Fehler wiedergeben (siehe Abbildungen 5 und 6). (Da echte Zahlen und Abhängigkeiten natürlich streng vertraulich sind, zeigen beide Bilder fiktive Daten. Jede Ähnlichkeit mit echten Zahlen und Abhängigkeiten ist rein zufällig.) Abbildung 5 zeigt ein mögliches zweischichtiges Netzwerk, Abbildung 6 die Häufigkeitsverteilung, die zu seinem ersten Teilnetz gehört. Da in diesem Beispiel die Batterieschadensrate für Fahrzeuge mit Klimaanlage und elektrischem Schiebedach deutlich höher ist als für solche mit keinem oder nur einem dieser Ausstattungsmerkmale, kann man vermuten, daß der durch sie hervorgerufene erhöhte Stromverbrauch zu häufigeren Batterieausfällen führt.

Hier ist zu bemerken, daß das Lernen eines probabilistischen Netzwerkes (da die verwendete Datenbank präzise Beschreibungen enthält, bringen possibilistische Netzwerke wenig ein) mit der obigen Struktur dem Lernen eines „Waldes“ von Entscheidungsbäumen aus folgenden Gründen vorzuziehen ist: Erstens sind Fehler selten. Entscheidungsbäume müssen aber in einem Blatt mindestens eine Fehlerhäufigkeit von 50% erreichen, damit sie als Klassifikation „Fehler“ ausgeben, anderenfalls wird die Verzweigung eliminiert. Folglich müssen die Fehlerhäufigkeiten in den Daten manipuliert werden, um überhaupt Entscheidungsbaumlerner anwenden zu können. Diese Manipulationen führen aber entweder zu einer starken Verringerung der Zahl der zur Verfügung stehenden Datensätze oder zu Verzerrungen des Datenmaterials. Zweitens ist, anders als in Entscheidungsbäumen, die Beschreibung der Abhängigkeiten in den Teilnetzen eines probabilistischen Netzwerks symmetrisch in bezug auf die Elternattribute. Dadurch können aus dem Ergebnis mit geringem Aufwand die Abhängigkeiten für eine verringerte Zahl von Elternattributen berechnet werden, was die nachfolgende Analyse der Ergebnisse erheblich erleichtert.

Obwohl spezifische Ergebnisse streng vertraulich sind, können wir hier bemerken, daß das Programm INES mittlerweile als ein zusätzliches Hilfsmittel für reale Ursachenanalysen im Nutzfahrzeugbereich bei Mercedes-Benz dient. Zweck des Einsatzes ist die Eingrenzung von Fehlerursachen, um u.a. aufwendige technische Prüfungen zu minimieren. An zwei typischen Anwendungsbeispielen wollen wir dies veranschaulichen. Die erste Anwendung betrifft ein Getriebeproblem, dessen Ursache zwar bereits bekannt war, dessen Analyse für die Fahrzeugexperten jedoch einen hohen Zeitaufwand bedeutet hatte. Auf der Grundlage der gleichen Informationen, die den Experten zur Verfügung gestanden hatten, konnte INES die auf den verursachenden Fehler deutende Abhängigkeit ohne Schwierigkeiten in erheblich kürzerer Zeit finden.

Die zweite Anwendung von INES wurde zu einem aktuellen Problem im Nutzfahrzeugbereich durchgeführt. Bei dieser Ursachenanalyse war zum Zeitpunkt der Analyse nicht bekannt, ob es eine Abhängigkeit gibt, die auf die mögliche Fehlerursache schließen läßt. In dieser Anwendung wurde von

INES keine Abhängigkeit zwischen Bauzuständen und aufgetretenen Fehlern gefunden, die auf eine mögliche Ursache hindeuteten. Nicht zuletzt wegen dieses Ergebnisses konzentrierte sich die Ursachenanalyse dann auf andere Bereiche. Am Ende wurde die Ursache bei einem Zulieferer ausgemacht, der nachträglich die Qualität eines Schmierstoffes geändert hatte. Diese Information war nicht in der Datenbank, auf die INES angesetzt wurde, abgelegt; die entsprechende Abhängigkeit konnte daher nicht gefunden werden.

Diese Beispiele zeigen, daß mit Hilfe von INES mögliche Fehlerursachen effizient und effektiv eingegrenzt werden können. Die Experten sind nun in der Lage, eine viel größere Kombination von Bauzuständen zu analysieren, als dies „konventionell" machbar wäre. Die von den Experten gesehenen Vorteile sind im wesentlichen die Möglichkeit einer zielgerichteteren und automatischen Suche und die enorme Zeitersparnis gegenüber der „manuellen" Analyse durch einzelne Anfragen an die Datenbank.

Der Einsatz von INES belegt, daß durch das automatisierte Lernen von Schlußfolgerungsnetzen aus Daten, wie in den beschriebenen Anwendungen, die Ursachenanalyse eines Experten unterstützt werden kann. Wie das zweite Anwendungsbeispiel allerdings auch deutlich macht, kann ein solches Verfahren in einer offenen Welt immer nur Hilfmittel sein.

5 Zusammenfassung

Wir haben in diesem Aufsatz versucht, einen wenn auch sehr knappen Überblick über das Lernen probabilistischer und possibilistischer Schlußfolgerungsnetze zu geben. (Den an Details interessierten Leser müssen wir auf die unten angegebene Literatur verweisen.) Die angesprochenen Verfahren eignen sich besonders, wenn Abhängigkeiten zwischen einer großen Zahl von Attributen untersucht werden sollen, denn sie versuchen, die bestehenden Abhängigkeiten in kleinen Unterräumen zu beschreiben und so die Komplexität des betrachteten Weltausschnitts handhabbar zu machen. Daß sie nicht nur von theoretischem Interesse, sondern auch von pratischer Bedeutung sind, zeigt die Anwendung bei Mercedes-Benz, in der sie zur Lösung realer Probleme eingesetzt werden.

