Computergestützte Analyse onkologischer Daten mithilfe Graphischer Modelle

Bachelorarbeit

zur Erlangung des akademischen Grades

Bachelor of Science (B. Sc.)

eingereicht an der Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von Alexander Dockhorn,
geboren am 17. April 1991 in Halle (Saale)

Gutachter: Prof. Dr. Rudolf Kruse
Dr.-Ing. Matthias Steinbrecher

Inhaltsverzeichnis

Kurzfassung

Die vorliegende Arbeit befasst sich mit der Generierung von Hypothesen im Bereich der Onkologie, welche als Grundlage zukünftiger klinischer Studien verwendet werden können. Hierfür wird ein interaktives Verfahren präsentiert, welches die Patientenmenge in Untergruppen einteilt, um so die Anzahl der Attributkombinationen einer jeden Gruppe zu reduzieren. Dies soll die Anzahl generierter Hypothesen auf diejenigen beschränken, welche für den derzeitigen Untersuchungsgegenstand relevant sind. Um einen Überblick über die Patientendaten zu geben, werden Markov-Netze unterschiedlicher Patientengruppen ermittelt. Eine vergleichende Darstellung hebt Unterschiede und Gemeinsamkeiten der berechneten Modelle hervor und gibt Hinweise auf zu untersuchende Abhängigkeiten. Eine anschließende Assoziationsanalyse wird auf voneinander abhängige Patientenattribute reduziert. Anhand eines Testdatensatzes wird gezeigt, dass durch das vorgestellte Verfahren dem Datensatz zugrundeliegende Assoziationsregeln extrahiert werden können, wobei der Suchaufwand im Vergleich zu einer reinen Assoziationsanalyse erheblich reduziert werden konnte.

Abstract

The presented thesis addresses the generation of hypotheses in the area of oncology, which can be used as foundation for future clinical studies. Therefor an interactive procedure is presented, which divides the set of patients into subsets to reduce the number of attribute combinations in each subset. This should restrict the number of generated hypotheses to those relevant for the current object of investigation. Markovnetworks of diverse subsets are computed to get an overview of available patient data. Differences and commonalities of calculated models can be highlighted in a comparison view. This can provide suggestions for further investigations of certain dependencies. A subsequent association analysis will be limited to dependent attributes. The capabilities of the proposed procedure will be illustrated on the basis of a testdataset, where known underlying association rules could be extracted, while the search efforts could be significantly reduced in comparison to a standard association analysis.

In the anticipated symbiotic partnership, men will set the goals,
formulate the hypotheses, determine the criteria, and perform the evaluations.

Computing machines will do the routinizable work that must be done to
prepare the way for insights and decisions in technical and scientific thinking.

Joseph Carl Robnett Licklider

1 Einleitung und Motivation

In diesem Kapitel werden Motivation, Aufgabenstellung und Inhalt der vorliegenden Arbeit besprochen.

1.1 Motivation

Der technische Fortschritt ermöglicht es heutzutage Daten in großen Mengen automatisiert zu erfassen. Bereits jetzt werden Daten in vielen Bereichen des alltäglichen Lebens gespeichert, z. B. werden Kundendaten eines Supermarktes an der Kasse erfasst, Webseitenzugriffe täglich gespeichert und Bewertungen von Filmen auf Portalen wie der Internet Movie Database (WWW: IMDB) gesammelt.

Auch in medizinischen Bereichen wird versucht, Erfahrungen mit bisherigen Patienten für die zukünftige Behandlung neu erkrankter Personen nutzbar zu machen. Kommt es beispielsweise bei der Behandlung eines Patienten zu Komplikationen, so wird versucht mögliche Ursachen zu detektieren und anhand von klinischen Studien deren Einfluss nachzuweisen.

Allein durch die Menge an messbaren Attributen und den jeweils individuellen Patientenhistorien übersteigt der Aufwand einer manuellen oder computergestützten Auswertung oft die Möglichkeiten bisheriger Systeme. Neue Perspektiven bieten moderne Datenbanktechnologien wie *SAP HANA*, welche *In-Memory*-Technologien mit relationalen Datenbanken verbinden. Durch den damit verbundenen Performancegewinn (FÄRBER

et al., 2012) lassen sich entsprechende Datensätze in kürzerer Zeit auswerten.

Die Entwicklung des in dieser Arbeit vorgestellten Systems ist geprägt von Lickliders Vision der *Man-Computer Symbiosis* (LICKLIDER, 1992). Diese beschreibt den Versuch, Stärken und Schwächen von Mensch und Computer miteinander zu verbinden. Der Mensch entscheidet hierbei über die Bearbeitungsstrategie, wobei der Computer die notwendigen Berechnungen durchführt. Durch diese interaktive Zusammenarbeit soll eine schnelle und gesteuerte Ergebnissuche ermöglicht werden.

1.2 Aufgabenstellung und Zielsetzung

Im Rahmen dieser Bachelorarbeit soll eine Erweiterung der *SAP HANA* Anwendung *SAP Patient Data Explorer* entwickelt werden, mit deren Hilfe Attributabhängigkeiten in medizinischen Daten analysiert und Regelmengen (un-)abhängiger Attributmengen abgeleitet werden können. Durch dieses Verfahren soll die Hypothesengenerierung als Grundlage für klinische Studien vereinfacht werden.

Weiterhin sollen leicht verständliche Visualisierungen bereits ermittelte Ergebnisse darstellen und, im Sinne von Lickliders Vision der *Man-Computer-Symbiosis*, durch Interaktion mit dem Anwender nachfolgende Analyseschritte festgelegt werden.

Die Entwicklung geschah unter Betreuung des SAP Innovation Center Potsdam (ICP). Verwendete Datensätze wurden vom Nationalen Centrum für Tumorerkrankungen Heidelberg (NCT) bereitgestellt und beinhalten Daten aus Patientenakten vergangener Jahre, welche in den folgenden Kapiteln näher erläutert werden. Aus Datenschutzgründen wurden für die Entwicklung des Prototyps von SAP generierte Testdaten verwendet,welche auf Wahrscheinlichkeitsverteilungen basieren, die aus dem Originaldatensatz abgeleitet wurden. Dargestellte Informationen basieren jeweils auf diesen Testdaten und erheben keinen Anspruch auf medizinische Korrektheit.

1.3 Aufbau der Arbeit

Die nachfolgende Arbeit ist wie folgt aufgebaut: In Kapitel 2 werden nötige mathematische und algorithmische Grundlagen vermittelt. Zum Ende des Kapitels werden der *SAP Patient Data Explorer* und weitere verwandte Arbeiten vorgestellt. Kapitel 3 erläutert den zur Verfügung gestellten Datensatz und darauf aufbauende Projektvorschläge. Nachfolgend wird das in Kooperation mit dem NCT und SAP definierte Projekt näher beschrieben und dessen geplanter Lösungsansatz von bisherigen Arbeiten in diesem Gebiet abgegrenzt. Die daraufhin erstellte Implementierung und deren Evaluierung werden in den Kapitel 4 und 5 beschrieben. Abschließend werden die gewonnenen Erkenntnisse in Kapitel 6 festgehalten und mögliche zukünftige Arbeiten betrachtet.

2
Grundlagen

Die für die Arbeit nötigen fachlichen Grundlagen der Themen Graphentheorie, Graphische Modelle und Assoziationsanalyse werden in den Abschnitten 2.1 - 2.3 erläutert. Verwendete Algorithmen werden jedoch erst im späteren Verlauf der Arbeit beschrieben. Des Weiteren wird das Tool *SAP Patient Data Explorer* und sein bisheriger Funktionsumfang in Abschnitt 2.4 kurz vorgestellt. Abschnitt 2.5 schließt mit einen Überblick über verwandte Arbeiten im Bereich der Krebsdatenanalyse ab.

2.1 Graphentheorie

In der vorliegenden Arbeit werden Bayes- und Markov-Netze verwendet. Etablierte Darstellungsformen sind einfache gerichtete und ungerichtete Graphen. Aus diesem Grund sollen hierfür nötige Grundbegriffe im folgenden Abschnitt eingeführt werden. Wenn nicht weiter angegeben, wurden Definitionen dieses Abschnitts dem Buch Computational Intelligence (KRUSE et al., 2011) entnommen.

Definition 1 ((Einfacher) Graph) *Ein Graph sei definiert als ein 2-Tupel* $G = (V, E)$.[1] *Im Folgenden gilt, sei V eine endliche Menge von n Knoten definiert als*

$$V = \{A_1, ..., A_n\}$$

[1] Die Verwendung von V für die Menge der Knoten und E für die Menge der Kanten leiten sich aus dem Englischen ab. Hierbei steht V für „Vertices“ und E für „Edges“.

und E eine Menge an Kanten definiert als

$$E \subseteq (V \times V) \setminus \{ (A, A) \mid A \in V \}$$

Wir bezeichnen Graph G zudem als einfachen Graphen, da die Kantenmenge weder Mehrfachkanten noch Schleifen (Kanten mit gleichen Anfangs- und Endknoten) enthält.

Die Kanten eines einfachen Graphs können in den Ausprägungen gerichtet und ungerichtet vorkommen. Diese sind wie folgt definiert:

Definition 2 (gerichtete Kante) *Sei $G = (V, E)$ ein Graph. Wir bezeichnen eine Kante $e = (A, B) \in E$ als eine gerichtete Kante, wenn gilt:*

$$(A, B) \in E \Rightarrow (B, A) \notin E$$

Eine solche Kante sei im Folgenden repräsentiert durch $A \rightarrow B$. Knoten A wird hierbei als Elternknoten von B und Knoten B als Kindknoten von A bezeichnet.

Definition 3 (ungerichtete Kante) *Sei $G = (V, E)$ ein Graph. Eine Kante $e = (A, B) \in E$ sei eine ungerichtete Kante, wenn gilt:*

$$(A, B) \in E \Rightarrow (B, A) \in E$$

Eine ungerichtete Kante sei im Folgenden repräsentiert durch A — B oder B — A. Beide Fälle sind aufgrund der Symmetrie-Eigenschaft als gleichbedeutend anzusehen.

Abbildung 2.1 zeigt Beispielhaft einen Graphen mit gerichteten und ungerichteten Kanten. Die Struktur des Graphen ist gegeben durch:

$$V = \{Vater,\ Tochter,\ Sohn\}$$

$$E = \{(Vater,\ Sohn),\ (Vater,\ Tochter),\ (Tochter,\ Sohn),\ (Sohn, Tochter)\}$$

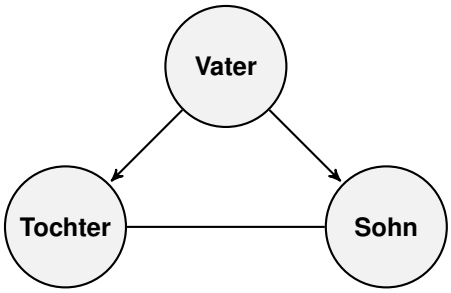


Abbildung 2.1: Simpler Graph mit gerichteten und ungerichteten Kanten

Die Knoten Vater und Sohn, sowie Vater und Tochter sind, mit je einer gerichteten Kante ausgehend vom Knoten Vater, verbunden. Der Definition nach sprechen wir bei den Knoten Tochter und Sohn von Kindknoten des Knotens Vater. Eine ungerichtete Kante verbindet die Knoten Sohn und Tochter miteinander.

Definition 4 (Adjazenzmenge) *Sei $G = (V, E)$ ein Graph. Die Adjazenzmenge eines Knotens $A \in V$ sei die Menge der Knoten die von Knoten A über eine Kante erreichbar sind.*

$$\text{adj}(A) = \{B \mid (A, B) \in E\}$$

Die Adjazenzmenge eines Knotens ist somit abhängig von der Richtung der anliegenden Kanten. Dieser Zusammenhang ist dargestellt in Abbildung 2.2.

Definition 5 (Pfad) *Sei $G = (V, E)$ ein Graph. Eine Folge ρ von r paarweise verschiedenen Knoten*

$$\rho\langle A_{i_1}, \ldots, A_{i_r}\rangle$$

heißt Pfad von A_i nach A_j, falls gilt:

- $A_{i_1} = A_i$
- $A_{i_r} = A_j$
- $(A_{i_k}, A_{i_{k+1}}) \in E \quad oder \quad (A_{i_{k+1}}, A_{i_k}) \in E, \quad 1 \leq k < r$

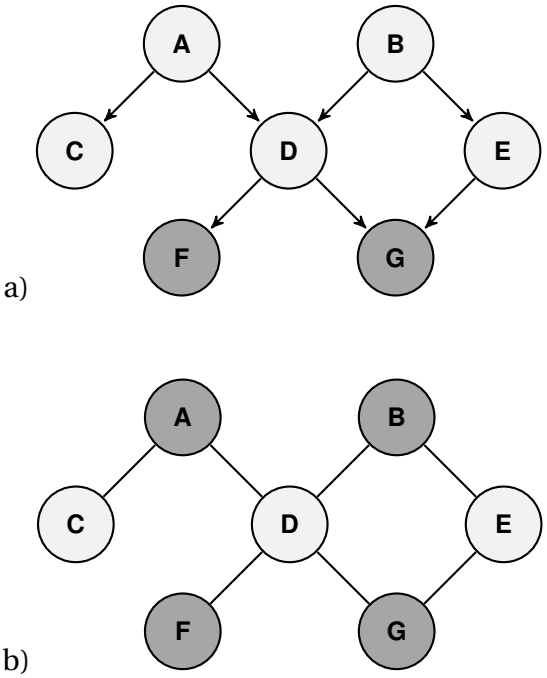


Abbildung 2.2: Vergleich der Adjazenzmengen des Knoten D in Graphen mit gerichteten (a) und ungerichteten Kanten (b). Dunkel gefärbte Knoten sind Element der Menge *adj(D)*.

Durch die Symetrieeigenschaft werden auch Pfade entgegen der Kantenrichtung erlaubt. Es werden zwei Unterkategorien von Pfaden unterschieden. Ungerichtete Pfade notieren wir mit

$$\rho = A_{i_1} - \cdots - A_{i_r}$$

Sie bestehen ausschließlich aus ungerichteten Kanten. Pfade mit ausschließlich gerichteten Kanten, welche nur in Kantenrichtung verlaufen, bezeichnen wir als gerichtete Pfade und notieren wir wie folgt

$$\rho = A_{i_1} \rightarrow \cdots \rightarrow A_{i_r}$$

Sind zwei Knoten A und B innerhalb eines Graphen G über einen gerichteten Pfad miteinander verbunden, so kennzeichnen wir dies mit der Kurznotation $A \overset{\rho}{\underset{G}{\rightsquigarrow}} B$. *Verbindet sie ein ungerichteter Pfad notieren wir* $A \overset{\rho}{\underset{G}{\nleftrightarrow}} B$.

Definition 6 (gerichteter azyklischer Graph) *Ein Graph $G = (V, E)$ wird als gerichtet bezeichnet, wenn für alle Kanten $e = (A, B) \in E$ gilt, e ist eine gerichtete Kante. Der Graph ist zudem azyklisch, wenn es keinen Pfad der Form $A \overset{\rho}{\underset{G}{\rightsquigarrow}} A, \forall A \in V$ gibt.*

Definition 7 (ungerichteter Graph) *Ein Graph $G = (V, E)$ wird als ungerichtet bezeichnet, wenn für alle Kanten $e = (A, B) \in E$ gilt, e ist eine ungerichtete Kante.*

Abbildung 2.2 zeigte bereits Beispiele für einen gerichteten azyklischen Graphen (a) und einen ungerichteten Graphen (b). In einem gerichteten Graphen definieren wir die nachfolgenden Knotenbeziehungen, welche in Abbildung 2.3 vergleichend dargestellt sind.

Definition 8 (Elternknoten) *Sei $G = (V, E)$ ein gerichteter Graph und $A \in V$ ein Knoten. Die Menge der Elternknoten von A ist definiert als:*

$$\mathrm{pa}(A) = \{B \in V \mid B \rightarrow A\}$$

Definition 9 (Kindknoten) *Sei $G = (V, E)$ ein gerichteter Graph und $A \in V$ ein Knoten. Die Menge der Kindknoten von A ist definiert als:*

$$\mathrm{ch}(A) = \{B \in V \mid A \rightarrow B\}$$

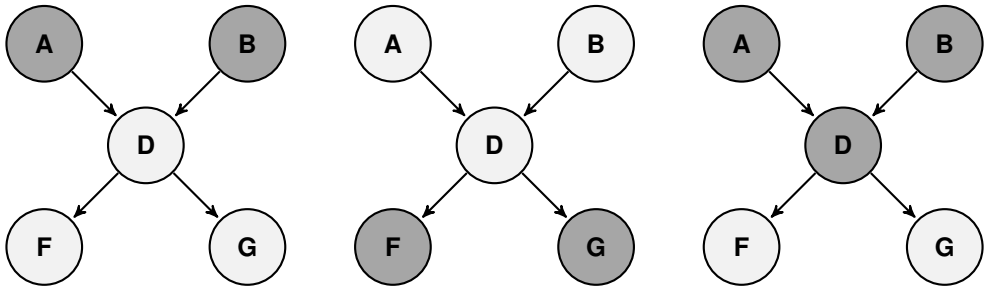
Definition 10 (Familie) *Sei $G = (V, E)$ ein gerichteter Graph und $A \in V$ ein Knoten. Die Familie des Knoten A sei definiert als:*

$$\mathrm{fa}(A) = \{A\} \cup \mathrm{pa}(A)$$

Definition 11 (Clique) *Sei $G = (V,\ E)$ ein ungerichteter Graph. Wir bezeichnen eine Knotenmenge $A \subseteq V$ als Clique, wenn für jedes Knotenpaar $A_i,\ A_j$ in A gilt:*

$$i \neq j \Rightarrow (A_i,\ A_j) \in E$$

Mit anderen Worten: ist jeder Knoten in A mit allen anderen Knoten in A verbunden, so ist A eine Clique.



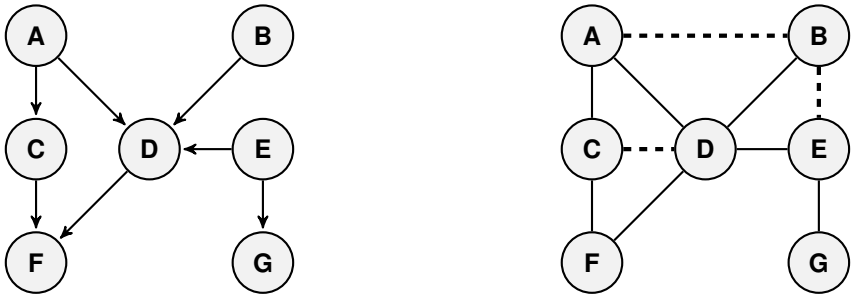
(a) Elternknoten (b) Kindknoten (c) Familie

Abbildung 2.3: Vergleich der Mengen $pa(D)$, $ch(D)$ und $fa(D)$

Die Definition von Cliquen wird später bei der Definition Graphischer Modelle (siehe Abschnitt 2.2) Anwendung finden.

