Buchempfehlung

Dieses Projekt behandelt ein Uni-Projekt bei dem ein Bayes Netz für einen vorgegeben Datensatz implementiert wird.

Das Repository kann auf Github über den folgenden Link gefunden warden.
https://github.com/leunark/wbs-bayesian-network

Abhängigkeiten

- dlib library: http://dlib.net/
- sqlite library: https://www.sqlite.org/cintro.html

Build

Falls sie das Projekt selber weiter entwickeln oder bauen wollen, dann müssen die folgenden beiden Schritte gemacht werden:

1. Add $(LocalDebuggerWorkingDirectory) to Configuration Properties / VC++ Directories / Include Directory
2. Add /bigobj to: Configuration Properties / C/C++ / Command Line / Additional Options

Aufgabenbeschreibung

"Es soll eine Buchempfehlung ausgesprochen werden. Zur Wahl steht nur eine begrenzte Zahl an Büchern. Als mögliche Informationen können die Angaben zu Altersgruppe, Geschlecht, Familienstand, Kinderzahl, Einkommen, Bildungsstand und Beruf vorliegen. Diese müssen nicht vollständig sein. Erstellen Sie ein Bayes Netz, welches die Zusammenhänge modelliert und plausibel auf Basis der Beispieldaten gefüllte CPTs nutzt. Legen Sie diesem Bayes Netz Beispieleingaben vor und geben Sie das Klassifikationsergebnis geeignet aus.
Entwickeln Sie eine Software, welche bei Eingabe (Datei, vgl. Beispielformat) von Testdaten die entsprechenden Klassifikationen mit Hilfe der Bayes Netz Implementierung geeignet bestimmt und ausgibt."

Ausführung

Das Programm benötigt die Datei "P001.csv" im selben Verzeichnis wie die .exe.
Dann kann das Programm ausgeführt und nur mit den Zahlen 0-9 bedient werden.
Momentan ist dieses Programm nur für die Beispieldatei ausgelegt, kann aber bei Bedarf einfach erweitert werden, da bis auf die Darlegung der Struktur alles vollautomatisch läuft. Der Datei können neue Beispieldatensätze hinzugefügt werden, die Struktur darf sich aber nicht verändern.

Strategie

Es handelt sich bei den Daten um eine CSV-Datei mit Einträgen zu dem Bayes Netz. Jede Spalte repräsentiert ein Knoten im Netz (außer Nr), die miteinander verstrickt sind. Die Struktur für ein Bayes Netz erwartet Expertenwissen, das selber manuell erstellt werden muss. Dafür wurde das Tool Netica verwendet. Theoretisch ist es möglich, wenn ausreichend Daten zur Verfügung stehen, maschinelles Lernen anzuwenden, um Struktur und Kantenstärke herauszufinden. Dies würde aber den Scope sprengen.

Die Datei "buchempfehlung.neta" ist ein Netica File, das mit Netica geöffnet werden kann, um die zugrunde liegende Struktur dieser Aufgabe zu sehen, alternativ ist auch ein Bild beigefügt ("buchempfehlung.jpg").

Sind einmal die verschiedenen Nodes und States in dem Programm eingepflegt, wird der Rest vollautomatisch generiert.

Im ersten Schritt werden die Daten aus "P001.csv" gelesen und in eine SQLite Datenbank geschrieben.

Danach wird für jeden Knoten alle möglichen Kombinationen mit den Elternknoten bestimmt, um die CPT (Condition Probability Table) zu erstellen. Die Generation der CPT findet in der Methode "*startGeneratingCPT()*" der Klasse *Node* statt. Tatsächlich ist sie aber nur ein Trigger für den Aufruf der privaten rekursiven Methode "*generateCPT(...)*".
Nun wird für jede CPT jede Kombination der States durchiteriert, um die Wahrscheinlichkeiten zu berechnen. Hier kommt auch die lokale Datenbank zum Einsatz, da für die Berechnung der CPT, Häufigkeiten von den Kombinationen bestimmt werden müssen. Dafür wird immer ein SQL Statement als String gebaut und dann auf der Datenbank ausgeführt. Ein Callback verarbeitet dann die Daten, um die errechneten Wahrscheinlichkeiten zurückgeben zu können. Sind keine Daten in der Datenbank, bedeutet dies, dass keine Informationen in das Netz geladen werden können. In diesem Fall sind alle Zustände des Knoten gleichwahrscheinlich für die jeweilige Kombination. Schließlich kann die Wahrscheinlichkeit dann in die Logik der dlib/bayesiannetwork-Bibliothek geschrieben werden.