Literatur

[1] S.K. Andersen, K.G. Olesen, F.V. Jensen, und F. Jensen. HUGIN — A shell for building Bayesian belief universes for expert systems. *Proc. 11th Int. J. Conf. on Artificial Intelligence*, 1080–1085, 1989

[2] C. Borgelt, J. Gebhardt, und R. Kruse. Concepts for Probabilistic and Possibilistic Induction of Decision Trees on Real World Data. *Proc. of the EUFIT'96*, Vol. 3:1556–1560, 1996

[3] C. Borgelt und R. Kruse. Evaluation Measures for Learning Probabilistic and Possibilistic Networks. *Proc. of the FUZZ-IEEE'97*, Vol. 2:pp.669–676, 1997

[4] C. Borgelt und R. Kruse. Some Experimental Results on Learning Probabilistic and Possibilistic Networks with Different Evaluation Measures. *Proc. of the ECSQARU/FAPR'97*, 1997

[5] L. Breiman, J.H. Friedman, R.A. Olshen, und C.J. Stone. *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984

[6] C.K. Chow und C.N. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Trans. on Information Theory* 14(3):462–467, IEEE 1968

[7] G.F. Cooper und E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9:309–347, Kluwer 1992

[8] J. Gebhardt und R. Kruse. A Possibilistic Interpretation of Fuzzy Sets in the Context Model. *Proc. IEEE Int. Conf. on Fuzzy Systems*, 1089-1096, San Diego 1992.

[9] J. Gebhardt und R. Kruse. POSSINFER — A Software Tool for Possibilistic Inference. In: D. Dubois, H. Prade, and R. Yager, eds. *Fuzzy Set Methods in Information Engineering: A Guided Tour of Applications*, Wiley 1995

[10] J. Gebhardt und R. Kruse. Learning Possibilistic Networks from Data. *Proc. 5th Int. Workshop on Artificial Intelligence and Statistics*, 233–244, Fort Lauderdale, 1995

[11] J. Gebhardt und R. Kruse. Tightest Hypertree Decompositions of Multivariate Possibility Distributions. *Proc. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems*, 1996

[12] R.V.L. Hartley. Transmission of Information. *The Bell Systems Technical Journal* 7:535–563, 1928

[13] D. Heckerman. *Probabilistic Similarity Networks*. MIT Press 1991

[14] D. Heckerman, D. Geiger, und D.M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243, Kluwer 1995

[15] M. Higashi und G.J. Klir. Measures of Uncertainty and Information based on Possibility Distributions. *Int. Journal of General Systems* 9:43–58, 1982

[16] G.J. Klir und M. Mariano. On the Uniqueness of a Possibility Measure of Uncertainty and Information. *Fuzzy Sets and Systems* 24:141–160, 1987

[17] I. Kononenko. On Biases in Estimating Multi-Valued Attributes. *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining*, 1034–1040, Montreal, 1995

[18] R.E. Krichevsky und V.K. Trofimov. The Performance of Universal Coding. *IEEE Trans. on Information Theory*, IT-27(2):199–207, 1983

[19] R. Kruse, E. Schwecke, und J. Heinsohn. *Uncertainty and Vagueness in Knowledge-based Systems: Numerical Methods*. Series: Artificial Intelligence, Springer, Berlin 1991

[20] R. Kruse, J. Gebhardt, und F. Klawonn. *Foundations of Fuzzy Systems*, John Wiley & Sons, Chichester, England 1994

[21] S. Kullback und R.A. Leibler. On Information and Sufficiency. *Ann. Math. Statistics* 22:79–86, 1951

[22] S.L. Lauritzen und D.J. Spiegelhalter. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, 2(50):157–224, 1988

[23] R. Lopez de Mantaras. A Distance-based Attribute Selection Measure for Decision Tree Induction. *Machine Learning* 6:81–92, Kluwer 1991

[24] H.T. Nguyen. Using Random Sets. *Information Science* 34:265–274, 1984

[25] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition)*. Morgan Kaufman, New York 1992

[26] J.R. Quinlan. Induction of Decision Trees. *Machine Learning* 1:81–106, 1986

[27] J.R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993

[28] J. Rissanen. A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics* 11:416–431, 1983

[29] J. Rissanen. Stochastic Complexity and Its Applications. *Proc. Workshop on Model Uncertainty and Model Robustness*, Bath, England, 1995

[30] A. Saffiotti und E. Umkehrer. PULCINELLA: A General Tool for Propagating Uncertainty in Valuation Networks. *Proc. 7th Conf. on Uncertainty in AI*, 323–331, San Mateo 1991

[31] G. Shafer und P.P. Shenoy. Local Computations in Hypertrees. Working Paper 201, School of Business, University of Kansas, Lawrence 1988

[32] C.E. Shannon. The Mathematical Theory of Communication. *The Bell Systems Technical Journal* 27:379–423, 1948

[33] P.P. Shenoy. Valuation-based Systems: A Framework for Managing Uncertainty in Expert Systems. Working Paper 226, School of Business, University of Kansas, Lawrence, 1991

[34] X. Zhou und T.S. Dillon. A statistical-heuristic Feature Selection Criterion for Decision Tree Induction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-13:834–841, 1991