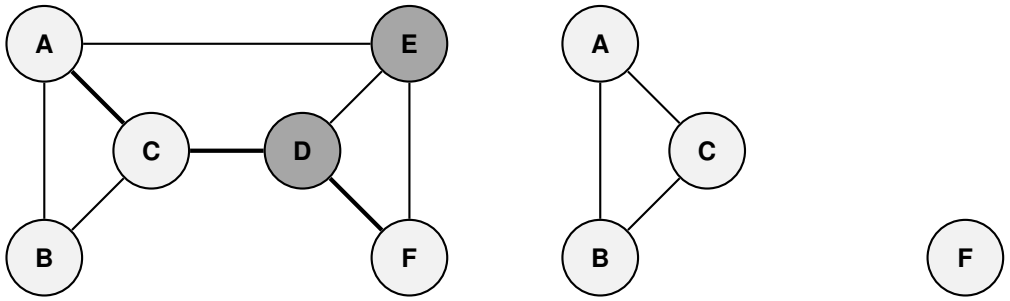
Definition 12 (Moralgraph, Moralisierung, cf. (CASTILLO, 1997))
Sei $G = (V, E)$ ein gerichteter azyklischer Graph. Wir erhalten seinen Moralgraphen G', indem wir zuerst jedes Knotenpaar mit mindestens einem gemeinsamen Kind mit einer gerichteten Kante verbinden und danach die Kantenausrichtung aller Kanten fallen lassen. Wir bezeichnen diesen Vorgang auch als Moralisierung des Graphen G

Als Beispiel ist die Moralisierung eines gerichteten azyklischen Graphens G in seinen Moralgraphen G' in Abbildung 2.4 dargestellt. Die Moralisierung wird benötigt um Bayes-Netze in Markov-Netze zu transformieren.



(a) gerichteter azyklischer Graph G (b) Moralgraph G'; hinzugefügte Kanten gestrichelt dargestellt

Abbildung 2.4: Erstellung eines Moralgraphen G'



(a) Knoten der Menge Z sind dunkelgrau hervorgehoben. Der eingezeichnete Pfad $A-C-D-F$ ist blockiert durch Knoten D aus Z.

(b) Durch entfernen der Knoten und Kanten bezüglich Z entstehen zwei voneinander getrennte Knotenmengen.

Abbildung 2.5: Grafischer Test der u-Separierbarkeit; ungerichteter Graph mit $X = \{A,\ B,\ C\}$, $Y = \{F\}$ und $Z = \{D, E\}$

Definition 13 (u-Separation) *Sei $G = (V,\ E)$ ein ungerichteter Graph und $X,\ Y,\ Z \subseteq V$ drei disjunkte Teilmengen der Knoten in G. Wir bezeichnen einen Knoten als blockierenden Knoten, wenn er in Z enthalten ist und einen Pfad als blockiert, wenn er einen Knoten aus Z enthält. Die Menge X gilt als u-separiert von Y durch Z, wenn jeder Pfad eines Knotens aus X zu einen Knoten aus Y blockiert ist. Dies schreiben wir in Kurzform als $X \perp\!\!\!\perp_G Y \mid Z$.*

Die Definition der u-Separation lässt sich leichter grafisch verdeutlichen. Entfernen wir die Knoten der Menge Z aus V und ihre dazugehörigen Kanten in E, so erhalten wir zwei getrennte Graphen, wenn gilt $X \perp\!\!\!\perp_G Y \mid Z$. Der Vorgang wird in Abbildung 2.5 dargestellt. In dem dargestellten Beispiel gilt: $\{A,\ B,\ C\} \perp\!\!\!\perp_G \{F\} \mid \{D,\ E\}$. Würden wir die zusätzliche Kante $(B,\ F)$ einfügen so wären die Mengen X und Y nicht u-separiert durch Z.

Für gerichtete Graphen werden wir die Kantenrichtung bei der Analyse der Separation zweier Knotenmengen mit einbeziehen. Entlang eines Pfades können wir Knoten weiter anhand der ein- und ausgehenden Kantenrichtungen differenzieren. Wir unterscheiden zwischen seriellen, divergierenden und konvergierenden Knoten. Tabelle 2.1 listet deren Unterscheidungen auf. Darauf basierend definieren wir die d-Separation wie folgt:

Bezeichnung	Kantenausrichtung entlang des Pfades ρ
seriell	$\cdots \leftarrow A \leftarrow \ldots$
seriell	$\cdots \rightarrow A \rightarrow \ldots$
divergierend	$\cdots \leftarrow A \rightarrow \ldots$
konvergierend	$\cdots \rightarrow A \leftarrow \ldots$

Tabelle 2.1: Knotenbezeichnungen

Definition 14 (d-Separation, cf. (FLESCH und LUCAS, 2007))
Sei $G = (V, E)$ ein gerichteter Graph und X, Y, $Z \subseteq V$ drei disjunkte Teilmengen der Knoten in G. Wir bezeichnen einen Knoten als blockierenden Knoten bezüglich eines Pfades ρ, wenn er entlang des Pfades

- *seriell oder divergierend ist und in Z liegt*
- *konvergierend ist und weder er, noch einer seiner Nachfahren in Z liegt*

Ein Pfad gilt als blockiert, wenn er einen blockierenden Knoten beinhaltet. Sind alle möglichen Pfade von X nach Y blockiert, so gilt die Menge X als d-separiert von Y durch Z. Im Folgenden verwenden wir hierfür die Kurzform $X \perp\!\!\!\perp_G Y \mid \emptyset$.

Die Anwendung der u- und d-Separation werden im nächsten Abschnitt erneut aufgegriffen.

2.2 Graphische Modelle

Die in diesem Abschnitt vorgestellten graphischen Modelle nutzen jeweils die Grundlagen des vorherigen Abschnitts und setzen Kenntnisse über bedingte Unabhängigkeiten in Wahrscheinlichkeitsverteilungen voraus. Auch dieser Abschnitt orientiert sich an dem Buch Computational Intelligence von (KRUSE et al., 2011). Die für die vorliegende Arbeit wichtigen Definitionen wurden der Buchquelle entnommen und sind hier entsprechend des Kontextes nochmals zusammengetragen.

Wir werden die Definitionen der u- und d-Separation nutzen, um bedingte Unabhängigkeit zweier Attributen innerhalb eines Graphen zu kodieren.

Definition 15 (Unabhängigkeitskarte) *Sei* $(\cdot \perp\!\!\!\perp_p \cdot \mid \cdot)$ *eine dreistellige Relation, die die bedingten Unabhängigkeiten einer gegebenen Verteilung p über der Attributmenge V repräsentiert.*

Ein ungerichteter (gerichteter) Graph $G = (V, E)$ *heißt bedingter Unabhängigkeitsgraph oder Unabhängigkeitskarte bezüglich p genau dann, wenn für alle disjunkten Teilmengen* $X, Y, Z \subset V$

$$X \perp\!\!\!\perp_G Y \mid Z \Rightarrow X \perp\!\!\!\perp_p Y \mid Z$$

gilt, d. h. wenn G durch u-Separation (d-Separation) nur solche bedingten Unabhängigkeiten beschreibt, die auch in p gelten.

Die Definition der Unabhängigkeitskarte besagt, dass sämtliche kodierten bedingten Unabhängigkeiten auch in p gelten, jedoch können weitere in p existieren. Durch die Graphoid-Axiome (siehe hierfür KRUSE et al., 2011, Seite 387) lassen sich Abhängigkeitsbeziehungen ableiten.

Definition 16 (zerlegbar bezüglich eines ungerichteten Graphen) *Die Wahrscheinlichkeitsverteilung* p_V *über einer Menge* $V = \{A_1, \ldots, A_n\}$ *von Attributen heißt zerlegbar oder faktorisierbar bezüglich eines ungerichteten Graphen* $G = (V, E)$ *genau dann, wenn sie als Produkt von nichtnegativen Funktionen auf den Cliquen von G geschrieben werden kann. Genauer: Sei* $\mathcal{C}$ *eine Familie (Menge) von Teilmengen von V, sodass sie durch die Mengen* $C \in \mathcal{C}$ *induzierten Teilgraphen die Cliquen von G sind. Sei außerdem* $\mathcal{E}_C$ *die Menge der Ereignisse, die sich durch Zuweisung von Werten an alle Attribute in C beschreiben lassen. Dann heißt* p_V *zerlegbar oder faktorisierbar bezüglich G, wenn es Funktionen* $\phi_C : \mathcal{E}_C \to \mathbb{R}_0^+, C \in \mathcal{C}$, *gibt, so dass gilt:*

$$\forall a_1 \in \mathrm{dom}(A_1) : \cdots \forall a_n \in \mathrm{dom}(A_n) :$$
$$p_V\left(\bigwedge_{A_i \in V} A_i = a_i\right) = \prod_{C \in \mathcal{C}} \phi_C\left(\bigwedge_{A_i \in C} A_i = a_i\right)$$

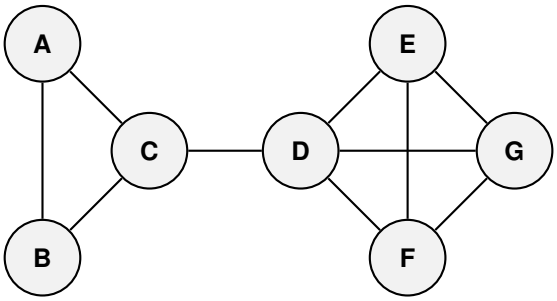


Abbildung 2.6: Beispiel eines Markov-Netzes

Die Definition der Zerlegbarkeit bezüglich eines ungerichteten Graphen sei anhand des Graphens aus Abbildung 2.6 verdeutlicht. Dieser beinhaltet die Cliquen:

$$C_1 = \{A, B, C\}, \quad C_2 = \{C, D\}, \quad C_3 = \{D, E, F, G\}$$

Die durch den Graphen induzierte Zerlegung lautet:

$$\begin{aligned} \forall a \in \mathrm{dom}(A) : \cdots \forall g \in \mathrm{dom}(G) : \\ p_V\big(A = a, \ldots, G = g\big) &= \phi_{C_1}(A = a, B = b, C = c) \\ &\quad \cdot \phi_{C_2}(C = c, D = d) \\ &\quad \cdot \phi_{C_3}(D = d, E = e, F = f, G = g) \end{aligned}$$

Abschließend verbinden wir die bisherigen Definitionen und geben eine Definition für Bayes- und Markov-Netze an.

Definition 17 (Bayes-Netz) *Ein Bayes-Netz ist ein gerichteter bedingter Unabhängigkeitsgraph einer Wahrscheinlichkeitsverteilung p_V zusammen mit einer Familie von bedingten Wahrscheinlichkeiten der durch den Graphen induzierten Faktorisierung.*

Da Bayes-Netze in dieser Arbeit lediglich als Zwischenschritt dienen, wird an dieser Stelle auf deren Funktionsweise nicht weiter eingegangen. Weitere Informationen sind in dem Buch (KRUSE et al., 2011) zu finden

Definition 18 (Markov-Netz) *Ein Markov-Netz ist ein ungerichteter bedingter Unabhängigkeitsgraph $G = (V, E)$ einer Wahrscheinlichkeitsverteilung p_V zusammen mit einer Familie von nichtnegativen Funktionen ϕ_M der durch den Graphen induzierten Faktorisierung.*

In dieser Arbeit werden Markov-Netze genutzt, um die Abhängigkeiten des Datensatzes in leicht lesbarer Form zu repräsentieren. Abschnitt 4.3 behandelt die Berechnung und Darstellung von Markov-Netzen innerhalb des entwickelten Programms.

2.3 Assoziationsanalyse

Im Folgenden werden Grundlagen der Assoziationsanalyse erläutert, welche das Finden von Mustern und deren „interessanten“ Beziehungen in zumeist großen Datensätzen beschreibt. Ein typisches Anwendungsgebiet ist die Warenkorbanalyse an deren Beispiel die folgenden Begriffe erläutert werden. Ziel dieser ist es, häufige Produktkombinationen zu ermitteln, um mittels gezielter Produktplatzierung die Gewinne zu maximieren.

2.3.1 Frequent Item Sets

Als Grundlage der Assoziationsanalyse müssen zuerst häufige Muster extrahiert werden. Diese werden wir im Folgenden als Frequent Item Sets bezeichnen. Hier aufgelistete Definitionen entstammen der ursprünglichen Definition in AGRAWAL et al. (1993). Abweichungen werden entsprechend erläutert. Die folgenden Begriffe werden jeweils mit ihren Äquivalent in einer Warenkorbanalyse verdeutlicht.

Definition 19 (Itembasis) *Sei $B = \{i_1, ..., i_m\}$ eine Menge bestehend aus Items. Im folgenden bezeichnen wir die Menge B als Itembasis.*

Definition 20 (Item Set) *Jedes Subset $I \subseteq B$ nennen wir ein Item Set.*

Im Gegensatz zur Problemdefinition in AGRAWAL et al. (1993) sei an dieser Stelle die Einschränkung von Items als ausschließlich binäre Attribute aufgehoben, da eine Umformung von mehrwertigen Attributen in eine Menge binärer Attribute, wie in Abbildung 2.7 dargestellt, leicht umgesetzt werden kann.

(a)

Attributname	mögliche Attributwerte	Beispiel Belegung
Obst	{Apfel, Birne, Banane, Kiwi}	{Apfel, Birne}

(b)

Attributname	mögliche Attributwerte	Beispiel Belegung
Apfel	{True, False}	{True}
Birne	{True, False}	{True}
Banane	{True, False}	{False}
Kiwi	{True, False}	{False}

Abbildung 2.7: Umformung eines mehrwertiger Attributes (a) in mehrere binäre (b) Attribute

Am Beispiel der Warenkorbanalyse beschreiben die unterschiedlichen, zur Verfügung stehenden Waren einzelne Items. Die Itembasis setzt sich somit aus dem gesamten Produktspektrum des Warenhauses zusammen.

Definition 21 (Transaktion) *Eine Transaktion sei präsentiert durch ein Tupel t bestehend aus einer eindeutigen Transaktions-ID und einer Menge an Items* $t = (t_{id}, \{I \subseteq B\})$.

In der Warenkorbanalyse stellt eine Transaktion den Einkauf von Waren eines einzelnen Kunden dar. Jeder abgeschlossene Einkauf kann somit in einer Datenbank zusammengefasst werden.

Definition 22 (Transaktionsdatenbank) *Sei* $T = (t_1, ..., t_n)$ *mit* $t_k \subseteq B$, $1 \leq k \leq n$ *ein Vektor aus Transaktionen* t_k *über B. Dieser Vektor wird Transaktionsdatenbank (über B) genannt. Es reicht hierbei auch die Menge der Items einer Transaktion* t_k *in der Datenbank zu erfassen, da der Vektorindex einer Transaktion bereits zur eindeutigen Identifikation dieser verwendet werden kann.*

Tabelle 2.2 listet Transaktionen einer Warenhaus-Transaktionsdatenbank in Tabellenform auf. Diese entsprechen den Käufen einzelner Kunden. In diesem Beispiel sei das Warenspektrum auf Orangensaft, Milch, Eier und Toast beschränkt. Die Auflistung aller in der Transaktionsdatenbank vorkommenden Waren bildet implizit die Itembasis B.

Käufer	gekaufte Waren
1	{Orangensaft, Milch}
2	{Eier, Toast, Milch}
3	{Orangensaft, Toast}
4	{Orangensaft, Milch}
5	{Orangensaft, Marmelade}
6	{Orangensaft, Toast}
7	{Eier, Toast}
8	{Orangensaft, Toast, Milch}
9	{Eier, Toast}
10	{Orangensaft, Milch}

Tabelle 2.2: Beispiel einer Transaktionsdatenbank mit Käuferdaten

Definition 23 ((relativer/absoluter) Support) *Sei $I \subseteq B$ ein Item Set und $t \in T$ eine Transaktion. Die Transaktion t beinhaltet I, wenn gilt $i \in t, \forall i \in I$.*

Wir bezeichnen den absoluten Support eines Item Sets in Bezug auf eine Transaktionsbasis T als die Anzahl der Transaktionen, die Item Set I beinhalten. In dieser Arbeit wird Support gleichbedeutend mit absoluten Support verwendet.

$$s(I)_T = |\{t \mid I \subseteq t,\ t \in T\}|$$

Der relative Support von I in Bezug auf eine Transaktionsbasis T sei definiert als der Anteil der I beinhaltenden Transaktionen $t \in T$.

$$\sigma(I)_T = \frac{1}{|T|}|\{t \mid I \subseteq t,\ t \in T\}|$$

Im Folgenden wird der Index T ausgelassen, wenn die zugrundeliegende Transaktionsdatenbank aus dem Kontext hervorgeht.

Aus der Definition des Supports geht hervor, dass der Support der leeren Menge stets die Anzahl der Transaktionen in der Transaktionsdatenbank entspricht beziehungsweise deren relativer Supportwert 1 ist.

Zur Verdeutlichung der Definition ermitteln wir den absoluten und relativen Supportwert des Item Sets {Eier, Toast} unter Verwendung der Transaktionsdatenbank aus Tabelle 2.2:

$$s(\{\text{Eier, Toast}\})_T = |\{t \mid \{\text{Eier, Toast}\} \subseteq t,\ t \in T\}| = |\{t_2, t_7, t_9\}| = 3$$

$$\sigma(\{\text{Eier, Toast}\})_T = \frac{1}{|T|}|\{t \mid \{\text{Eier, Toast}\} \subseteq t,\ t \in T\}| = \frac{1}{10}|\{t_2, t_7, t_9\}| = 0.3$$

In unserem Beispiel beschreibt der Support wie oft eine Produktkombination gekauft wurde. Dies kann bereits verwendet werden, um deren zukünftigen Umsatz abzuschätzen. Da wir an Abhängigkeiten von Waren interessiert sind, fokussieren wir die weitere Analyse auf ausschließlich häufige Produktkombinationen, welche wie folgt definiert sind:

Definition 24 (Frequent Item Set) *Gegeben sei eine Transaktionsdatenbank $T = \{t_1, ..., t_n\}$ über einer Itembasis $B = \{i_1, ..., i_m\}$ und ein minimaler Support $s_{\min} \in \mathbb{N},\ 0 < s_{\min} \leq n$.*

Wir bezeichnen ein Item Set I als Frequent Item Set [2], wenn gilt:

$$s(I)_T \geq s_{\min}$$

Die Menge der Frequent Item Sets bezeichnen wir als

$$F_T(s_{\min}) = \{I \subseteq B \mid s_T(I) \geq s_{\min}\}$$

Es ist weiterhin möglich den minimalen Support als relativen Support zu definieren. Folgende Umformungen müssen hierfür beachtet werden.

minimaler Support:

$$\sigma_{\min} \in \mathbb{R},\ 0 < \sigma_{\min} < 1$$

Frequent Item Set:

$$\sigma_T(I) \geq \sigma_{\min}$$

Menge der Frequent Item Sets:

$$\Phi_T(\sigma_{\min}) = \{I \subseteq B \mid \sigma_T(I) \geq \sigma_{\min}\}$$

Es sei darauf hingewiesen, dass Angaben eines minimalen Supports oder eines minimalen relativen Supports durch

$$s_{\min} = \lceil n\sigma_{\min} \rceil \quad \textit{und} \quad \sigma_{\min} = \frac{1}{n} s_{\min}$$

frei untereinander transformiert werden können.