Im letzten Schritt werden Eingaben für verfügbare Evidenzen vom User abgefragt, um sie in das Netz einzupflegen. Das Buch mit der höchsten Wahrscheinlichkeit wird empfohlen.

Es wird zusätzlich eine Log-Datei erstellt, um den Vorgang etwas nachvollziehen zu können, und ein Einblick in die Wahrscheinlichkeiten des Bayes Netzes zu gewähren. Nach dem Ausführen des Programms erscheint eine Log-Datei im selben Pfad von wo die .exe gestartet wurde!

Ergebnis

Für die Validierung des Ergebnisses wurden die CPTs stichpropenartig manuell berechnet. So kann die Richtigkeit des Programmentwurfs garantiert werden.
Bayes Netze dienen der kompakten Speicherung und Verarbeitung unsicheren Wissens. Neue Informationen an einer Stelle des Netzes, in Form von Wahrscheinlichkeiten, wirken sich unter Umständen auf das gesamte Netz aus.
(vgl. Marc Wagner, Vortrag im Seminar " Bayes Netze und Data-Mining", 10. Februar 2000)
Sind Evidenzen vorhanden und werden diese in das Netz eingepflegt kann sichtlich erkannt werden, dass sich die Wahrscheinlichkeiten anderer Knoten anpassen, je nach Knoten sogar durch das ganze Netz.

Verschiedene Tests mit unterschiedlichen Evidenzen zeigten jedoch, dass das Bayes Netz sehr stark von seiner Struktur abhängt. Bei der Erstellung einer Struktur für die Knoten wurde in der Anfangsphase nur auf das Allgemeinwissen vertraut, wie bestimmte Knoten von anderen abhängen.