0 Items	1 Item		2 Items	
∅: 10	{Orangensaft}	: 7	{Orangensaft, Toast}	: 3
	{Eier}	: 3	{Orangensaft, Milch}	: 4
	{Toast}	: 6	{Eier, Toast}	: 3
	{Milch}	: 5		

Tabelle 2.3: Frequent Item Sets ($s_{\min} = 3$, $\sigma_{\min} = 0.3 = 30\%$) der Transaktionsdatenbank 2.2

Tabelle 2.3 listet alle Frequent Item Sets der Transaktionsdatenbank aus Tabelle 2.2 auf, welche einen minimalen Support von $s_{\min} = 3$ erfüllen.

2.3.2 Induktion von Assoziationsregeln

Für die ermittelten Frequent Item Sets wollen wir nun Assoziationsregeln aufstellen. Diese seien wie folgt definiert:

Definition 25 (Assoziationsregel, cf. (AGRAWAL et al., 1993))
Eine Assoziationsregel r sei beschrieben durch zwei Item Sets $X, Y \subset B$ für die gilt: $X \cap Y = \emptyset$ und $X \neq \emptyset$, $Y \neq \emptyset$. Wir notieren $X \Rightarrow Y$. Mit dieser Regel bezeichnen wir den Zusammenhang „Wenn X gilt, dann gilt auch Y". Mit X bezeichnen wir die Antezedenz der Assoziationsregel und Y nennen wir die Konsequenz der Regel.

Im Gegensatz zur Originaldefinition in (AGRAWAL et al., 1993) seien Antezedenz und Konsequenz hier als nichtleere Mengen festgelegt. Wir werden uns im Verlaufe der Arbeit auf Assoziationsregeln mit einelementiger Konsequenz beschränken, da Hauptziel der Analyse das Finden von hypothetischen Ursachen eines spezifischen Symptoms sein wird.

Eine Assoziationsregel kann beispielsweise Rückschlüsse über das Kaufverhalten von Kunden zum Ausdruck bringen. Die Regel „Wenn ein Kunde Eier kauft, so kauft er auch Toast" ($Eier \Rightarrow Toast$) sei hierfür ein Beispiel. Um die Relevanz einer Regel zu messen, werden in der vorliegenden Arbeit die Maße Support, Konfidenz und Lift einer Regel verwendet.

[2] In AGRAWAL et al. (1993) noch als *large Item Set* bezeichnet

Definition 26 (Support (einer Assoziationsregel), cf. GÖRZ (2000))
Sei r eine Assoziationsregel der Form $X \Rightarrow Y$. Der absolute Support der Regel r sei definiert als

$$s(r) = |\{t \in T \mid X \cup Y \in t\}| = s(X \cup Y)$$

Entsprechend der Definition des relativen Supports von Item Sets gilt auch für Assoziationsregeln:

$$\sigma(r) = \frac{|\{t \in T \mid X \cup Y \subseteq t\}|}{|T|} = \sigma(X \cup Y)$$

Definition 27 (Konfidenz, cf. GÖRZ (2000)) *Sei r eine Assoziationsregel der Form $X \Rightarrow Y$. Die Konfidenz der Regel r sei definiert als*

$$c(r) = \frac{|\{t \in T \mid X \cup Y \subseteq t\}|}{|\{t \in T \mid X \subseteq t\}|} = \frac{s(X \cup Y)}{s(X)}$$

Definition 28 (Lift, cf. (MCNICHOLAS et al., 2008)) *Sei r eine Assoziationsregel der Form $X \Rightarrow Y$. Der Lift einer Regel r sei definiert als*

$$\text{lift}(r) = \frac{|T|^2}{|T|} \frac{|\{t \in T \mid X \cup Y \in t\}|}{|\{t \in T \mid X \in t\}| \cdot |\{t \in T \mid Y \in t\}|} = \frac{\sigma(X \cup Y)}{\sigma(X) \cdot \sigma(Y)} = \frac{c(r)}{\sigma(Y)}$$

Der Support einer Regel ($X \Rightarrow Y$) beschreibt die relative Häufigkeit der in der Transaktionsdatenbank vorhanden Fälle, in denen die Regel anwendbar ist. Die relative Häufigkeit der Fälle in denen die Regel richtig ist, wird durch die Konfidenz angegeben.

Die Menge an auszugebenden Regeln wird oft beschränkt auf Regeln mit einer hohen Anwendbarkeit, welche häufig zum richtigen Ergebnis führen. Durch die vorhergehende Suche nach Frequent Item Sets ist die Anwendbarkeit (Mindestsupport) bereits sichergestellt. Wir können im folgenden für jedes Frequent Item Set I alle Regeln der Form $X \Rightarrow Y$ aufstellen, für die gilt:

$$X \cup Y = I, \quad X \cap Y = \emptyset, \quad X \neq \emptyset \quad \text{und} \quad Y \neq \emptyset$$

Aus der in Tabelle 2.3 bereits als Frequent Item Set dargestellten Menge {Orangensaft, Milch} lassen sich die in Tabelle 2.4 abgeleiteten Regeln mit

Regel	Konfidenz	Lift
{Orangensaft} ⇒ {Milch}	$4/7 \approx 0.57$	$0.4/0.7 \cdot 0.5 = 1.14$
{Milch} ⇒ {Orangensaft}	$4/5 = 0.80$	$0.4/0.5 \cdot 0.7 = 1.14$

Tabelle 2.4: aufstellbare Assoziationsregeln des Frequent Item Sets {Orangensaft, Milch} und deren Konfidenzwerte

ihren zugehörigen Konfidenzwerten berechnen. Daraus lässt sich ableiten, dass 80 % der Kunden, die Milch gekauft haben, auch Orangensaft kauften, wohingegen nur rund 57 % der Orangensaftkäufer auch zu Milch griffen.

Der Lift wird in (MCNICHOLAS et al., 2008) beschrieben als ein symmetrisches Maß, welches angibt, wie sehr die Mengen X und Y nicht unabhängig voneinander sind. Das heißt, zu welchen Ausmaß die Gleichung

$$P(X,Y) = P(X) \cdot P(Y)$$

nicht wahr ist. Im Gegensatz zu Support und Konfidenz einer Assoziationsregel ist der Wertebereich des Lift nicht beschränkt auf $0 \leq x \leq 1$. Ein Wert unter (über) 1 deutet darauf hin, dass Antezedenz und Konsequenz der Regel gemeinsam seltener (öfter) auftreten als erwartet. Der Lift kann wie auch der Support und die Konfidenz einer Regel genutzt werden, um das Interesse an dieser zu beschreiben.

Die Generierung der Assoziationsregeln wird in Abschnitt 4.5 erläutert.

2.4 SAP Patient Data Explorer

Die in der vorliegenden Arbeit entwickelten Module gliedern sich in das von SAP seit Anfang 2013 in Entwicklung befindliche Webapplication *SAP Patient Data Explorer*[3] ein (WWW: SAP). In Zusammenarbeit mit dem Nationalen Centrum für Tumorerkrankungen Heidelberg werden neue

[3] Hier beschriebene Funktionen und Abbildungen spiegeln den Projektstand des 10.01.2014 wieder. Das Tool befand sich zum Zeitpunkt der Abgabe dieser Arbeit weiterhin in kontinuierlicher Entwicklung.

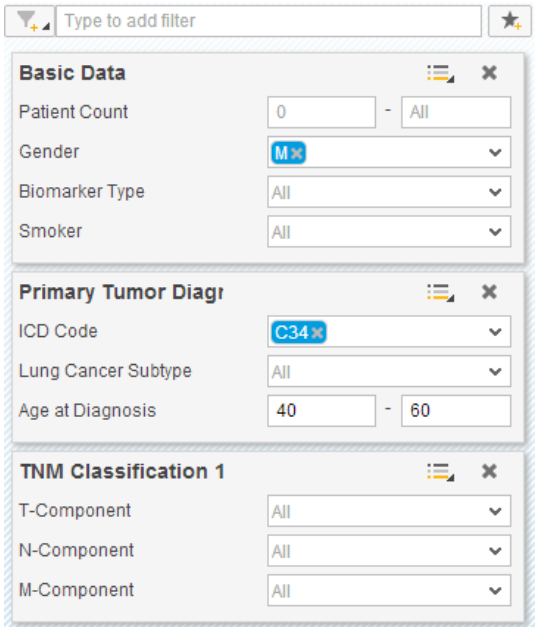


Abbildung 2.8: Darstellung der Filtereinstellungen (männlich, Lungenkrebsdiagnose, Alter 40-60); weitere Filterkarten über obere Eingabeleiste hinzufügbar

Wege gesucht, um die enorme Menge an Patientendaten zeitnah zu verarbeiten.

Beispielsweise müssen in Vorbereitung einer klinischen Studie geeignete Patienten gefunden werden, um die Wirksamkeit einer neuen Therapieform zu testen. Dies umfasst die Analyse sämtlicher verfügbarer Informationen, um mögliche Gefahren oder Störeinflüsse zu vermeiden. Hierfür wurde ein einheitliches Datenmodell entwickelt, welches es ermöglicht, Informationen aus unterschiedlichen Quellen zu aggregieren. Bisher integriert wurden Daten aus Tumordatenbanken, Biobanken, Arztbriefen und *IS-H / i.s.h.med*[4].

Das Programm ermöglicht es, die Gesamtheit der Patienten nach Attributen zu filtern und diese auszugeben. Abbildung 2.8 zeigt eine mögliche Filtereinstellung, in welcher die Grundgesamtheit der Patienten auf ausschließlich Männer im Alter von 40-60 Jahren mit vorangegangener Lungenkrebsdiagnose eingeschränkt ist.

[4] Produktbezeichnungen von Krankenhausinformationssystemen

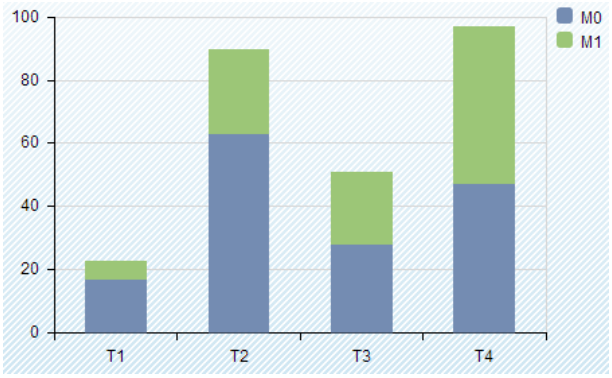


Abbildung 2.9: Attributwertdarstellung als Balkendiagramm

Die Verteilungen der Attribute dieser gefilterten Menge können in unterschiedlichen Darstellungen angezeigt und verglichen werden. Abbildung 2.9 zeigt die Verteilung der Metastasierung aufgeteilt nach gemessener Tumorengröße des in Abbildung 2.8 dargestellten Filters.

2.5 Verwandte Arbeiten

Zum Abschluss des Kapitels sollen vorangegangene Arbeiten im Bereich der Krebsdatenanalyse vorgestellt werden. Die hier aufgelisteten Arbeiten gaben Hinweise über Möglichkeiten der Analyse medizinischer Daten und beeinflussten die Projektdefinitionsphase.

Data-Mining-Techniken fanden bereits Anwendung in der Diagnose und Prognose von Krebserkrankungen. Die Arbeiten von (ZUBI und SAAD, 2011), (AHMEDMEDJAHED et al., 2013) und (AHMED et al., 2013) zeigten den erfolgreichen Einsatz von Cluster- und Klassifikationsalgorithmen zur Früherkennung von Lungen-, Brust- und Hautkrebs.

Im Bereich der Prognose zeigten die Arbeiten von (BURKE et al., 1997) und (RAVDIN und CLARK, 1992) wie Neuronale Netzwerke genutzt werden können, um die Überlebensrate von Brustkrebspatienten zu schätzen. Ähnliche Ergebnisse erzielte die Arbeit von (SARVESTANI et al., 2010), welche zeigte, wie Naïve-Bayes-Classifier, Neuronale Netze und Entscheidungs-

bäume zur Abschätzung der verbleibenden Lebenszeit eines Patienten genutzt werden können.

Einen anderen Ansatz verfolgte die Arbeit von (AGRAWAL und CHOUDHARY, 2011). In dieser wurden Daten von Lungenkrebspatienten der SEER Datenbank mittels einer Assoziationsanalyse untersucht. Hierbei wurden Regeln bezüglich der Überlebensdauer extrahiert, um Aussagen über Lebens verlängernde/verkürzende Patienteneigenschaften zu treffen. Als Ergebnis konnten bereits bekannte und neue Aussagen vorgewiesen werden, welche Anreize für weitere Untersuchungen gaben.

Die Schätzung der Überlebenszeit wurde in den bisher vorgestellten Arbeiten immer abhängig vom derzeitigen Zustand des Patienten ermittelt. Es gibt jedoch auch Ansätze den Erfolg von noch kommenden Therapien abzuschätzen, um so die erfolgversprechendste Behandlung zu wählen. So wurde in der Studie von (YADAV et al., 2013) versucht, das Ansprechen der Patienten auf eine Chemotherapie mittels Clustering der Patientengruppen abzuschätzen.

Chemotherapien sind auch Thema von anderen Analysen. So wird beispielsweise versucht, die Zusammenstellung eines Chemotherapieprotokolls zu optimieren. Dies ist durch die Anzahl der Freiheitsgrade ein aufwändiger Prozess. Die Arbeiten von (TAN et al., 2002) und (PETROVSKI und MCCALL, 2001) zeigten, dass Evolutionäre Algorithmen verwendet werden können, um die Intervalle der zu verabreichenden Medikamente und deren Dosierungen zu ermitteln.

Zur Beantwortung von allgemeinen medizinischen Fragestellungen wurde das von IBM entwickelte Programm *Watson* (FERRUCCI et al., 2010) angepasst. Ärzte können dem System Fragestellungen in natürlicher Sprache stellen. Darauf aufbauend werden Hypothesen formuliert, in deren Bewertung Daten aus Patientenakten, klinischen Studien, Behandlungsrichtlinien und weiteren Quellen mit einbezogen werden. Positiv bewertete Hypothesen und deren Sicherheit werden dem Arzt ausgegeben und können die Diagnose und die Wahl einer geeigneten Therapieform unterstützen. Eine Weiterentwicklung im Bereich der Onkologie wurde in (WWW: WATSON ONCOLOGY) offiziell bestätigt.

Die vorgestellten Arbeiten decken bereits weite Felder der Krebsdatenanalyse ab. In Abschnitt 3.2 werden eigene Ansätze vorgestellt, welche anhand der vorgegebenen Daten denkbar wären. Das im Rahmen dieser Arbeit erstellte Programm wird in Kapitel 4 vorgestellt. Die durchgeführte Evaluation (siehe Kapitel 5) zeigt Unterschiede zu alternativen Lösungsansätzen auf und hebt die Leistungsfähigkeit des Systems hervor.

3

Datenanalyse und Projektdefinition

Zu Beginn der Kooperation des NCT und SAP wurden Wahrscheinlichkeitsverteilungen der Attribute auf Basis vollständig anonymisierten Patientendaten abgeleitet. Für die Entwicklung eines *Proof of Concepts* des *SAP Patient Data Explorers* wurden synthetische Patientendaten erstellt, welche den von SAP aggregierten Originaldatensatz im Aufbau gleichen, jedoch zufällig aus abgeleiteten Attributsverteilungen generiert wurden.

Attribute des Datensatzes werden in Abschnitt 3.1 erläutert. Abschnitt 3.2 beschreibt mögliche darauf aufbauende Projekte, über deren Umsetzung in gemeinsamer Abstimmung mit dem NCT entschieden wurde. Ein Entwurf des Systems wird in Abschnitt 3.3 näher beschrieben und von anderen Verfahren abgegrenzt.

3.1 Datenanalyse

Attribute der von SAP generierten Datenbank basieren jeweils auf abgeleiteten Attributsverteilungen des anonymisierten Originaldatensatzes. Für die Generierung von Patienten der mit am häufigsten vorkommenden Krebsarten Lungen-, Brust- und Darmkrebs wurden deren bedingten Wahrscheinlichkeitsverteilungen verwendet. Weitere Krebsarten wurden mittels allgemein ermittelten Attributsverteilungen erstellt.

Vitalstatus Das Attribut Vitalstatus gibt den Zustand des Patienten zum zuletzt abgefragten Datum an. Es wird zwischen „Lebt“, „Verstor-

ben“ und „Unbekannt“ unterschieden. Diese Daten werden durch regelmäßige Anfragen bei früheren Patienten gesammelt.

Geburtsdatum Die Geburtsdaten wurden zufällig entsprechend der aus dem Originaldaten abgeleiteten Verteilung generiert. Das Ergebnis wurde im Format „Tag.Monat.Jahr“ abgespeichert.

Geschlecht Gibt das Geschlecht des Patienten (M/W) an.

Weitere Daten beziehen sich jeweils auf eine Erstdiagnose. Erkrankt ein Patient erneut an der selben Tumorart oder einer anderen, so werden diese Informationen gesondert von vorher notierten Fällen aufgenommen.

ICD-Klassifikation Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-Klassifikation). Die in der Datenbank vorhandenen Klassifikationscodes C00-C98 beschreiben die Lokalisation bösartiger Neubildungen (maligne Tumoren). Der eingetragene Wert entspricht jeweils einer Erstdiagnose, zu welcher ebenfalls das Datum angegeben ist.

TNM-Klassifikation Die TNM-Klassifikation beruht auf statistischen Aussagen über Größe und Art der Tumorausprägung. Der Zustand der Tumorerkrankung wird nach einem dreistufigen System bewertet und kann zur Krankheitsprognose verwendet werden. [1]

Tumor Ausdehnung des Primärtumors, Einstufung 0 (Primärtumor unbekannt oder bereits durch Vorbehandlung bekämpft), 1 - 4 (aufsteigende Größe und Ausbreitung des Tumors)

Nodes (Lymphknoten) Vorhandensein von regionären Lymphknotenmetastasen, Einstufung von 0 (keine) bis 3 (höchste Befallsstufe)

Metastasen Vorhandensein von hämatogenen Fernmetastasen, Einstufung von 0 (keine) bis 1 (vorhanden)

[1] (SOBIN et al., 2011) beinhaltet aktuelle Beschreibungen der TNM-Klassifikation

Biomarker Eine Eigenschaft, welche objektiv messbar und evaluierbar als ein Indikator für biologische Prozesse, pathogene Prozesse oder eine pharmakologische Reaktion auf eine therapeutische Intervention dient[2]. Die Integration von Biomarkern in der Datenbank ist nicht abgeschlossen und derzeit auf 200 Einträge beschränkt.

Behandlungshistorie Der Start- und Endzeitpunkt der Behandlung wird jeweils im Format „Tag.Monat.Jahr“ abgespeichert.

Therapieinformationen Zur Behandlung von Tumoren bieten sich je nach Art und Stadium des Tumors unterschiedliche Behandlungsformen an. Durchgeführte Therapien werden je Patient abgespeichert. Es ist möglich, dass mehrere Therapien (der gleichen Form) durchgeführt wurden.

Chemotherapien Informationen zu durchgeführten Chemotherapien; beinhaltet pro Eintrag Start- und Enddatum der Therapie sowie das eingesetzte Chemoprotokoll

Operative Therapie Datum der durchgeführten Operation

Strahlentherapien Datum der durchgeführten Strahlentherapie

Der Datensatz wurde noch um abgeleitete Attribute erweitert. Ursprung und Berechnung des jeweiligen Attributes werden nachfolgend erläutert.

Alter zum Zeitpunkt der Erstdiagnose Aus dem angegebenem Datum der Erstdiagnose und dem Geburtsdatum des betroffenen Patienten wird das Alter zum Zeitpunkt der Erstdiagnose ermittelt.

Zeitpunkt des Todes Ist der Vitalstatus eines Patienten mit „Verstorben“ angegeben, so kann das Alter zum Zeitpunkt des Todes nur durch das Datum der letzten Abfrage und dem vorhandenen Geburtsdatum geschätzt werden. Dieser Zeitpunkt dient jedoch nur als Schätzung, da präzise Aussagen zum Zeitpunkt des Todes nicht bei der Ermittlung des Vitalstatus angegeben werden müssen.

[2] sinngemäß übersetzt aus (ATKINSON et al., 2001)

Alter zum Zeitpunkt des Todes Das Alter zum Zeitpunkt des Todes wird geschätzt durch das in der Datenbank vorhandene Geburtsdatum und dem abgeleiteten Zeitpunkt des Todes. Diese Angabe wird auf Jahre gerundet, da eine höhere Genauigkeit durch den Abfragerhythmus des Vitalstatus nicht legitimiert werden kann.

Sämtliche Attribute wurden entsprechend des zu erwartenden Wertebereichs gefiltert und deren Verteilung im Gespräch mit Ärzten des NCT einer Plausibilitätsüberprüfung unterzogen. Auf diese Weise konnten einige Datenfehler gefunden und behoben werden.

3.2 Projektdefinition

Nach erster Sichtung der Daten wurden in Gesprächen mit Fachärzten des NCT die verfügbaren Daten und mögliche darauf aufbauende Analysesysteme besprochen. Folgende Systemvorschläge wurden zur Diskussion gestellt:

Überprüfen von Behandlungsrichtlinien Die Behandlung der Patienten folgt teils strengen, genormten Richtlinien. Diese können als Entscheidungsbäume dargestellt werden, mit deren Hilfe beispielsweise die anzuwendende Therapie bestimmt werden kann. Einflussfaktoren hierfür sind unter anderem die TNM-Klassifikation, das Alter und die Lokalisation des Tumors. Anhand der gegebenen Datenbank ließe sich überprüfen, ob die verbindlichen Behandlungsrichtlinien eingehalten werden.

Abschätzung der zu erwartenden Überlebenszeit Bei der Behandlung von Krebs werden betroffene Patienten teils vor schwere Entscheidungen gestellt. Sind mehrere Therapieformen denkbar, muss abgeschätzt werden, ob das mit diesen Formen verbundene Risiko tragbar ist. Beispielsweise kann eine Chemotherapie mit einer enormen körperlichen Belastung einhergehen, aber die zu erwartende Lebenszeit bei Behandlungserfolg beträchtlich steigern. Durch

Clusteranalysen, Kaplan-Meier- und Cox-Modelle könnte der Einfluss der möglichen Behandlungen auf die zu erwartende Überlebenszeit abgeschätzt werden, um den Arzt in der Beratung des Patienten zu unterstützen.

Abhängigkeitsanalyse der Attribute Gibt es mögliche Risiko- oder Hinweisfaktoren für Krebserkrankungen, so wird deren Aussagekraft in klinischen Studien untersucht. Die verfügbare Datenbank kann bereits Hinweise auf solche Attributabhängigkeiten liefern. Diese könnten durch das System ermittelt werden und als Grundlage für die Gestaltung von zukünftigen Studien dienen. Durch statistische Modelle und Assoziationsanalysen könnten Attributabhängigkeiten des Datensatzes extrahiert und bewertet werden.

Nach Vorstellung und Besprechung der anzuwendenden Methoden, konnte das überprüfen der Behandlungsrichtlinien durch Mitarbeiter des NCT ausgeschlossen werden, da ein wichtiger Indikator in Behandlungsrichtlinien das Befinden des Patienten ist und dieses nicht in der Datenbank erfasst wird. Ein direkter Vergleich zu bisherigen Richtlinien ist daher nicht möglich.

Eine Integration von Kaplan-Meier und Cox-Modellen wurde bereits von SAP vorbereitet und wird in einer kommenden Version des *SAP Patient Data Explorers* hinzugefügt. Eine Erweiterung durch eine Clusteranalyse konnte durch einen Prototypen ausgeschlossen werden, da die Performance in der Schätzung der Überlebenszeit weit unter der Leistung der statistischen Modelle lag.

In Abstimmung mit dem NCT wurde entschieden, dass die Umsetzbarkeit von Vorschlag drei mittels eines Prototypens abgeklärt werden soll. Hierbei wurde von beiden Seiten festgelegt, dass das zu entwickelnde System die Hypothesengenerierung und -überprüfung innerhalb der Patientendaten des NCTs unterstützen soll. Patientengruppen für eine entsprechende Studie können dann mit Hilfe der anderen Module des *SAP Patient Data Explorers* exportiert wurden.

Folgendes Szenario wäre für das zu entwickelnde System denkbar:

> Ein Arzt nimmt Auffälligkeiten bei der Behandlung von Patienten mit einem Biomarker wahr. Im Laufe der Behandlung kam es zu gesteigerten Vorkommen von Übelkeit und Erbrechen.

Zu beantwortende Frage:

> Existieren bisher nicht näher untersuchte statistische Abhängigkeiten zwischen dem Biomarker, den wahrgenommenen Nebenwirkungen und weiteren Attributen der Patienten?

Den Mitarbeitern des NCT ist es hierbei wichtig die zu untersuchende Grundgesamtheit frei nach den Attributen der Patienten und der durchgeführten Behandlungen in eine zu untersuchende Teilmenge einzuschränken. Statistische Abhängigkeiten der zugrunde liegenden Teilmenge sollen automatisiert ermittelt und in einer graphischen Darstellung ausgegeben werden

3.3 Planung des Systemaufbaus

Die Generierung von Hypothesen soll durch eine Assoziationsanalyse realisiert werden. Durch die Vielfalt an vorhandenen Merkmalskombinationen ist jedoch ein Übermaß an Assoziationsregeln zu erwarten (siehe Evaluation in Kapitel 5). Diese Menge muss bestmöglich nach den Bedürfnissen des Benutzers eingeschränkt werden. Gespräche mit zukünftigen Anwendern haben gezeigt, dass oft nur Teilmengen zu untersuchen sind. So ist beispielsweise ein Spezialist der Lungenkrebsbehandlung nur an Regeln dieses Tumors interessiert.

Diese Teilmengen unterscheiden sich teils beträchtlich in ihren Attributverteilungen. So wird Lungenkrebs zumeist erst spät diagnostiziert und in aufgenommenen TNM-Messungen dominieren hohe T-Wertungen die Verteilung. Im Gegensatz dazu stehen Tumoren wie Brust- und Hautkrebs, welche bereits in frühen Stadien diagnostiziert und behandelt

werden. Bei diesen sind nur selten große Tumoren und Metastasen zu entdecken. Entsprechende Unterschiede sollen transparent in Bayes- oder Markov-Netzen dargestellt werden. Durch einen Vergleich der Netzstruktur können Änderungen detektiert werden, die Grundlagen für zu formulierende Hypothesen bieten.

Die in der Netzstruktur dargestellten Abhängigkeiten sollen wiederum genutzt werden, um die Menge der zu untersuchenden Attribute innerhalb der Assoziationsanalyse einzuschränken. Hierfür wird die Verbindung zweier Attribute durch eine Kante in einem Bayes- bzw. Markov-Netz verwendet.

Durch die vorgestellten Filtermaßnahmen soll eine deutliche Reduzierung der Regelmenge erreicht werden, in welcher sich aufgrund der definierten Teilmenge weiterhin für den Anwender relevante Assoziationsregeln befinden.

Der Suchprozess wird somit interaktiv gestaltet. Auf diese Weise verschafft der Benutzer sich mithilfe der Graphischen Modelle einen Überblick über vorhandene Abhängigkeiten, kann einzelne wählen und durch die Assoziationsanalyse Hypothesengrundlagen generieren. Gefundene Ergebnisse können wiederum Anreize für zukünftige klinische Studien geben.

4 Implementation

Die Umsetzung der in Abschnitt 3.3 geplanten Erweiterung des *SAP Patient Data Explorers* wird in den folgenden Abschnitten beschrieben. Hierfür verwendete Systeme werden in Abschnitt 4.1 gelistet. Abschnitt 4.2 beschreibt anhand eines Beispielszenarios die Abfolge der entwickelten Module, deren Funktionsweise in den Abschnitten 4.3 - 4.5 erklärt werden.

4.1 Verwendete Systeme

Die in dieser Arbeit hauptsächlich verwendeten Systeme sollen in diesem Abschnitt kurz vorgestellt werden. Dies umfasst das *SAP HANA* Datenbanksystem (Unterabschnitt 4.1.1) zur Verwaltung der Daten, Analyseprogramme in der Programmiersprache *R* (Unterabschnitt 4.1.2) und dynamische Webseiten Inhalte, welche mit Hilfe der *JavaScript*-Bibliothek *D3* (Unterabschnitt 4.1.3) erstellt wurden.

4.1.1 SAP HANA

SAP HANA ist ein von SAP entwickeltes Datenbanksystem. Es zeichnet sich durch die konsequente Verwendung der *In-Memory*-Technologie aus, bei welcher die Daten ausschließlich im Hauptspeicher (statt auf Festplatten) vorgehalten werden. Daten werden in spaltenorientierter und komprimierter Form abgespeichert. Dieser Umstand und die schnel-

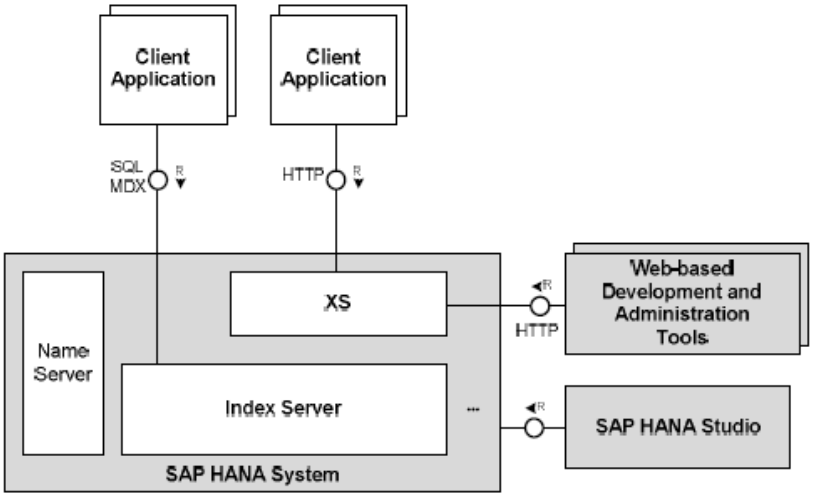


Abbildung 4.1: SAP HANA Architektur (SAP AG, 2013)

leren Zugriffszeiten führen zu einer erheblichen Leistungssteigerung gegenüber bisherigen Festplattensystemen.

Des Weiteren wird die Parallelisierung von Operationen auf Mehrkernprozessorsystemen sowie die Verteilung der *SAP HANA* auf mehrere Server vollständig unterstützt.

Die Datenbank des NCT umfasst die von SAP aggregierten Patientendaten unterschiedlicher Datenquellen, welche zur Analyse durch den *SAP Patient Data Explorer* bereit stehen. Dies geschieht über einen in *SAP HANA* integrierten Webservice, genannt *XS-Engine*, welcher über *HTTP* gestellte *SQL*-Anfragen beantwortet.

4.1.2 R

R ist eine weiterhin in der Entwicklung befindliche Software Umgebung und Programmiersprache für statistische Berechnungen(R CORE TEAM, 2013). Sie wurde aufgrund des reichhaltigen Angebots an bereits vorhandenen Datenanalysepaketen verwendet.

Außerhalb der *SAP HANA*-Architektur wurde ein weiterer Server bereitgestellt, welcher die *R*-Umgebung zur Verfügung stellt. Diese Trennung

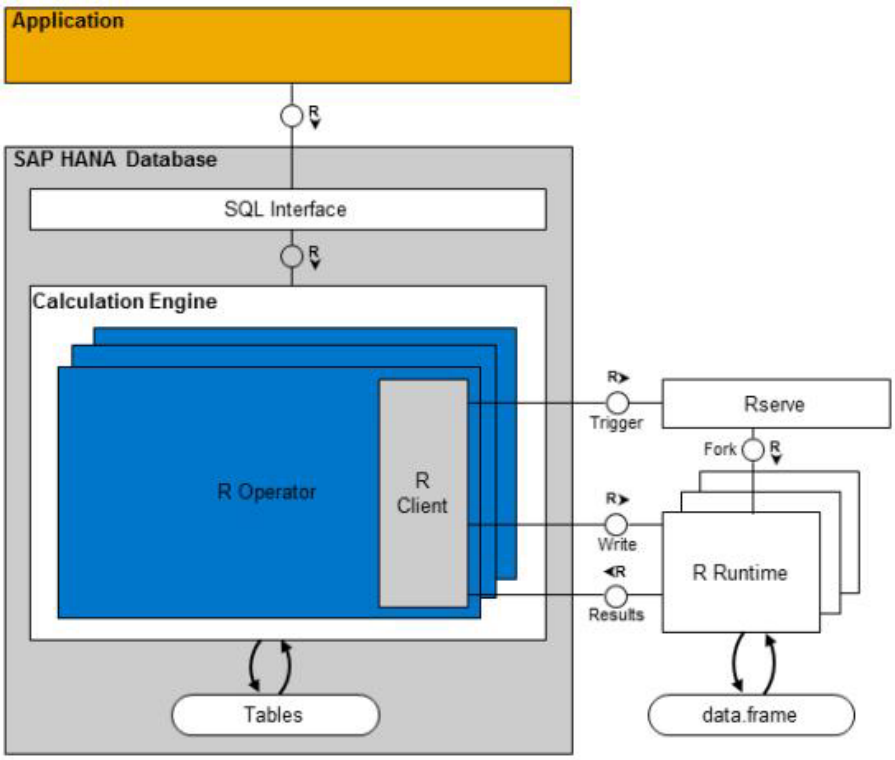


Abbildung 4.2: SAP HANA/RServe-Architektur (SAP AG, 2013)

wurde aus lizenzrechtlichen Gründen umgesetzt, da „The R Project for Statistical Computing“ unter GPL-Lizenz veröffentlicht wurde.

Der *R*-Server kann, von der *SAP HANA* aus, über Prozeduren angesprochen werden. Diese können wie in einer *R*-Skript-Datei Befehlsfolgen beinhalten. Als Ein- und Ausgabeparameter der Prozeduren sind nur Tabellen zulässig, welche innerhalb der Calculation Engine in Dataframes konvertiert werden. Hierbei werden *SQL*-Datentypen in ihre jeweiligen Äquivalente der *R*-Laufzeitumgebung umgewandelt. Als Ausgabe der R-Anfrage müssen die Daten in einem *Dataframe* mit gleichen Spaltennamen vorliegen. Abbildung 4.2 zeigt den strukturellen Aufbau zwischen der *SAP HANA*-Umgebung und der *R*-Umgebung.

Tabelle 4.1 listet die in der vorliegenden Bachelorarbeit verwendeten zusätzlichen Packages und ihren Verwendungszweck auf.

Package	Verwendungszweck	Referenzen
arules	Assoziationsanalyse	HAHSLER et al. (2005)
		HAHSLER et al. (2013)
bnlearn	Berechnung der Bayesnetzstruktur	SCUTARI (2010)
		NAGARAJAN et al. (2013)

Tabelle 4.1: Verwendete Packages innerhalb der Arbeit

4.1.3 D3

D3 (Abkürzung für Data-Driven Documents) ist eine *JavaScript*-Bibliothek, welche die Erstellung von datengesteuerten dynamischen Inhalten auf Webseiten unterstützt. Hierbei werden *HTML*, *SVG* und *CSS* als Ausgabeformate verwendet.

In der vorliegenden Arbeit wird *D3* zur Visualisierung der Datenbankinhalte verwendet. Dies beinhaltet Tabellen der *SAP HANA*-Umgebung und von *R*-Skripten zurückgegebene Daten in *JavaScript Object Notation* (JSON).

4.2 Anwendungsszenario

Zur Erläuterung des entwickelten Programms soll an dieser Stelle ein Überblick der integrierten Module und ihrer Verknüpfungen anhand einer Beispielanfrage gegeben werden. Details zu deren Funktionsweisen sind in den Abschnitten 4.3 bis 4.5 zu finden.

Dargestellte Informationen basieren auf von SAP generierten Testdaten und beanspruchen keinerlei Aussagekraft über medizinische Sachverhalte.

Abbildung A.1 zeigt den Startbildschirm des *SAP Patient Data Explorers*. In unserem Beispielszenario schränken wir die Menge der Patientendaten auf Fälle mit Brustkrebs (ICD-Code: C50) als Erstdiagnose ein. Diese Einstellung können wir in dem Patientenfilter auf der rechten Seite vornehmen. Eine Änderung führt zu einem automatischen Update der derzeit ausgewählten Visualisierung, in unserem Fall ein Balkendiagramm,

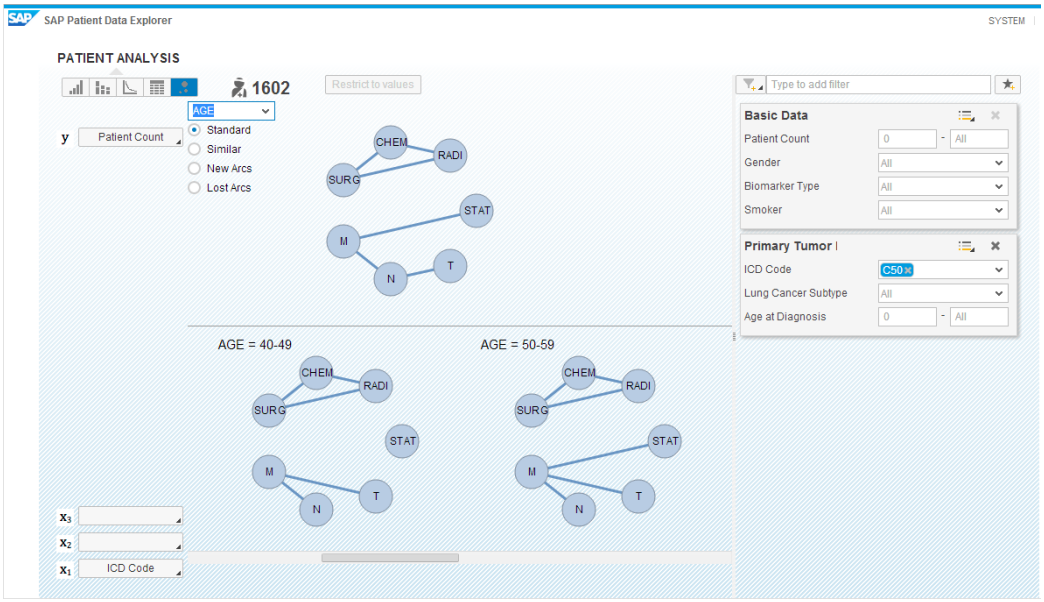


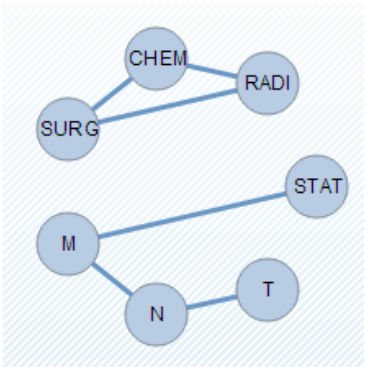
Abbildung 4.3: Aufteilung der Daten nach Alterskohorten und Generierung von Teilmengennetzen; oben: Netz aller Brustkrebspatienten, unten Teilmengennetze per Attributausprägung der Alterskohorten

welches nun die Anzahl der Brustkrebspatienten anzeigt (siehe Abbildung A.2).