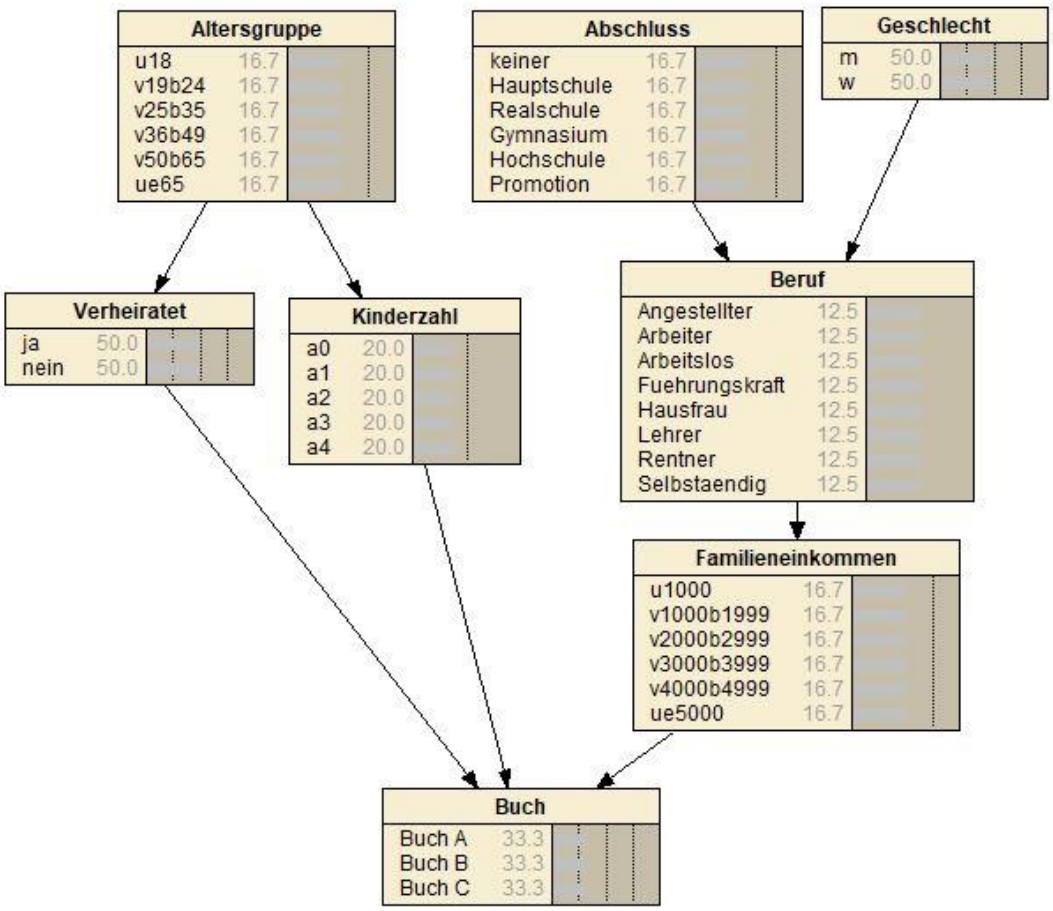

Abb. 1: Netica Buchempfehlung 1

Diese Abbildung stellt unseren ersten Entwurf dar für den Datensatz "P001.csv". Hier wurde in erster Linie Altersgruppe als Bedingung für Verheiratet und Kinderzahl gesetzt, weil mit zunehmenden Alter natürlich die Wahrscheinlichkeit von Kindern und einer Ehe logischerweise zunimmt.
Abschluss ist hier Elternknoten von Beruf, denn nur manche Berufe z.B. Lehrer können nur mit einem guten Abschluss erreicht werden. Wiederum Geschlecht beeinflusst den Beruf insofern, dass v.a. Führungskräfte eher weniger Frauen sind, oder speziell als Hausfrau kann mit der Bezeichnung wohl kein Mann tätig sein. Der Abschluss beeinflusst wiederum das Familieneinkommen und schließlich zeigen die untersten Knoten auf Buch.

Diese Struktur erwies sich jedoch als sehr problematisch nachdem die Implementierung des Programmes vollendet war. In dem Datensatz kann beispielsweise gesehen werden, dass ausschließlich junge Männer das Buch B leihen.

Nr	Altersgru	Geschlech	Verheirat	Kinderzah	Abschluss	Beruf	Familiene	Buch
21	19-24	m	ja	0	Gymnasium	Angestellter	2000-2999	Buch_B
25	19-24	m	nein	0	Gymnasium	Angestellter	3000-3999	Buch_B
90	19-24	m	nein	0	Gymnasium	Arbeiter	2000-2999	Buch_B
0	19-24	m	nein	0	keiner	Angestellter	3000-3999	Buch_B
37	<18	m	nein	0	keiner	Angestellter	2000-2999	Buch_B
52	<18	m	ja	0	keiner	Angestellter	2000-2999	Buch_B
55	19-24	m	nein	0	keiner	Arbeiter	1000-1999	Buch_B
56	<18	m	nein	0	keiner	Angestellter	3000-3999	Buch_B
73	19-24	m	nein	0	keiner	Arbeitslos	1000-1999	Buch_B
74	<18	m	nein	0	keiner	Angestellter	2000-2999	Buch_B
77	19-24	m	nein	0	keiner	Arbeiter	3000-3999	Buch_B
80	19-24	m	nein	0	keiner	Angestellter	2000-2999	Buch_B
85	19-24	m	nein	0	keiner	Arbeiter	2000-2999	Buch_B
94	19-24	m	nein	0	keiner	Arbeiter	2000-2999	Buch_B
11	<18	m	nein	0	Realschule	Angestellter	3000-3999	Buch_B
32	<18	m	nein	0	Realschule	Angestellter	1000-1999	Buch_B
65	19-24	m	nein	0	Realschule	Arbeitslos	2000-2999	Buch_B
76	19-24	m	nein	0	Realschule	Arbeiter	2000-2999	Buch_B
93	<18	m	nein	0	Realschule	Arbeiter	2000-2999	Buch_B
96	19-24	m	nein	1	Realschule	Arbeitslos	1000-1999	Buch_B

Abb. 2: P001.csv gefiltert nach Buch B

Wählt man aber diese Eigenschaften als Evidenzen für die ersterstellte Struktur, dann werden die Wahrscheinlichkeiten unerwartet anders berechnet.