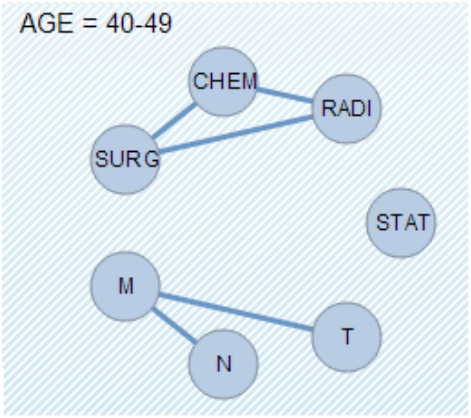
Die in dieser Arbeit beschriebenen Module sind erreichbar über den in Abbildung A.2 markierten Button. Durch einen Klick auf diesen wird das Markov-Netz-Modul gestartet. Hierbei wird für eine Auswahl an Attributen ein Markov-Netz für die Patienten der aktuellen Filtereinstellungen berechnet.

Die *Combobox* auf der linken Seite erlaubt die Wahl eines Attributes, nach welchem die Patientendaten in jeweils eine Teilmenge pro Attributausprägung aufgeteilt werden. Nach kurzer Rechenzeit werden die Markov-Netze der jeweiligen Teilmengen ebenfalls angezeigt. Abbildung A.3 zeigt die Aufteilung nach Alterskohorten.

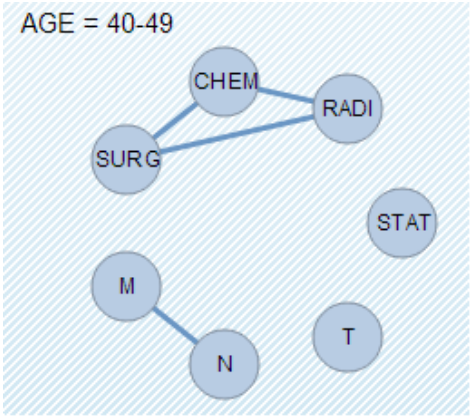
Die *Radio Button Group* unterhalb der *Combobox* lässt Gemeinsamkeiten und Veränderungen der Teilmengennetze im Bezug zum Netz der Grundgesamtheit hervorheben. Die vier wählbaren Optionen sind in Abbildung 4.4 vergleichend dargestellt.



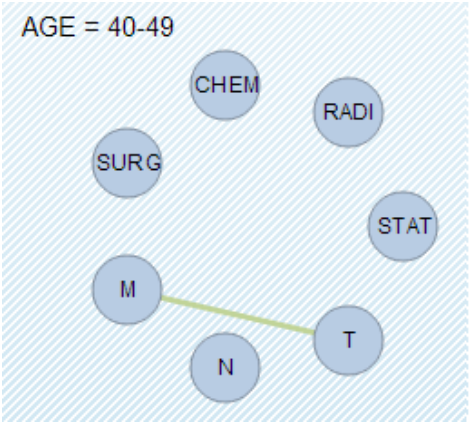
(a) Ansicht des Hauptmengennetzes



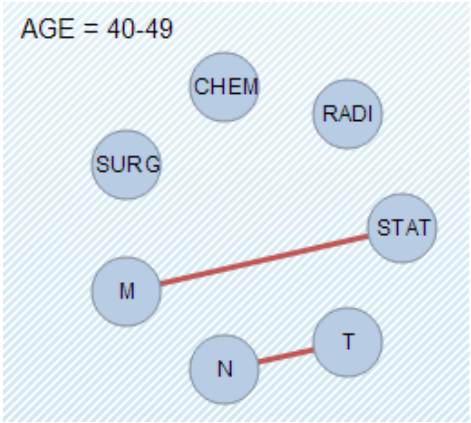
(b) Standardansicht des Netzes



(c) nur gleiche Kanten des Netzes



(d) nur neue Kanten des Netzes



(e) nur gelöschte Kanten des Netzes

Abbildung 4.4: Vergleichsansichten des Hauptnetzes und des Teilmengennetzes für Patienten im Alter von 40-49 Jahren

Durch die Wahl eines Teilmengennetzes wird dieses dem Hauptnetz direkt gegenübergestellt. Wir wählen das Teilmengennetz der Alterskohorte „40-49 Jahre“ aus, da wir die Änderungen der Netzstruktur bezüglich der Knoten *T, N, M* und *STAT* näher untersuchen möchten. Es erfolgt der Aufruf des Vergleichsmoduls, welches die Marginalverteilungen der dargestellten Attribute beider Netze berechnet und in Balkendiagrammen ausgibt. Durch einen *Mouseover*-Effekt werden für den jeweiligen Abschnitt die Wahrscheinlichkeit des Hauptnetzes, des Teilmengennetzes sowie deren Verhältnis (Lift) angezeigt. Die Marginalverteilungen und die Detailansicht sind in Abbildung A.5 dargestellt.

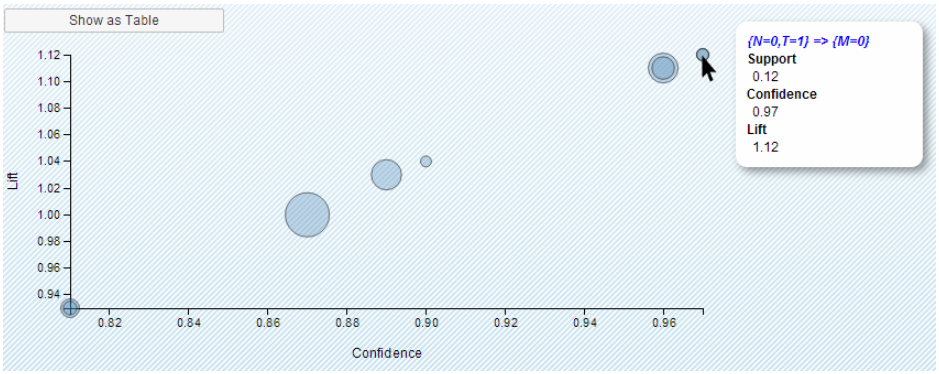
Bisher zeigten wir ungerichtete Attributverknüpfungen und gaben einen Überblick über die zugrunde liegenden Daten. Befinden wir uns in der Vergleichsansicht, so wird eine *Mouseover*-Funktion für das aktuelle Teilmengennetz aktiviert. Fahren wir nun mit der Maus über einen Knoten, wird dieser und seine Adjazenzmenge hervorgehoben. Durch einen Mausklick wird das Assoziationsregel-Modul für die hervorgehobene Menge an Attributen aufgerufen. Wir wählen das Attribut *M* aus, um die Änderungen der Abhängigkeiten in Bezug zum Hauptmengennetz nachzuvollziehen. Es werden die Assoziationsregeln der Attributmenge $\{M,\ N,\ T\}$ ermittelt.

Die Ergebnisse der Assoziationsanalyse werden in einer Tabelle dargestellt. Ausgegeben werden Antezedenz, Konsequenz, Support, Konfidenz und Lift der gefundenen Assoziationsregeln. Die Schaltfläche über der Tabelle erlaubt den Wechsel zu einer alternativen Darstellung als *Bubblechart*. Abbildung 4.5 zeigt die Regelmenge der Attribute *T*, *N* und *M* vergleichend in Tabellen- und *Bubblechart*-Darstellung. Die Abbildungen A.6 und A.7 zeigen diese jeweils im Kontext des Programms.

Um die Übersichtlichkeit des Beispiels zu gewährleisten, wurden nicht alle möglichen Übergänge erläutert. Zur Vervollständigung sind diese in Abbildung 4.6 dargestellt.

Antecedent	Consequent	Support	Confidence	Lift
{N=1}	{M=0}	0.28	0.81	0.93
{N=1,T=2}	{M=0}	0.16	0.81	0.93
{}	{M=0}	0.87	0.87	1
{T=2}	{M=0}	0.55	0.89	1.03
{T=3}	{M=0}	0.10	0.90	1.04
{N=0}	{M=0}	0.54	0.96	1.11
{N=0,T=2}	{M=0}	0.37	0.96	1.11
{T=1}	{M=0}	0.14	0.97	1.12
{N=0,T=1}	{M=0}	0.12	0.97	1.12

(a) Tabellendarstellung



(b) Bubblechartdarstellung

Abbildung 4.5: Vergleich der Regelvisualisierungen

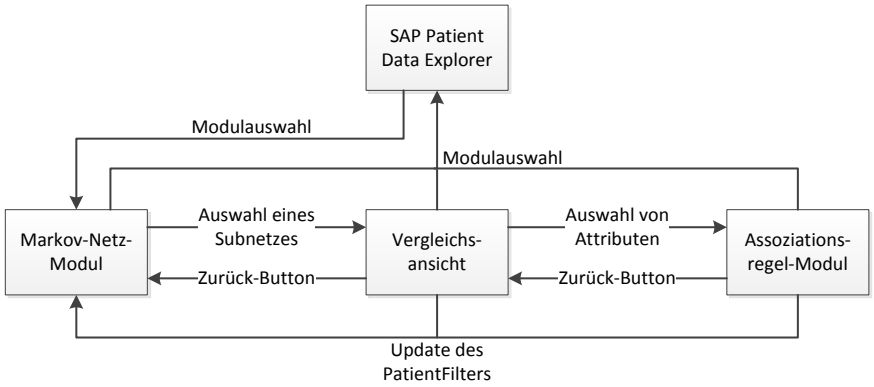


Abbildung 4.6: Übergangsdiagramm

4.3 Markov-Netz-Modul

Um einen schnellen Überblick über Abhängigkeiten der Attribute zu vermitteln, wurde das Markov-Netz-Modul entwickelt. Hierbei wird der durch die Einstellungen des Patientenfilters gefilterte Datensatz als Grundmenge verwendet. Mittels des in Unterabschnitt 4.3.1 vorgestellten Verfahrens werden Markov-Netz-Strukturen der Hauptmenge und deren Teilmengen ermittelt. Unterabschnitt 4.3.2 stellt eine Auswahl an Graphlayouts vor und nimmt Bezug auf die Vergleichbarkeit präsentierter Netzstrukturen.

4.3.1 Strukturlernen von Bayes-/Markov-Netzen

Algorithmen des Strukturlernens von Bayes- und Markov-Netzen verfolgen unterschiedliche Strategien um den Suchraum möglichst effizient zu durchsuchen. Deren Hauptgruppen *constraint-based* (einschränkungsbasierende) und *score-based* (bewertungsorientierte) Algorithmen wurden in (SCUTARI, 2009) unterschieden.

Bei der Wahl des Algorithmus wurde Wert darauf gelegt, dass eine Startstruktur vorgegeben werden kann. Auch wenn diese Funktion bisher nicht implementiert ist, ist es denkbar, dass diese noch umgesetzt wird, um schon bekannte Sachverhalte vorzugeben. Im bisherigen Programmablauf wird der Algorithmus stets mit einer leeren Netzwerkstruktur aufgerufen.

Die Wahl fiel auf den *Hill-Climbing*-Algorithmus, welcher zu den bewertungsorientierten Verfahren zählt. Es handelt sich hierbei um ein *Greedy*-Verfahren, bei welchem die derzeitige Bewertung der Netzstruktur gegen die aller Strukturen mit einer hinzugefügten, gelöschten oder umorientierten Kante gewertet wird. Wird eine bessere Netzstruktur gefunden, so wird der Vorgang mit dieser wiederholt. Die Gefahr nur ein lokales Maximum zu finden bleibt hierbei bestehen. Der Ablauf sei in Algorithmus 1 vereinfacht dargestellt.

Algorithmus 1 *Hill-Climbing*-Algorithmus (NAGARAJAN et al., 2013)

Wähle eine Netzwerkstruktur G über V
Berechne $Score(G)$
Setze $maxScore = Score(G)$ und $maxScore' = maxScore$
So lange $maxScore' > maxScore$ **führe aus**
  Setze $maxScore = maxScore'$
  Für alle Hinzufügungen, Richtungsänderungen, Löschungen von Kanten des Graphen G die zu einem nicht azyklischen Graphen G' führen **führe aus**
    Berechne $Score(G')$
    Wenn $Score(G') > Score(G)$ **dann**
      setze $G = G'$ und $Score(G') = Score(G)$
    Wenn Ende
  Für alle Ende
  Setze $maxScore' = Score(G)$
So lange Ende
Rückgabe: Ausgabe des gerichteten azyklischen Graphens G

Durch das Hinzufügen von zusätzlichen Kanten, kann die Wahrscheinlichkeitsverteilung mit höherer Genauigkeit repräsentiert werden. Jedoch ist nach dem Prinzip der Parsimonie (auch bekannt unter *Occam's razor*), bei einer Wahl aus mehreren Modellen mit gleicher Aussagekraft, das einfachste zu bevorzugen. Die Bewertung eines Graphen (*Score(G)*) wird durch das *Bayesian Information Criterion* (BIC) umgesetzt. Dieses schätzt die Wahrscheinlichkeit des induzierten Modells bezüglich der Daten mittels einer *Maximum-Likelihood*-Schätzung und führt einen Strafterm für die Komplexität des Modells ein (SCHWARZ, 1978).

Während einer Vorstellung eines bereits fortgeschrittenen Prototyps bewerteten Mitarbeiter des NCT die Darstellung von gerichteten Verbindungen als negativ. Um weitgreifende Veränderungen am Programm zu vermeiden, wurde der Strukturlernalgorithmus nicht auf das Lernen von Markov-Netzen angepasst. Stattdessen wurde der bis dato entwickelte Ablauf um eine Moralisierung der Graphen erweitert. Auch wenn dies zusätzlichen Rechenaufwand bedeutet, so konnte die gewünschte Funktionalität auf diese Weise erreicht werden.

4.3.2 Darstellung von Markov-Netzen

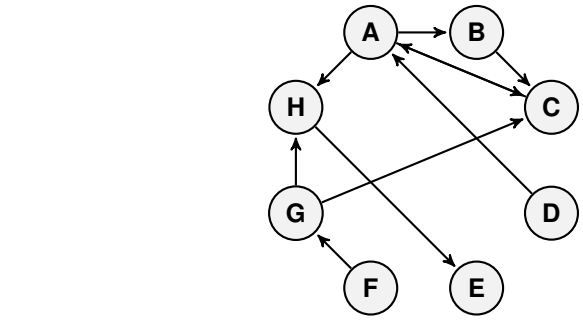
In diesem Abschnitt möchte ich die Problemstellung der Visualisierung ermittelter Netzstrukturen näher erläutern. Zur Darstellung von ungerichteten und gerichteten Graphen wurden eine Vielzahl von Darstellungsformen entwickelt. Eine Auswahl aus dem Buch „Handbook of Graph Drawing and Visualization“ (TAMASSIA, 2007) soll im Folgenden näher betrachtet werden.

Circular (Zirkular) Layout In dieser Darstellung werden alle Knoten in gleichmäßigen Abstand auf einem Kreis platziert. Die Knoten sind somit jeweils einem Eckpunkt auf einem regelmäßigen Polygon zugeordnet. Einzuzeichnende Kanten können sich hierbei jedoch schneiden. Abbildung 4.7a zeigt einen Graphen in zirkularen Layout.

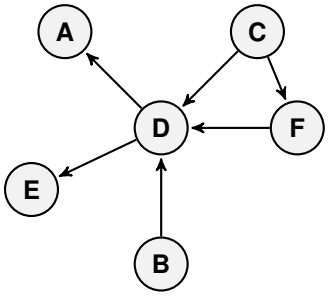
Force-Based (Kräftebasierendes) Layout Statt eine Positionierung vorzugeben, werden in Kräftebasierenden Layouts Anziehungs- und Abstoßungskräfte definiert. Knotenpositionen werden zufällig initialisiert und eine Simulation gestartet. Kanten zwischen zwei Knoten halten diese zusammen, während alle Knoten sich voneinander abstoßen. Mit der Zeit kann der Einfluss der Kräfte gesenkt werden. Sobald sich die Darstellung nicht mehr verändert, kann die Grafik ausgegeben werden. In Abbildung 4.7b ist ein Ergebnis eines Kräftebasierenden Layouts dargestellt.

Tree Layout Hierarchisch organisierte Graphen können in einem Tree Layout dargestellt werden. Hierbei ist eine topologische Ordnung der Knoten vonnöten, welche ebenfalls von Algorithmen wie dem K2-Algorithmus (COOPER und HERSKOVITS, 1992) vorausgesetzt wird. Ein Beispiel für einen in Tree Layout dargestellten Graph ist in Abbildung 4.7c zu sehen.

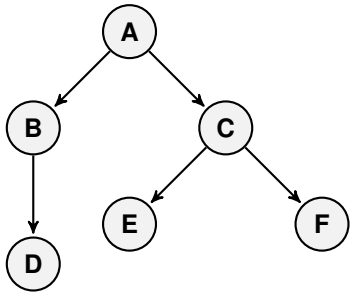
Bei der Wahl der Visualisierung wurde besonderer Wert auf die Vergleichbarkeit der Netze gelegt, welche aus einer festen Knotenmenge und variabler Kantenmenge bestehen.



(a) Circular Layout



(b) Force based Layout



(c) Tree Layout

Abbildung 4.7: Beispiele für Graphenlayouts; unterschiedliche Graphen für bessere Darstellung der Layouteigenschaften

Das Tree Layout wurde ausgeschlossen, da bei der vorhandenen Attributmenge keine eindeutige topologische Ordnung festgelegt werden kann. Kräftebasierende Layouts teilen diese Einschränkung nicht, jedoch ist die Positionierung der Knoten abhängig von aus- und eingehenden Kanten. Da sich diese in den zu vergleichenden Netzen stark ändern können, wurde auch von dieser Lösung abgesehen. Von den drei vorgestellten Layouts eignete sich das Circular Layout am besten. Bei kleinen Knotenmengen bleibt der Graph übersichtlich und durch die feste Positionierung der Knoten können mehrere Graphen leicht miteinander verglichen werden.

Der Anwender wird bei dem Vergleich der Netzstrukturen zusätzlich durch eine optionale Farbkodierung unterstützt. Gemeinsamkeiten und Änderungen können so farblich hervorgehoben werden. Abbildung A.4 zeigte bereits die möglichen Einstellungen. Durch das Hervorheben der Änderungen können stark veränderte Teilmengennetze schnell gesichtet werden und eine nähere Analyse kann eingeleitet werden.

4.4 Vergleichsansicht

Wurde ein interessantes Teilmengennetz per Mausklick ausgewählt, wird die Vergleichsansicht aufgerufen. Hierbei werden das Hauptmengennetz, das gewählte Teilmengennetz und deren Wahrscheinlichkeitsverteilungen direkt gegenübergestellt. Durch den Vergleich der Verteilungen können Änderungen der Netzstruktur nachvollzogen beziehungsweise starke Abhängigkeiten bezüglich des gewählten Split-Attributes festgestellt werden.

Der Vergleich der Wahrscheinlichkeitsverteilungen wurde durch ein gestapeltes Balkendiagramm per Attribut realisiert, vgl. Abbildung A.5. Die vorhandenen Attributsausprägungen sind farblich kodiert. Der jeweils obere Balken beschreibt die Verteilung des Hauptmengennetzes im Vergleich zum unteren Balken des Teilmengennetzes.

Fährt man mit der Maus über einen Balkenabschnitt, so wird dieser hervorgehoben und weitere Informationen in einem Tooltip auf der rechten Seite des Diagramms eingeblendet. Dieser beinhaltet den Namen der Attributausprägung, die Wahrscheinlichkeit des Ereignisses im Haupt- und Teilmengennetz (p_{main} und p_{sub}), sowie den Lift ($p_{\text{sub}} / p_{\text{main}}$).

4.5 Assoziationsregel-Modul

Das Assoziationsregel-Modul vollendet den bisher vorgestellten Ablauf mit der Ausgabe von Assoziationsregeln für im Netz dargestellte Abhängigkeiten. Diese können als Hypothese für kommende Untersuchungen herangezogen werden. Es wird hierbei versucht, die Menge an auszugebenden Regeln möglichst überschaubar zu halten, jedoch gleichzeitig relevante Regeln hervorzuheben.

Die Bestimmung aller Assoziationsregeln beziehungsweise der vorher zu ermittelnden Frequent Item Sets des gesamten Datensets wäre jedoch ein aufwändiger Prozess, da die Anzahl der möglichen Item Sets exponentiell abhängig von der Anzahl der Attribute der Itembasis B ist. Der in Unter-

abschnitt 4.5.2 vorgestellte *Apriori*-Algorithmus verfolgt bereits ein optimiertes Suchschema, welches die Menge an zu überprüfenden Item Sets einschränkt.

In dieser Arbeit werden zudem die Ergebnisse der vorhergehenden Analyseschritte zur Einschränkung der Assoziationsanalyse mit einbezogen. Die zuvor ermittelte Netzstruktur gibt bereits Hinweise auf Abhängigkeiten der Attribute. Wir nutzen dies als heuristische Einschränkung der zu untersuchenden Attributmenge. Innerhalb der Vergleichsansicht kann der Anwender durch den Vergleich der Verteilungen eine auffällige Veränderung eines Attributs A feststellen. Durch einen Klick auf den dazugehörigen Knoten wird das Assoziationsregel-Modul mit diesem und allen im Netz verbundenen Attributen aufgerufen. Wir reduzieren durch den Graphen des Teilmengennetzes $G = (V, E)$ die Itembasis auf die Menge

$$B' = \{A\} \cup \{B \mid (A, B) \in E\}$$

für welche gilt:

$$2^{|B|} \geq 2^{|B'|}$$

Die Berechnung der auf Itembasis B' basierenden Frequent Item Sets und Assoziationsregeln werden in den Unterabschnitten 4.5.1 und 4.5.2 beschrieben. Für die Ausgabe der Assoziationsregeln wurden zwei Darstellungen implementiert, welche in den Unterabschnitten 4.5.3 und 4.5.4 vorgestellt und verglichen werden.

4.5.1 Auswahl des Algorithmus

Das R-Package „arules“ bietet Schnittstellen für die Verwendung des *Apriori-* und des *Eclat*-Algorithmus[1] zur Detektierung aller Frequent Item Sets. Im Gegensatz zu durchgeführten Performancetests in (GARG und KUMAR, 2013) zeigte sich die *Apriori*-Implementation in der Generierung der Regeln in ausgewählten Testszenarios als schneller.

[1] Diese basieren jeweils auf C-Implementierungen von Christian Borgelt, abrufbar unter (WWW: FPM)

Es existieren jedoch noch viele weitere Algorithmen zur Generierung von Assoziationsregeln. Insbesondere wurden *RElim* (BORGELT, 2005b), *SODIM* (BORGELT und KÖTTER, 2011) und *FP-growth* (BORGELT, 2005a) als effektive Algorithmen in Performancetests hervorgehoben. Die Anwendung dieser innerhalb des Assoziationsregel-Moduls wurde durch den zusätzlichen Implementierungsaufwand jedoch vorerst ausgeschlossen. Für spätere Optimierungen wäre der Einsatz benannter Algorithmen empfehlenswert, da die Größe der Datenbank des NCT die der bisherigen Testsets in Zukunft deutlich übersteigen kann.

4.5.2 Apriori-Algorithmus

Der Apriori-Algorithmus macht sich eine charakteristische Eigenschaft des Supports von Item Sets zu nutze, um die Menge an zu untersuchenden Item Sets auf das Erreichen des Mindestsupportwertes zu reduzieren.

Fügen wir dem Item Set I ein Item $A_i \in B$ hinzu, so kann der Support des Item Sets nicht steigen. Es gilt:

$$s(I) \geq s(I \cup \{A_i\})$$

Aus dieser Eigenschaft, welche wir im folgenden als Apriori Eigenschaft bezeichnen, lässt sich ableiten, dass ein n-elementiges Item Set nur ein Frequent Item Set sein kann, wenn auch alle seine (n-1)-elementigen Teilmengen Frequent Item Sets sind.

Algorithmus 2 zeigt die Generierung aller Frequent Item Set Kandidaten der Größe (k+1) (E_{k+1}) gegeben der Menge der Frequent Item Sets der Größe k (F_k) unter Ausnutzung der Apriori Eigenschaft. Algorithmus 3 zeigt die anschließende Bestimmung der Supportwerte aller Kandidaten in E. Übersteigen diese den minimalen Support, werden sie der Menge F hinzugefügt[2].

Der *Apriori*-Algorithmus (vgl. Algorithmus 4) führt diese Schritte abwechselnd durch und gibt anschließend die Menge aller Frequent Item Sets

[2] Pseudo-code Beschreibungen der Algorithmen 2 - 4 von Christian Borgelt, abrufbar unter (WWW: APRIORI)

Algorithmus 2 Kandidatengenerierung (nach (AGRAWAL und SRIKANT, 1994))

Eingabe: Frequent Item Sets der Größe k F_k
Initialisiere Kandidatenmenge $E = \emptyset$
Für alle $f_1, f_2 \in F_k$ **mit** $f_1 = \{i_1, \ldots, i_{k-1}, i_k\}$ **and** $f_2 = \{i_1, \ldots, i_{k-1}, i'_k\}$ **and** $i_k < i'_k$
führe aus
$f = f_1 \cup f_2 = \{i_1, \ldots, i_{k-1}, i_k, i'_k\}$
Wenn $\forall i \in f : f - \{i \in F_k\}$ **dann**
$E = E \cup \{f\}$
Wenn Ende
Für alle Ende
Rückgabe: E

Algorithmus 3 Pruning (nach (AGRAWAL und SRIKANT, 1994))

Eingabe: Kandidatenmenge E, Transaktionsdatenbank T, minimaler Support $s_{\min}$
Für alle $e \in E$ **führe aus**
$s_T(e) = 0$
Für alle Ende
Für alle $t \in T$ **führe aus**
Für alle $e \in E$ **führe aus**
Wenn $e \subseteq t$ **dann**
$s_T(e) = s_T(e) + 1$
Wenn Ende
Für alle Ende
Für alle Ende
$F = \emptyset$
Für alle $e \in E$ **führe aus**
Wenn $s_T(e) \geq s_{\min}$ **dann**
$F = F \cup e$
Wenn Ende
Für alle Ende
Rückgabe: F

gegeben einer Transaktionsbasis T über B und einem minimalen Support Wert s_{min} aus.

Algorithmus 4 *Apriori*-Algorithmus (nach (AGRAWAL und SRIKANT, 1994))

Eingabe: Itembasis B, Transaktionsdatenbank T, minimaler Support $s_{\min}$
Setze Itemsetgröße $k = 1$
Initialisiere zu testende Kandidaten; $E_k := \cup_{i \in B}$
Frequent Item Sets der Größe k; $F_k := \mathit{Pruning}(E_k, T, s_{\min})$
So lange $F_k \neq \emptyset$ **führe aus**
$E_{k+1} = \mathit{Kandidatengenerierung}(F_k)$
$F_{k+1} = \mathit{Pruning}(E_k, T, s_{\min})$
$k = k + 1$
So lange Ende
Rückgabe: $\cup_{j=1}^{k} F_j$

Gegeben einer Menge an Frequent Item Sets können wir nun die zugehörigen Assoziationsregeln bestimmen. Da wir, der Aufgabenstellung nach, auf der Suche nach Ursachen von beispielsweise einer Krebserkrankung sind, suchen wir lediglich nach Assoziationsregeln mit einelementiger Konsequenz.

Algorithmus 5 zeigt die Generierung aller möglichen Assoziationsregeln und deren Überprüfung auf die vom Benutzer angegebene Mindestkonfidenz. Die Berechnung der Konfidenz einer Regel lässt sich optimieren, wenn die Supportwerte der Frequent Item Sets während ihrer Generierung abgespeichert wurden, da mögliche Antezedenzen und Konsequenzen ebenfalls in F enthalten sind.

4.5.3 Tabellendarstellung

Die Ausgabe des Assoziationsregel-Moduls beinhaltet die gefundenen Regeln und die zugehörigen Maße Support, Konfidenz und Lift. Diese werden in einer fünfspaltigen Tabelle ausgegeben (siehe Abbildung A.6), wobei die Antezedenz und Konsequenz der Regel in zwei separate Spalten aufgeteilt werden.

Algorithmus 5 Regelgenerierung

Eingabe: Frequent Item Sets F, minimale Konfidenz c_{min}
  Für alle Item Sets $I \in F$ **führe aus**
    Für alle Item $i \in$ Item Set I **führe aus**
      Wenn $c(I \setminus \{i\} \Rightarrow \{i\}) \geq c_{min}$ **dann**
        Füge Regel $I \setminus \{i\} \Rightarrow \{i\}$ zur Regelmenge R hinzu
      Wenn Ende
    Für alle Ende
  Für alle Ende
  Rückgabe: R

Die Ausgabe lässt sich nach jeder Spalte sortieren. So können schnell Regeln mit besonders hohen Support-, Konfidenz- und Liftwerten detektiert werden. Weiterhin kann der Benutzer für jede Spalte Filter angeben. Auf diese Weise können die generierten Regeln z.B. auf Regeln für noch lebende Patienten (Status = „Lebend“) eingeschränkt werden, um einen Überblick der überlebensratesteigernden Umstände zu erhalten.

4.5.4 Lift-Chart

Bei einer großen Anzahl an Regeln zeigte sich die Tabellendarstellung als unpraktisch. Es wurde daher eine alternative Visualisierung nach (KRUSE und STEINBRECHER, 2010) implementiert. Hierbei werden die Regeln in einem *Bubblechart* anhand der Werte Konfidenz (X-Achse), Lift (Y-Achse) und Support (Radius) eingezeichnet. Die Achsen werden jeweils auf die für die aktuelle Regelmenge relevanten Wertebereiche verkürzt, um die Übersichtlichkeit zu bewahren.

Die genauen Werte werden in einem Tooltip eingeblendet, wenn der Benutzer mit der Maus über einen Kreis fährt. Die Darstellung erlaubt es schnell Regeln mit außergewöhnlich hohen Messwerten zu identifizieren oder Gruppen von Regeln mit ähnlichen Werten zu erkennen. Anhand von Abbildung 4.8 ist zu sehen wie sich Regeln bündeln können. So besitzt zum Beispiel die Regelgruppe auf der rechten Seite jeweils die gleiche Konsequenz. Die Ausreißer auf der linken Seite des Plots zeigen, wie

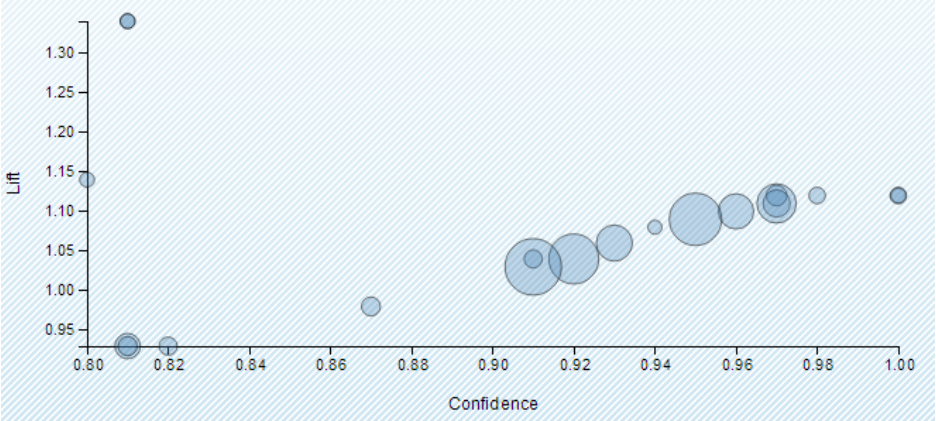


Abbildung 4.8: Beispiel Regelvisualisierung

deutlich sich Regeln abheben kann. Dies kann dem Anwender Anreize geben, entsprechende Regeln näher zu untersuchen.

5 Evaluation

Um die Leistungsfähigkeit des Programms zu überprüfen, wurde das in Abschnitt 5.1 näher beschriebene Testszenario entwickelt. In diesem wird der Arbeitsablauf und dessen Ergebnisse mit einer reinen Assoziationsanalyse verglichen.

5.1 Test auf generierten Daten

Das hier präsentierte Testszenario basiert auf einem thematisch motivierten, synthetischen Datensatz. Dieser beinhaltet Patienten mit Lungenkrebs, Tuberkulose und Bronchitis, sowie Attribute wie Raucher und Atembeschwerden. Es soll geprüft werden, ob in diesem Datensatz vorhandene Abhängigkeiten und Assoziationsregeln mithilfe des in der vorliegenden Arbeit vorgestellten Programms detektiert werden können.

Als Grundlage der Datengenerierung wurde ein synthetischer Datensatz aus (LAURITZEN und SPIEGELHALTER, 1988) verwendet. Abbildung 5.1 zeigt die bedingten Verteilungen des Datensatzes. Mithilfe des Tools „Induction of Network Structures“ (INeS) (BORGELT, 2002) [1] wurde ein Sample mit 100.000 Einträgen erstellt. Hierbei kam es zu vernachlässigbaren Abweichungen von der zugrundeliegenden Verteilung.

Der Datensatz wurde von Lauritzen und Spiegelhalter wie folgt beschrieben:

[1] Programm abrufbar unter (WWW: INES)

D	Atemnot	T	Tuberkulose
L	Lungenkrebs	B	Bronchitis
A	Asienaufenthalt	S	Raucher
X	Röntgenthoraxaufnahme	E	$T \vee L$

$P(A)$		
	A_{no}	0,99
	A_{yes}	0,01

$P(S)$		
	S_{no}	0,50
	S_{yes}	0,50

$P(T \mid A)$	A_{no}	A_{yes}
T_{no}	0,99	0,95
T_{yes}	0,01	0,05

$P(L \mid S)$	S_{no}	S_{yes}
L_{no}	0,99	0,90
L_{yes}	0,01	0,10

$P(B \mid S)$	A_{no}	A_{yes}
B_{no}	0,70	0,40
B_{yes}	0,30	0,60

$P(E \mid L, T)$	L_{no}		L_{yes}	
	T_{no}	T_{yes}	T_{no}	T_{yes}
E_{no}	1	0	0	0
E_{yes}	0	1	1	1

$P(D \mid E, B)$	E_{no}		E_{yes}	
	B_{no}	B_{yes}	B_{no}	B_{yes}
D_{no}	0,90	0,20	0,30	0,10
D_{yes}	0,10	0,80	0,70	0,90

$P(X \mid E)$	E_{no}	E_{yes}
X_{no}	0,95	0,02
X_{yes}	0,05	0,98

Abbildung 5.1: Wahrscheinlichkeitsverteilungen des Testdatensatzes aus (LAURITZEN und SPIEGELHALTER, 1988)

Antezedenz	⇒ Konsequenz	Antezedenz	⇒ Konsequenz
T_{yes}	$\Rightarrow D_{yes}$	$T_{no} \wedge L_{no} \wedge B_{no}$	$\Rightarrow D_{yes}$
L_{yes}	$\Rightarrow D_{yes}$	$T_{yes} \wedge L_{yes} \wedge B_{yes}$	$\Rightarrow D_{yes}$
B_{yes}	$\Rightarrow D_{yes}$	A_{yes}	$\Rightarrow T_{yes}$
$T_{yes} \wedge L_{yes}$	$\Rightarrow D_{yes}$	S_{yes}	$\Rightarrow L_{yes}$
$L_{yes} \wedge B_{yes}$	$\Rightarrow D_{yes}$	S_{yes}	$\Rightarrow B_{yes}$
$B_{yes} \wedge T_{yes}$	$\Rightarrow D_{yes}$		

Tabelle 5.1: Regelmenge des Testdatensatzes

> Atemnot kann durch Tuberkulose, Lungenkrebs, Bronchitis, keines von diesen oder von mehr als einem ausgelöst werden. Ein kürzlicher Aufenthalt in Asien erhöht die Chance einer Tuberkulose Erkrankung, während Rauchen als ein Risikofaktor für Lungenkrebs und Bronchitis bekannt ist. Die Ergebnisse einer einzelnen Röntgenthoraxaufnahme lassen nicht zwischen Tuberkulose und Lungenkrebs unterscheiden, genauso wie das Vorkommen oder die Abwesenheit von Kurzatmigkeit dies nicht ermöglicht.[2]

Die Attributmenge des Datensatzes ist demnach semantisch ähnlich zu dem geplanten Anwendungsgebietes des entwickelten Programms. Dies ermöglicht ein praxisnahes Anwendungsbeispiel anhand der in Tabelle 5.1 aufgelisteten, aus dem Text abgeleiteten Regeln.

Zur Evaluierung des entwickelten Programms wird getestet, ob angegebene Regeln durch Einsatz des Programms ermittelt werden können. Der Aufwand wird hierbei verglichen mit einer reinen Assoziationsanalyse.