p(Altersgruppe=<18) = 0
p(Altersgruppe=19-24) = 1
p(Altersgruppe=25-35) = 0
p(Altersgruppe=36-49) = 0
p(Altersgruppe=50-65) = 0
p(Altersgruppe=>65) = 0
p(Verheiratet=nein) = 1
p(Verheiratet=ja) = 0
p(Kinderzahl=0) = 1
p(Kinderzahl=1) = 0
p(Kinderzahl=2) = 0
p(Kinderzahl=3) = 0
p(Kinderzahl=4) = 0
p(Geschlecht=m) = 1
p(Geschlecht=w) = 0
p(Abschluss=keiner) = 0.39
p(Abschluss=Hauptschule) = 0.06
p(Abschluss=Realschule) = 0.17
p(Abschluss=Gymnasium) = 0.1
p(Abschluss=Hochschule) = 0.25
p(Abschluss=Promotion) = 0.03
p(Beruf=Angestellter) = 0.408214
p(Beruf=Arbeiter) = 0.216071
p(Beruf=Arbeitslos) = 0.144405
p(Beruf=Fuehrungskraft) = 0.0835714
p(Beruf=Hausfrau) = 0

p(Beruf=Lehrer) = 0.0178571
p(Beruf=Rentner) = 0.0505952
p(Beruf=Selbstaendig) = 0.0792857
p(Familieneinkommen=<1000) = 0.0361012
p(Familieneinkommen=1000-1999) = 0.251161
p(Familieneinkommen=2000-2999) = 0.46352
p(Familieneinkommen=3000-3999) = 0.145824
p(Familieneinkommen=4000-4999) = 0
p(Familieneinkommen=5000 und mehr) = 0.103393
p(Buch=Buch_A) = 0.210981
p(Buch=Buch_B) = 0.260825
p(Buch=Buch_C) = 0.528193

Das Ergebnis empfiehlt Buch C anstatt Buch B. Der Grund hierfür liegt allein in der Struktur des Bayes Netzes. Werden die Daten genauer untersucht, kann man sehr starke Abhängigkeiten zwischen den Root-Knoten Altersgruppe, Geschlecht, Abschluss und dem Knoten Buch feststellen. Weil diese keine direkte Beziehung mit dem Knoten Buch haben und der gegebene Datensatz deutlich zu klein ist, fehlen wichtige Informationen für eine brauchbare Berechnung der Wahrscheinlichkeiten.

Der Code wird für Erweiterung der beiden Kanten angepasst:

```
// Set parent-child relations
nodes[Verheiratet]->parents = { nodes[Altersgruppe] };
nodes[Kinderzahl]->parents = { nodes[Altersgruppe] };
nodes[Beruf]->parents = { nodes[Abschluss], nodes[Geschlecht] };
nodes[Familieneinkommen]->parents = { nodes[Beruf] };
nodes[Buch]->parents = { nodes[Verheiratet], nodes[Kinderzahl],
nodes[Familieneinkommen], nodes[Altersgruppe], nodes[Geschlecht], nodes[Abschluss]};
```

… werden nun die direkten Abhängigkeiten in die Berechnung miteinbezogen.
Die Struktur würde folgendermaßen in Netica aussehen:

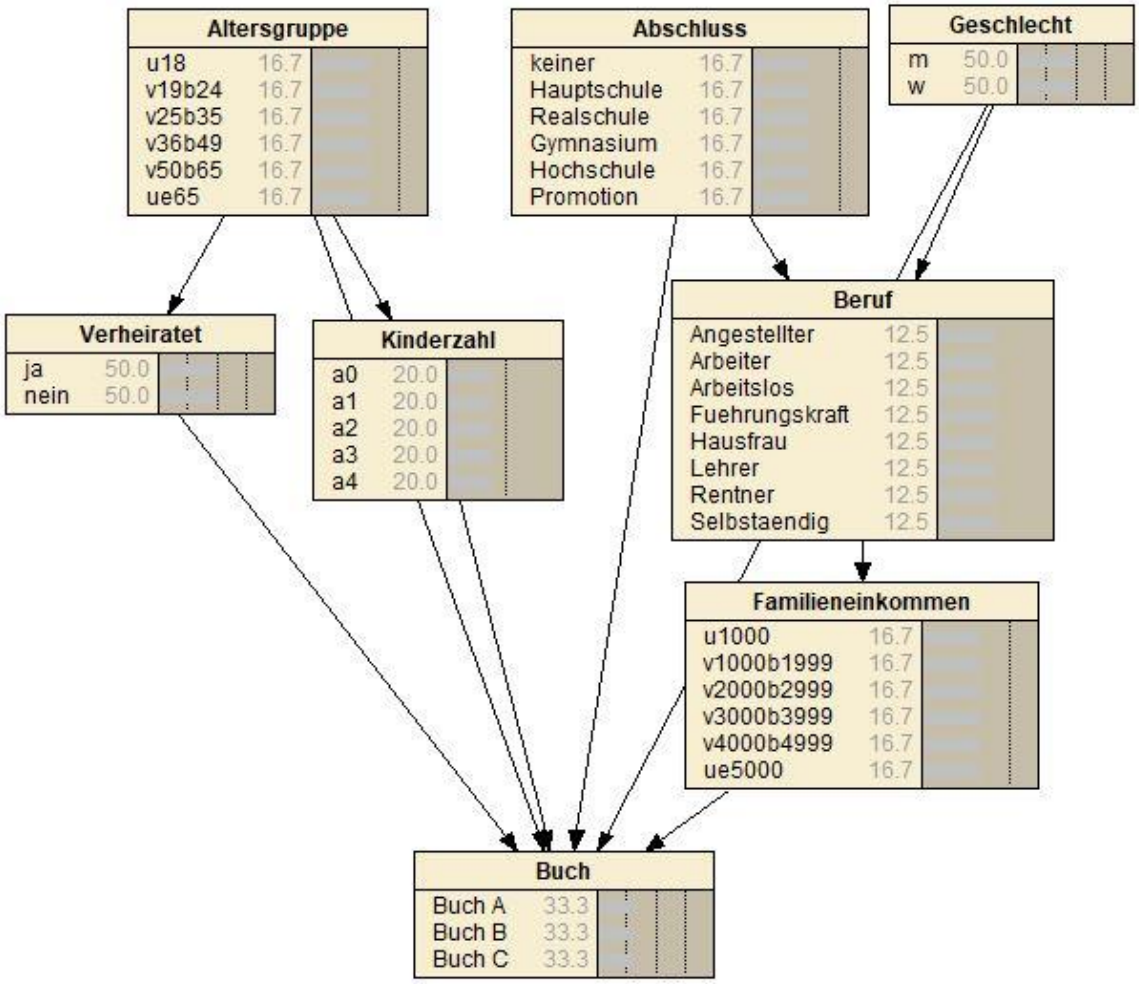

Abb: 3: Netica Buchempfehlung 2

Wird die vorherige Eingabe der Evidenzen für die erste Struktur nun auch für die zweite Struktur gemacht, ist zu erkennen, dass ein plausibles Ergebnis berechnet wird.

p(Altersgruppe=<18) = 0
p(Altersgruppe=19-24) = 1
p(Altersgruppe=25-35) = 0
p(Altersgruppe=36-49) = 0
p(Altersgruppe=50-65) = 0
p(Altersgruppe=>65) = 0
p(Verheiratet=nein) = 1
p(Verheiratet=ja) = 0
p(Kinderzahl=0) = 1
p(Kinderzahl=1) = 0
p(Kinderzahl=2) = 0
p(Kinderzahl=3) = 0
p(Kinderzahl=4) = 0
p(Geschlecht=m) = 1
p(Geschlecht=w) = 0
p(Abschluss=keiner) = 0.39
p(Abschluss=Hauptschule) = 0.06

p(Abschluss=Realschule) = 0.17
p(Abschluss=Gymnasium) = 0.1
p(Abschluss=Hochschule) = 0.25
p(Abschluss=Promotion) = 0.03
p(Beruf=Angestellter) = 0.408214
p(Beruf=Arbeiter) = 0.216071
p(Beruf=Arbeitslos) = 0.144405
p(Beruf=Fuehrungskraft) = 0.0835714
p(Beruf=Hausfrau) = 0
p(Beruf=Lehrer) = 0.0178571
p(Beruf=Rentner) = 0.0505952
p(Beruf=Selbstaendig) = 0.0792857
p(Familieneinkommen=<1000) = 0.0361012
p(Familieneinkommen=1000-1999) = 0.251161
p(Familieneinkommen=2000-2999) = 0.46352
p(Familieneinkommen=3000-3999) = 0.145824
p(Familieneinkommen=4000-4999) = 0
p(Familieneinkommen=5000 und mehr) = 0.103393
p(Buch=Buch_A) = 0.187721
p(Buch=Buch_B) = 0.661107
p(Buch=Buch_C) = 0.151172

Es wird eindeutig Buch B empfohlen mit einer Wahrscheinlichkeit von etwa 66%.