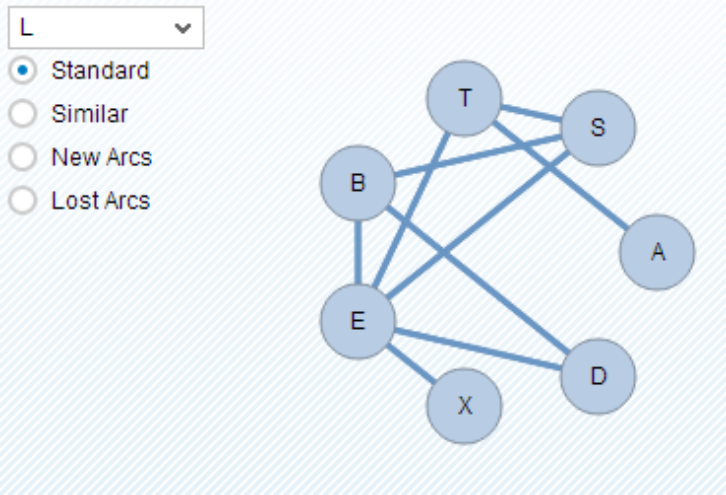
Die Schwierigkeit dieser Aufgabe liegt in den teils sehr geringen Supportwerten der angegebenen Assoziationsregeln. Der kleinste ermittelte Supportwert ist $s(\{T_{yes}, L_{yes}, B_{yes}, D_{yes}\}) = 0,00026$, während der Konfidenzwert der damit verbundenen Regel $c(\{T_{yes}, L_{yes}, B_{yes}\} \Rightarrow \{D_{yes}\}) = 1$ beträgt. Die niedrigen Supportwerte sind bedingt durch die Seltenheit bestimmter Attributausprägungen und ihrer Kombinationen, zum Beispiel $P(T_{yes}) = 0,01$, $P(L_{yes}) = 0,05$ und $P(A_{yes}) = 0,009$. Um diese Assoziations-

[2] sinngemäß übersetzt

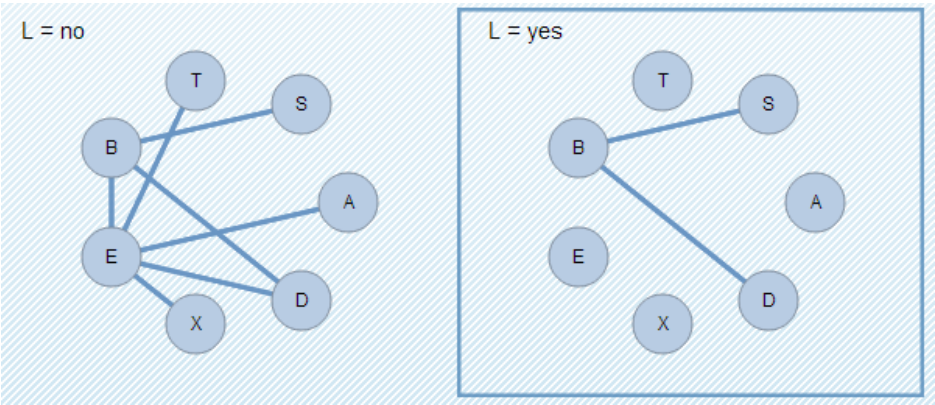
regeln zu ermitteln, müssten wir demnach einen minimalen Supportwert von $s_{\min} < 0,00026$ wählen. Dies hat jedoch zur Folge, dass eine enorm große Anzahl an weiteren Regeln ausgegeben wird. Bei einem gewählten minimalen Konfidenzwert von $c_{\min} = 0,8$ sind es bereits 3.802 Regeln.

Die im Folgenden dargestellte Suchprozedur steht beispielhaft für den Ablauf einer durch das Programm ermöglichten explorativen Analyse. Auf der Suche nach einem Risikofaktor für Lungenkrebs wählen wir L als Split-Attribut und die Menge der betroffenen Patienten (L_{yes}) als weiter zu verarbeitende Teilmenge. Für diese ist sichtbar, dass die Anzahl der Raucher im Vergleich zur Grundgesamtheit stark erhöht ist. Aus diesem Grund verfolgen wir die mit S verbundenen Attribute mithilfe des Assoziationsregel-Moduls weiter. Es werden 8 Regeln hervorgehoben, in denen auch die gesuchte $S_{yes} \Rightarrow L_{yes}$ enthalten ist. Von den gefundenen Regeln hat diese zusammen mit $L_{yes} \Rightarrow S_{yes}$ den höchsten Supportwert von 0,91. Abbildung 5.2 zeigt Ausschnitte des hier beschriebenen Suchablaufs.

Von den angegebenen Regeln konnten mithilfe des Programms und seiner Filtermöglichkeiten alle Regeln bis auf $\{A_{yes}\} \Rightarrow \{T_{yes}\}$ aus dem Datensatz ermittelt werden. Selbst in gefilterten Teilmengen betrug der Support des ItemSets $\{A, T\} = 0,05$, was für die Grundeinstellungen des verwendeten Apriori-Algorithmus ($s_{\min} = 0,1$) nicht ausreichend hoch ist. Durch Anpassungen des Mindestsupportwertes und des Mindestkonfidenzwertes könnten weitere Regeln ausgegeben werden. Jedoch würde dies auch in anderen Anwendungsbeispielen die Anzahl der hervorgehobenen Regeln deutlich erhöhen.



(a) Wahl des Splitting-Attributs L



(b) Wahl der erkrankten Patienten als zu verarbeitende Teilmenge

(c) Suche nach stark veränderten Wahrscheinlichkeitsverteilungen; hier S = Raucher

Antecedent	Consequent	Support	Confidence	Lift
{B=no,L=yes}	{S=yes}	0.37	0.84	0.93
{L=yes}	{S=yes}	0.91	0.91	1
{B=yes,L=yes}	{S=yes}	0.55	0.95	1.05
{S=yes,B=no}	{L=yes}	0.37	1	1
{B=no}	{L=yes}	0.42	1	1
{S=yes,B=yes}	{L=yes}	0.55	1	1
{B=yes}	{L=yes}	0.56	1	1
{S=yes}	{L=yes}	0.91	1	1

(d) Ansicht der Regelmenge der Attribute S, B und L; gesuchte Regel wurde blau hervorgehoben

Abbildung 5.2: Beispiel eines explorativen Analyseablaufs

6

Zusammenfassung

Die vorliegende Arbeit zeigt am Beispiel eines medizinischen Systems die von Licklider angestrebte Zusammenarbeit zwischen Mensch und Maschine. Es konnte gezeigt werden, wie durch das präsentierte interaktive Verfahren die Generierung von Hypothesen im Bereich der Onkologie umgesetzt werden kann.

Die Hypothesenerstellung wurde durch eine Assoziationsanalyse umgesetzt. Im Datensatz gefundene Assoziationensregeln geben Hinweise auf mögliche Zusammenhänge, welche es innerhalb einer klinischen Studie in einem kontrollierten Umfeld zu klären gilt. Aufgrund der hohen Anzahl an Attributsausprägungen innerhalb der Datenbank musste die Menge an zu untersuchenden Item Sets eingeschränkt werden.

Eine Filterung der Attribute konnte durch den Einsatz von Markov-Netzen erreicht werden. Durch das Voraussetzen einer Abhängigkeit innerhalb der berechneten Modelle wurde die Attributmenge der Assoziationsanalyse eingeschränkt. Des Weiteren wurde die Assoziationsanalyse lediglich auf Teilmengen angewandt. Diese Aufteilung ermöglichte die Extraktion von Assoziationsregeln für seltene Diagnosen (mit sehr geringen Supportwert), ohne den Anwender mit einer unnötig großen Zahl an Regeln häufigerer Diagnosen zu überlasten.

Eine Evaluation zeigte, dass das vorgestellte Verfahren die Menge an auszugebenden Assoziationsregeln deutlich senken kann, ohne für den Untersuchungsgegenstand relevante Regeln zu vernachlässigen. Der Suchprozess konnte gegenüber einer ungefilterten Assoziationsanalyse erheb-

lich beschleunigt werden. Es bleibt jedoch zu zeigen, ob das vorgestellte Verfahren allgemeine Anwendbarkeit finden kann.

Der Prototyp wurde in einer Präsentation vor Mitarbeitern des NCT positiv aufgenommen. Daraus ergab sich ein Folgeprojekt, welches die derzeitige Version weiterentwickeln wird. Erweiterungsvorschläge welche im Entwicklungsverlauf des Prototypen aufkamen seien im folgenden Abschnitt abschließend vorgestellt.

6.1 Ausblick

In dieser Arbeit wurde das entwickelte Programm in seinem jetzigen Zustand beschrieben. Hierbei wurden bereits mögliche Erweiterungen angesprochen, welche bisher nicht implementiert werden konnten. Diese sollen nachfolgend nochmals aufgelistet werden, um mögliche Weiterentwicklungen und ihre Bedeutung näher zu erläutern.

Vorgeben einer Netzwerkstruktur Das Vorgeben einer Netzstruktur wurde bereits bei der Auswahl des Algorithmus zur Strukturermittlung beachtet. So ist es denkbar, dass Studien bereits (Un-)Abhängigkeiten belegen, welche in der Netzstruktur enthalten sein sollten. Ein User-Interface, welches das vorgeben einer Netzstruktur ermöglicht, muss jedoch noch entwickelt werden.

Performancesteigerung durch Ermitteln eines Markov-Netzes Dieser Punkt wurde ebenfalls bereits angesprochen. Das bisherige Vorgehen war einer späten Designänderung zu schulden. Durch die direkte Berechnung der Struktur eines Markov-Netzes könnte der Vorgang der Moralisierung eines Bayes-Netzes eingespart werden.

Evidenzpropagation Die ermittelte Netzstruktur wird bisher als Filter der Attributmenge für das Assoziationsregel-Modul verwendet. Die in der Netzstruktur repräsentierten bedingten Unabhängigkeiten können jedoch auch genutzt werden, um mittels Evidenzpropagation direkte Aussagen über die Wahrscheinlichkeit von Attribut-

ausprägungen zu treffen. Hierbei könnte jedes durch bisherige Behandlungen bereits ermittelte Attribut genutzt werden, um übrige Attributsausprägungen besser abzuschätzen.

Dynamisierung der Teilmengendefinition Die bisherige Aufteilung in Teilmengennetze basiert auf nominalen Attributen des verwendeten Datensatzes. Für die Aufteilung nach Alterskohorten wurden die Patienten ihrem Alter nach gruppiert. Von besonderem Interesse wäre es, mehrere Patientengruppen mithilfe des Patientenfilters zu definieren und als Input für vorgestellte Module zu nutzen, um auch Abhängigkeiten sich stark unterscheidender Teilmengen zu vergleichen.

Vergleich der Assoziationsregeln unterschiedlicher Netze Das Assoziationsregel-Modul agiert im bisherigen Zustand lediglich mit der ausgewählten Teilmenge. In Zukunft könnte dies erweitert werden, um die in der Teilmenge ermittelten Regeln auf ihre Konsistenz in anderen Teilmengen oder der Hauptmenge zu überprüfen. Beispielsweise kann es von Interesse sein, inwiefern sich die Werte Support, Konfidenz und Lift bezüglich der zugrundeliegenden Transaktionsdatenbank ändern.

Einsatz schnellerer Verfahren zur Detektierung der Frequent Item Sets Evaluationen der Algorithmen *RElim* (BORGELT, 2005b), *SODIM* (BORGELT und KÖTTER, 2011) und *FPgrowth* (BORGELT, 2005a) zeigten teils deutliche Geschwindigkeitsvorteile gegenüber dem verwendeten *Apriori*-Verfahren zur Bestimmung der Frequent Item Sets. In der geplanten Verwendung des Programms als Hypothesengenerator zur Vorbereitung von klinischen Studien lag der Fokus bisher nicht auf der Optimierung der Geschwindigkeit. Der Algorithmus könnte jedoch bei Bedarf durch alternative Algorithmen ersetzt werden.

Weitere Maße für den Vergleich von Assoziationsregeln Die verwendeten Vergleichsmaße für Assoziationsregeln sind nur eine Auswahl aus vielen. Diese kann nach Bedarf noch ergänzt werden. Beispiele hierfür sind *Recall* und *Specificity* (LAVRAC et al., 1999). Die Tabellendarstellung der Regeln könnte um weitere Spalten erweitert werden. Die Regelvisualisierung als Ballondiagramm könnte dementsprechend um eine flexible Achsenbelegung erweitert werden.

A

Diagramme und Bildschirmaufnahmen

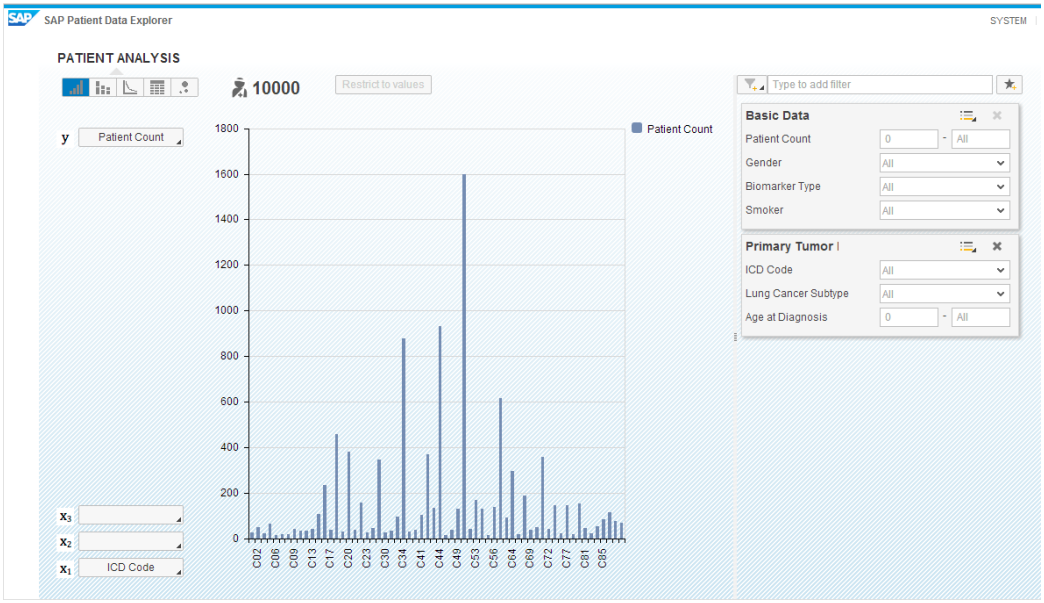


Abbildung A.1: Startbildschirm des *SAP Patient Data Explorers*

Abbildung A.2: Filtereinstellung für Brustkrebspatienten (ICD = C50); rot hervorgehoben: Übergang zu dem Markov-Netz-Modul

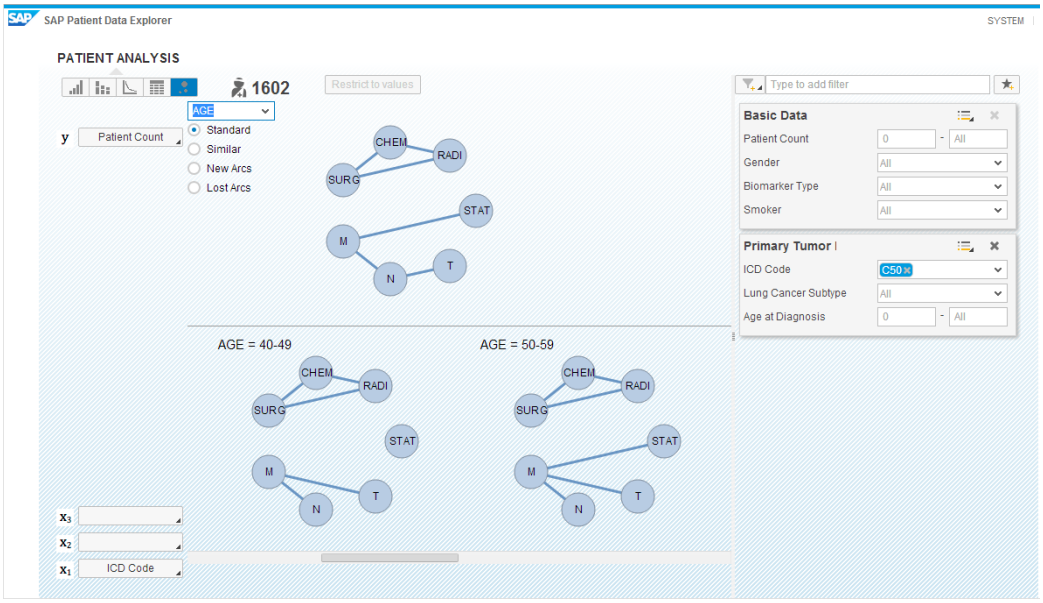
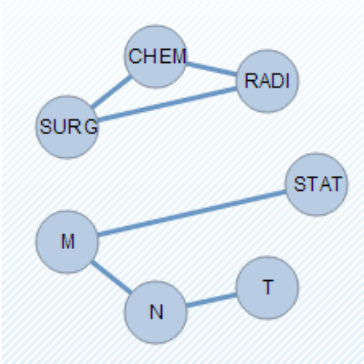
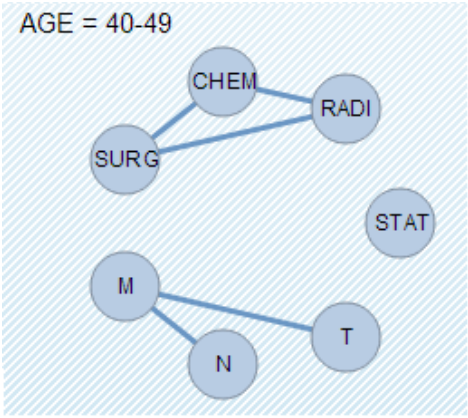


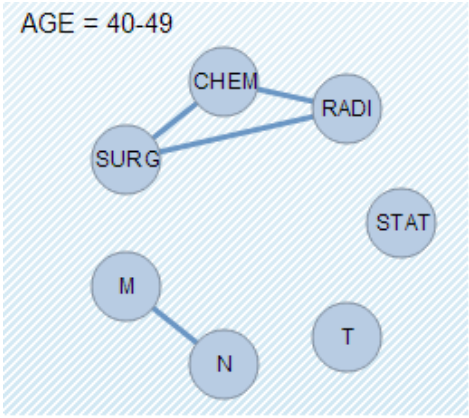
Abbildung A.3: Aufteilung der Daten nach Alterskohorten und Generierung von Teilmengennetzen; oben: Netz aller Brustkrebspatienten, unten Teilmengennetze per Attributausprägung der Alterskohorten, Wiederholung von Abbildung 4.3



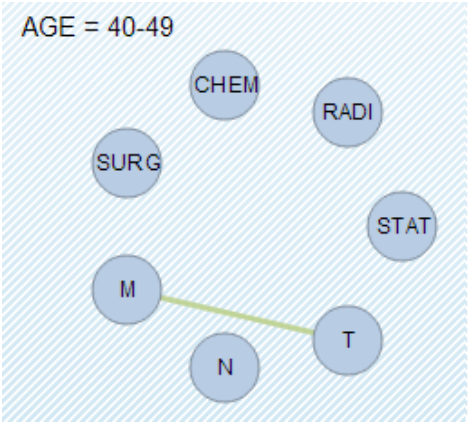
(a) Ansicht des Hauptmengennetzes



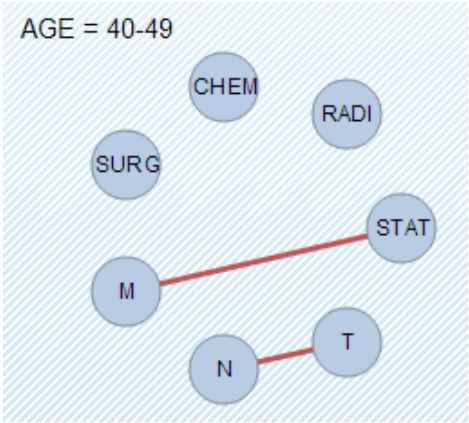
(b) Standardansicht des Netzes



(c) nur gleiche Kanten des Netzes



(d) nur neue Kanten des Netzes



(e) nur gelöschte Kanten des Netzes

Abbildung A.4: Vergleichsansichten des Hauptnetzes und des Teilmengennetzes für Patienten im Alter von 40-49 Jahren, Wiederholung von Abbildung 4.4

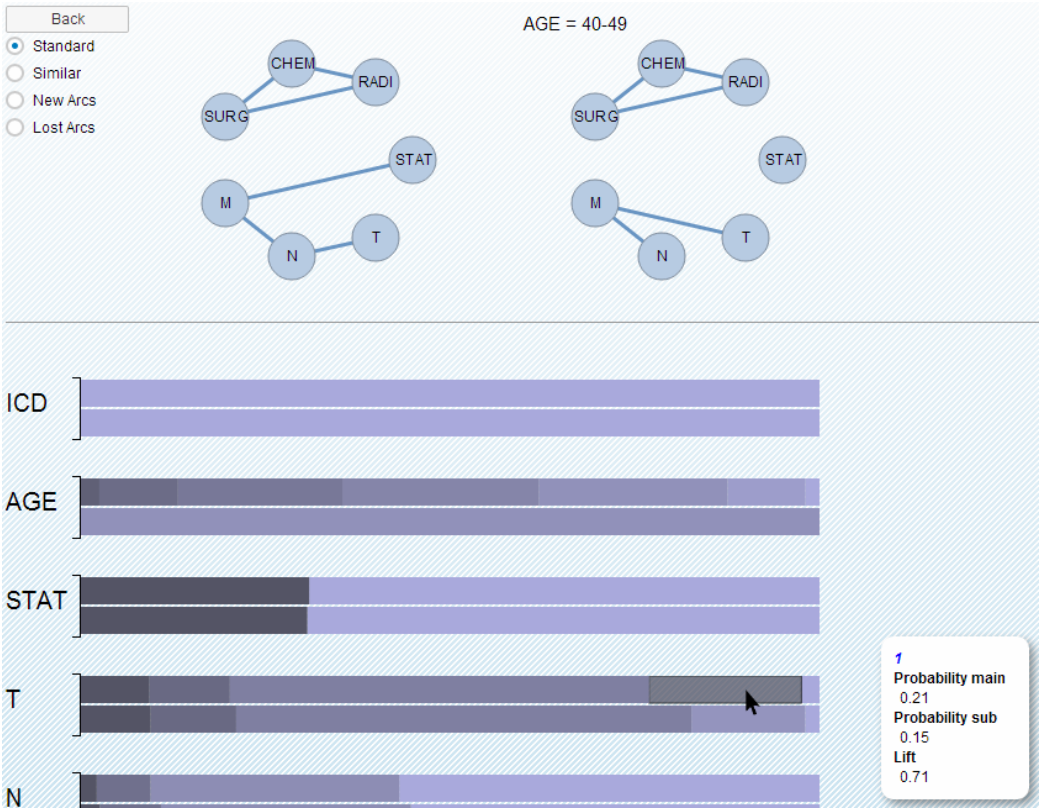
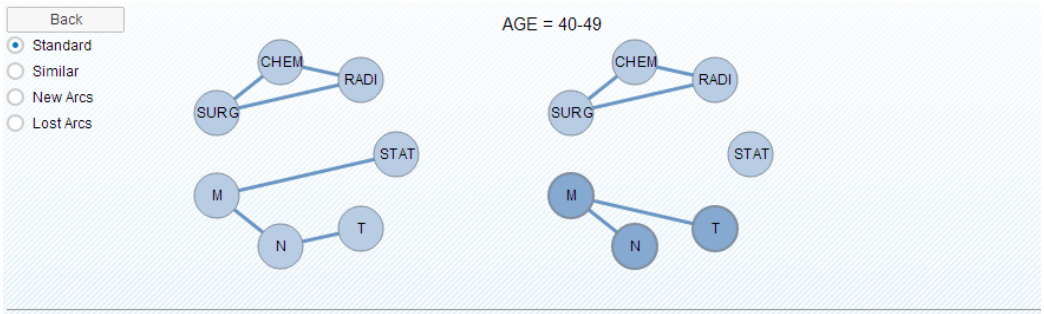


Abbildung A.5: Ausschnitt der Vergleichsansicht mit Tooltip; links oben: Markovnetz aller Brustkrebspatienten, rechts oben: Markovnetz der 40-49 jährigen Brustkrebspatienten, unten: Gestapelte Balkendiagramme per Attribut (oberer Balken: Verteilung des Hauptnetzes, unterer Balken: Verteilung des Teilmengennetzes), rechts unten: Tooltip



Show as Plot

Rule-Table Example

Antecedent	Consequent	Support	Confidence	Lift
{N=1}	{M=0}	0.28	0.81	0.93
{N=1,T=2}	{M=0}	0.16	0.81	0.93
{}	{M=0}	0.87	0.87	1
{T=2}	{M=0}	0.55	0.89	1.03
{T=3}	{M=0}	0.10	0.90	1.04
{N=0}	{M=0}	0.54	0.96	1.11
{N=0,T=2}	{M=0}	0.37	0.96	1.11
{T=1}	{M=0}	0.14	0.97	1.12
{N=0,T=1}	{M=0}	0.12	0.97	1.12

Abbildung A.6: Tabellendarstellung der Regeln für die Attributmenge $\{M,\ N,\ T\}$, generierte Regeln liegen dem Datensatz des Teilmengennetzes zugrunde

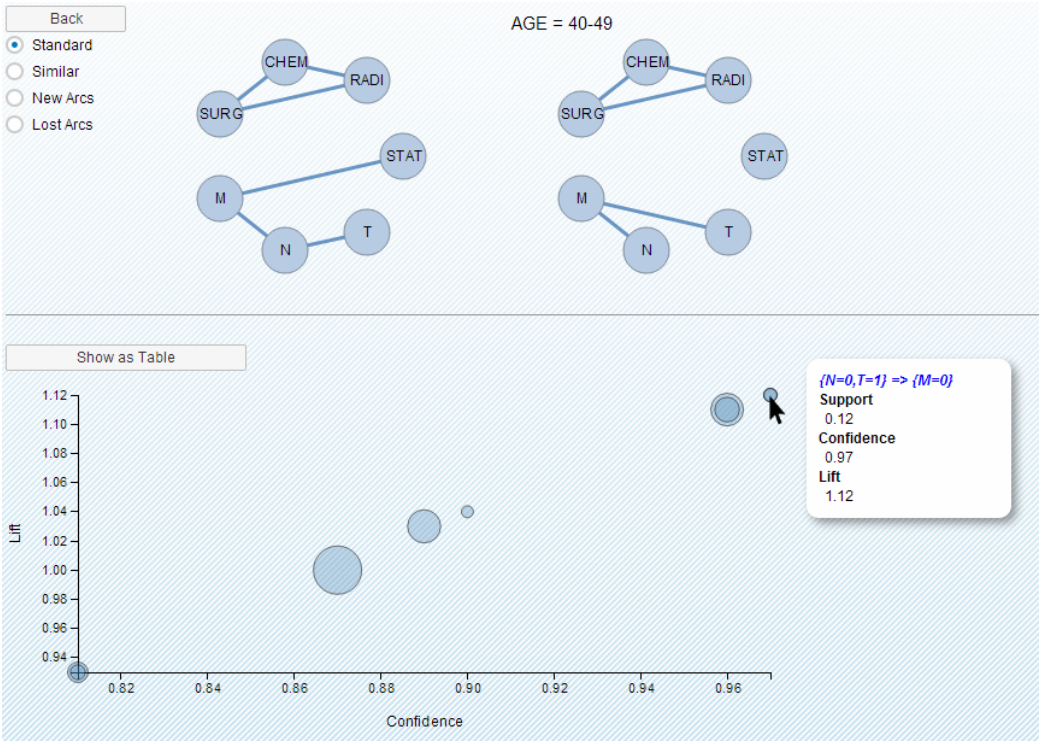


Abbildung A.7: Bubblechart der Assoziationsregeln für die Attributmenge $\{M, N, T\}$, generierte Regeln liegen dem Datensatz des Teilmengennetzes zugrunde, Tooltip blendet Details zur ausgewählten Regel ein

B
Listen

Abkürzungsverzeichnis

Akronym	Bedeutung
NCT	Nationales Centrum für Tumorerkrankungen Heidelberg
SQL	Structured Query Language
HTTP	Hypertext Transfer Protocol
XS	Extended Application Services
ICD	Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme
IS-H	Industry Solution Healthcare
RAM	Random-Access Memory
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
SVG	Scalable Vector Graphics
D3	Data-Driven Documents
JSON	JavaScript Object Notation

Abbildungsverzeichnis

C

Quellenverzeichnis

[AGRAWAL und CHOUDHARY 2011] A. Agrawal und A. Choudhary. **Identifying HotSpots in Lung Cancer Data Using Association Rule Mining**. pp. 995–1002. 2011, IEEE.

[AGRAWAL et al. 1993] R. Agrawal, T. Imieliński und A. Swami. **Mining association rules between sets of items in large databases**. pp. 207–216. 1993, ACM Press.

[AGRAWAL und SRIKANT 1994] R. Agrawal und R. Srikant. **Fast Algorithms for Mining Association Rules in Large Databases**. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, pp. 487–499. 1994, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[AHMED et al. 2013] K. Ahmed, T. Jesmin und M. Zamilur Rahman. **Early Prevention and Detection of Skin Cancer Risk using Data Mining**. International Journal of Computer Applications, Vol. 62(4):1–6, 2013.

[AHMEDMEDJAHED et al. 2013] S. AhmedMedjahed, T. Ait Saadi und A. Benyettou. **Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules**. International Journal of Computer Applications, Vol. 62(1):1–5, 2013.

[ATKINSON et al. 2001] A. J. Atkinson, W. A. Colburn, V. G. DeGrutolla, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley,

B. A. Spilker, J. Woodcock und S. L. Zeger. **Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework**. Clinical Pharmacology & Therapeutics, Vol. 69(3):89–95, 2001.

[BORGELT 2002] C. Borgelt. **Graphical models: methods for data analysis and mining**. J. Wiley, New York, 2002.

[BORGELT 2005a] C. Borgelt. **An implementation of the FP-growth algorithm**. pp. 1–5. 2005, ACM Press.

[BORGELT 2005b] C. Borgelt. **Keeping things simple: finding frequent item sets by recursive elimination**. pp. 66–70. 2005, ACM Press.

[BORGELT und KÖTTER 2011] C. Borgelt und T. Kötter. **Mining Fault-Tolerant Item Sets Using Subset Size Occurrence Distributions**. In: D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Gama, E. Bradley und J. Hollmén, Eds., Advances in Intelligent Data Analysis X, Vol. 7014, pp. 43–54. 2011. Springer Berlin Heidelberg, Berlin, Heidelberg.

[BURKE et al. 1997] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, J. Harrell, F E, J. R. Marks, D. P. Winchester und D. G. Bostwick. **Artificial neural networks improve the accuracy of cancer survival prediction**. Cancer, Vol. 79(4):857–862, 1997. PMID: 9024725.

[CASTILLO 1997] E. Castillo. **Expert systems and probabilistic network models**. Monographs in computer science. Springer, New York, 1997.

[COOPER und HERSKOVITS 1992] G. F. Cooper und E. Herskovits. **A Bayesian method for the induction of probabilistic networks from data**. Machine Learning, Vol. 9(4):309–347, 1992.

[FÄRBER et al. 2012] F. Färber, N. May, W. Lehner, P. Groß e, I. Müller, H. Rauhe und J. Dees. **The SAP HANA Database – An Architecture Overview.** IEEE Data Eng. Bull., Vol. 35(1):28–33, 2012.

[FERRUCCI et al. 2010] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer und C. A. Welty. **Building Watson: An Overview of the DeepQA Project**. AI Magazine, Vol. 31(3):59–79, 2010.

[FLESCH und LUCAS 2007] I. Flesch und P. J. Lucas. **Markov Equivalence in Bayesian Networks**. In: P. Lucas, J. A. Gámez und A. Salmerón, Eds., Advances in Probabilistic Graphical Models, Vol. 214, pp. 3–38. 2007. Springer Berlin Heidelberg, Berlin, Heidelberg.

[GARG und KUMAR 2013] K. Garg und D. Kumar. **Comparing the Performance of Frequent Pattern Mining Algorithms**. International Journal of Computer Applications, Vol. 69(25):21–28, 2013.

[GÖRZ 2000] G. Görz. **Handbuch der künstlichen Intelligenz**. Oldenbourg, München; Wien, 2000.

[HAHSLER et al. 2013] M. Hahsler, C. Buchta, B. Gruen und K. Hornik. **arules: Mining Association Rules and Frequent Itemsets**. 2013. R package version 1.0-15.

[HAHSLER et al. 2005] M. Hahsler, B. Gruen und K. Hornik. **arules – A Computational Environment for Mining Association Rules and Frequent Item Sets**. Journal of Statistical Software, Vol. 14(15):1–25, 2005.

[KRUSE et al. 2011] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, G. Ruß und M. Steinbrecher. **Computational Intelligence**. Vieweg Teubner, 2011.

[KRUSE und STEINBRECHER 2010] R. Kruse und M. Steinbrecher. **Visual data analysis with computational intelligence methods**. Bulletin of the Polish Academy of Sciences: Technical Sciences, Vol. 58(3), 2010.

[LAURITZEN und SPIEGELHALTER 1988] S. Lauritzen und D. J. Spiegelhalter. **Local computations with probabilities on graphical structures and their application to expert systems (with discussion)**. Journal of the Royal Statistical Society series B, Vol. 50:157–224, 1988.

[LAVRAC et al. 1999] N. Lavrac, P. A. Flach und B. Zupan. **Rule Evaluation Measures: A Unifying View**. In: Proceedings of the 9th International Workshop on Inductive Logic Programming, ILP '99, pp. 174–185. 1999, Springer-Verlag, London, UK, UK.

[LICKLIDER 1992] J. C. R. Licklider. **Man-Computer Symbiosis**. IEEE Ann. Hist. Comput., Vol. 14(1):24–, 1992.

[MCNICHOLAS et al. 2008] P. McNicholas, T. Murphy und M. O'Regan. **Standardising the lift of an association rule**. Computational Statistics & Data Analysis, Vol. 52(10):4712 – 4721, 2008.

[NAGARAJAN et al. 2013] R. Nagarajan, M. Scutari und S. Lèbre. **Bayesian networks in R with applications in systems biology**. Springer, New York, NY, 2013.

[PETROVSKI und MCCALL 2001] A. Petrovski und J. McCall. **Multi-objective Optimisation of Cancer Chemotherapy Using Evolutionary Algorithms**. In: Evolutionary Multi-Criterion Optimization, Vol. 1993. 2001. Springer Berlin Heidelberg, Berlin, Heidelberg.

[R CORE TEAM 2013] R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[RAVDIN und CLARK 1992] P. M. Ravdin und G. M. Clark. **A practical application of neural network analysis for predicting outcome of individual breast cancer patients**. Breast cancer research and treatment, Vol. 22(3):285–293, 1992. PMID: 1391994.

[SAP AG 2013] SAP AG. **SAP HANA System Landscape Guide**. Whitepaper 1.0, SAP AG, 2013.

[SARVESTANI et al. 2010] A. S. Sarvestani, A. A. Safavi, N. Parandeh und M. Salehi. **Predicting breast cancer survivability using data mining techniques**. 2010, IEEE.

[SCHWARZ 1978] G. Schwarz. **Estimating the Dimension of a Model**. The Annals of Statistics, Vol. 6(2):461–464, 1978.

[SCUTARI 2009] M. Scutari. **Learning Bayesian Networks with the bnlearn R Package**. arXiv:0908.3817 [stat], 2009. Journal of Statistical Software (2010), 35(3), 1-22.

[SCUTARI 2010] M. Scutari. **Learning Bayesian Networks with the bnlearn R Package**. Journal of Statistical Software, Vol. 35(3):1–22, 2010.

[SOBIN et al. 2011] L. H. Sobin, M. K. Gospodarowicz und C. Wittekind. **TNM Classification of Malignant Tumours**. John Wiley & Sons, 2011.

[TAMASSIA 2007] R. Tamassia. **Handbook of Graph Drawing and Visualization (Discrete Mathematics and Its Applications)**. Chapman & Hall/CRC, 2007.

[TAN et al. 2002] K. Tan, E. Khor, J. Cai, C. Heng und T. Lee. **Automating the drug scheduling of cancer chemotherapy via evolutionary computation**. Artificial Intelligence in Medicine, Vol. 25(2):169 – 185, 2002.

[WWW: APRIORI] **Pseudo-code des Apriori Algorithmus von Christian Borgelt**. `http://www.borgelt.net/docs/apriori.pdf`. Accessed January 2014.

[WWW: FPM] **Frequent Pattern Mining Implementierungen von von Christian Borgelt**. `http://www.borgelt.net//fpm.html`. Accessed January 2014.

[WWW: IMDB] **Internet Movie Database**. `http://www.imdb.com/`. Accessed January 2014.

[WWW: INES] **Induction of Network Structures**. `http://www.borgelt.net/ines.html`. Accessed February 2014.

[WWW: SAP] **SAP Healthcare Projects - Patient Data Explorer**. `http://www.sap-innovationcenter.com/2013/09/19/medical-explorer/`. Accessed January 2014.

[WWW: WATSON ONCOLOGY] **Einsatz von Watson im Gesundheitswesen**. `http://www-03.ibm.com/press/us/en/pressrelease/37235.wss`. Accessed February 2014.

[YADAV et al. 2013] R. Yadav, Z. Khan und H. Saxena. **Article: Chemotherapy Prediction of Cancer Patient by using Data Mining Techniques**. International Journal of Computer Applications, Vol. 76(10):28–31, 2013. Published by Foundation of Computer Science, New York, USA.

[ZUBI und SAAD 2011] Z. S. Zubi und R. A. Saad. **Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer**. In: Proceedings of the 10th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED'11, pp. 32–37. 2011, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA.

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Bachelorarbeit selbstständig, nur unter Zuhilfenahme der aufgeführten Quellen und Hilfsmittel, verfasst zu haben.

Die Arbeit wurde weder einer anderen Prüfungsbehörde vorgelegt noch veröffentlicht.

Datum: .....

(Unterschrift)