Dadurch, dass nun zwei weitere Kanten ergänzt wurden, also insgesamt 6 Knoten auf Buch zeigen, wird die CPT für Buch exorbitant größer als davor. Dies beeinträchtigt die Performance deutlich, wodurch sich zwei Optionen anbieten:
1. Das Programm könnte durch Parallelisierung effizienter gemacht werden. Vor allem die Berechnung der CPT könnte dabei auf anderen Threads stattfinden, wodurch Multikernsysteme profitieren würden.
2. Die Struktur könnte für den Datensatz geeigneter gewählt werden, um vor allem viele Elternknoten zu vermeiden.
Da eine Änderung der Struktur deutlich einfacher und passender ist, wird Option 2 gewählt.

Die finale Struktur stellt nun die folgende Abbildung dar:

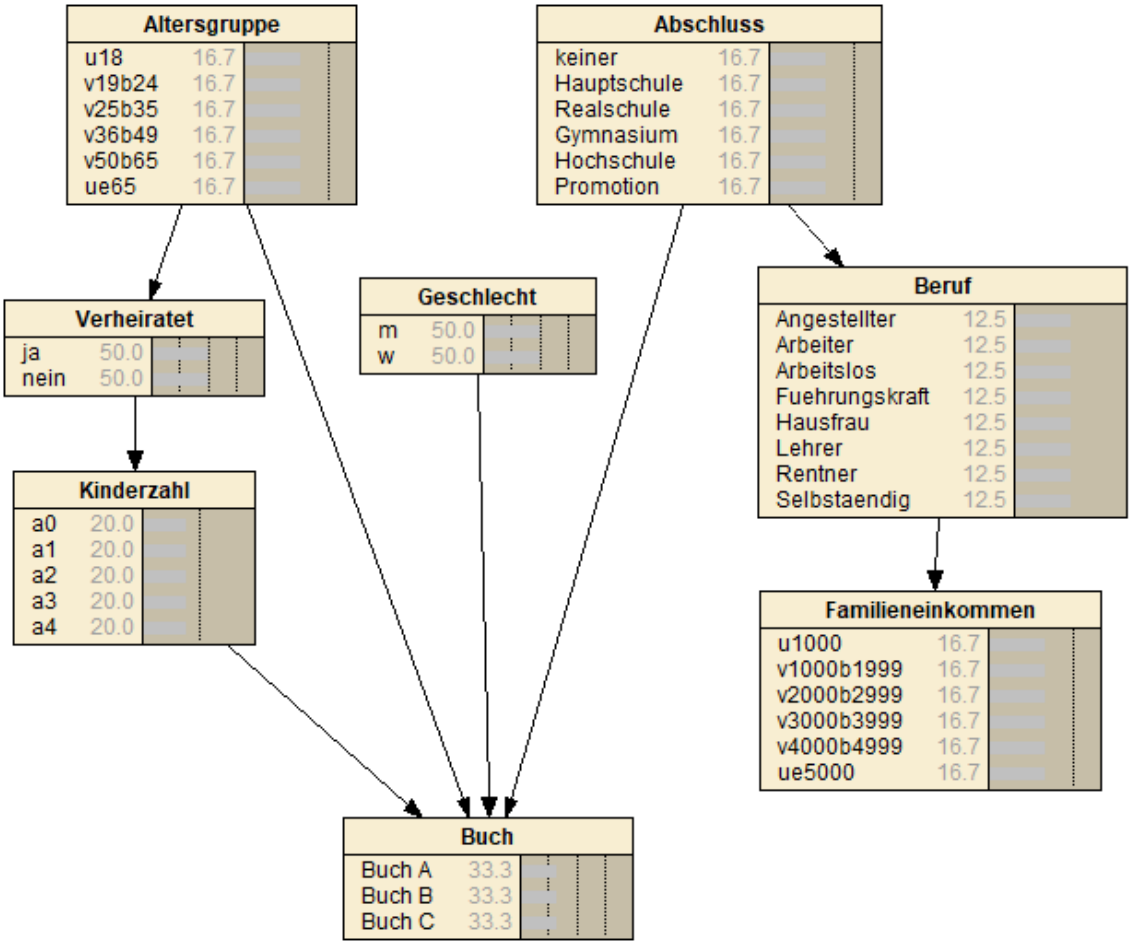

Abb: 3: Netica Buchempfehlung 3

Nur die relevanten Kanten zu Buch werden nun abgebildet. Geschlecht wurde losgelöst von dem Knoten Beruf, weil vergleichsweise für Hausfrau nur wenige Daten vorhanden sind. Des Weiteren wird Kinderzahl in Abhängigkeit von Verheiratet gemacht, da verheiratet eher weniger Einfluss auf die Buchempfehlung hat und Kinder eher bei verheirateten Personen auftreten. Auch Familieneinkommen hat nicht direkt mit dem Buch eine Verbindung, sodass nun nur noch 4 Kanten auf Buch zeigen.

Der Code wird für diese Struktur wieder angepasst:

```
// Set parent-child relations
nodes[Verheiratet]->parents = { nodes[Altersgruppe] };
nodes[Kinderzahl]->parents = { nodes[Verheiratet] };
nodes[Beruf]->parents = { nodes[Abschluss] };
nodes[Familieneinkommen]->parents = { nodes[Beruf] };
nodes[Buch]->parents = { nodes[Kinderzahl], nodes[Altersgruppe], nodes[Geschlecht],
nodes[Abschluss] };
```

Das Ergebnis für die gleiche Eingabe von Evidenzen ist jetzt:

p(Altersgruppe=<18) = 0
p(Altersgruppe=19-24) = 1
p(Altersgruppe=25-35) = 0
p(Altersgruppe=36-49) = 0

p(Altersgruppe=50-65) = 0
p(Altersgruppe=>65) = 0
p(Verheiratet=nein) = 1
p(Verheiratet=ja) = 0
p(Kinderzahl=0) = 1
p(Kinderzahl=1) = 0
p(Kinderzahl=2) = 0
p(Kinderzahl=3) = 0
p(Kinderzahl=4) = 0
p(Geschlecht=m) = 1
p(Geschlecht=w) = 0
p(Abschluss=keiner) = 0.39
p(Abschluss=Hauptschule) = 0.06
p(Abschluss=Realschule) = 0.17
p(Abschluss=Gymnasium) = 0.1
p(Abschluss=Hochschule) = 0.25
p(Abschluss=Promotion) = 0.03
p(Beruf=Angestellter) = 0.41
p(Beruf=Arbeiter) = 0.15
p(Beruf=Arbeitslos) = 0.12
p(Beruf=Fuehrungskraft) = 0.05
p(Beruf=Hausfrau) = 0.13
p(Beruf=Lehrer) = 0.02
p(Beruf=Rentner) = 0.04
p(Beruf=Selbstaendig) = 0.08
p(Familieneinkommen=<1000) = 0.04
p(Familieneinkommen=1000-1999) = 0.22
p(Familieneinkommen=2000-2999) = 0.47
p(Familieneinkommen=3000-3999) = 0.18
p(Familieneinkommen=4000-4999) = 0.02
p(Familieneinkommen=5000 und mehr) = 0.07
p(Buch=Buch_A) = 0.28
p(Buch=Buch_B) = 0.69
p(Buch=Buch_C) = 0.03

Das finale Ergebnis schlägt wieder richtig das Buch B mit 69% vor und ist deutlich schneller berechnet als die vorherigen Strukturen. Die letzte Struktur wird auch im aktuellen Programm verwendet. Bei Bedarf kann die Struktur im Code verändert werden, dann muss die Anwendung aber wieder gebaut werden.

Als letztes soll darauf hingewiesen werden, dass zu wenig Daten ein undefiniertes Ergebnis herbeiführen kann. Werden bestimmte Evidenzen festgelegt und sich diese in den Daten nicht wiederfinden, dann kann kein Buch empfohlen werden und es wird für alle Wahrscheinlichkeiten NaN ausgegeben. Kommt es zu so einem Fall, müssen die Evidenzen allgemeiner definiert oder der Datensatz erweitert werden.