Bayesian Networks for Risk Assessment

Society of Information Risk Analysts
14 November 2014

Norman Fenton
Queen Mary University of London
and
Agena Ltd
Outline

Overview of Bayes and Bayesian networks
Why Bayesian networks are needed for risk assessment
The challenges
Applications
“... although there have been several excellent books dedicated to Bayesian Networks and related methods, these books tend to be aimed at readers who already have a high-level of mathematical sophistication .... As such they are not really accessible to readers who are not already proficient in those subjects. This book is an exciting development because it addresses this problem”.

Judea Pearl, winner 2011 Turing Award for work on AI reasoning
A typical probability problem

What is the probability a person has the disease if they test positive?
Bayes Theorem

Have a prior $P(H)$ ("person has disease")
Now get some evidence $E$ ("test result positive")

We know $P(E|H)$

But we want the posterior $P(H|E)$

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\text{not } H) \cdot P(\text{not } H)}$$

$$P(H|E) = \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.05 \cdot 0.999} = \frac{0.001}{0.5005} \approx 0.02$$
Imagine 100,000 people
Out of whom 10 has the disease
But about 500 of the Remaining 9990 people without the disease test positive
So 10 out of 510 who test positive actually have the disease. That's just under 2%. That’s very different from the 95% assumed by most medics.
An alternative visual explanation

Possible people 100,000

- Have disease 100
  - Test Positive 100 (100%)
  - Test negative 0 (0%)

- Don't have disease 99,900
  - Test Positive 4,995 (5%)
  - Test negative 94,905 (95%)

So 100 out of 5,095 who test positive match actually have the disease, i.e. Under 2%
Bayesian Propagation

Applying Bayes theorem to update all probabilities when new evidence is entered
Intractable even for small BNs
Breakthrough in late 1980s - fast algorithms
Tools implement efficient propagation
A Classic BN: Marginals
Dyspnoea observed

- Visit to Asia?
  - Yes: 1.032%
  - No: 98.968%

- Smoker?
  - Yes: 63.4%
  - No: 36.6%

- Has tuberculosis
  - Yes: 1.885%
  - No: 98.115%

- Has lung cancer
  - Yes: 10.276%
  - No: 89.724%

- Has bronchitis
  - Yes: 83.397%
  - No: 16.603%

- Tuberculosis or cancer
  - Yes: 12.054%
  - No: 87.946%

- Positive X-ray?
  - Yes: 15.21%
  - No: 84.79%

- Dyspnoea?
  - Yes: 100%
  - No: 0%

Scenario: Yes
Also non-smoker
Positive x-ray

- Visit to Asia?
  - yes: 1.944%
  - no: 98.056%

- Smoker?
  - yes: 100%
  - no: Scenario 1: no

- Has tuberculosis
  - yes: 25.563%
  - no: 74.437%

- Has lung cancer
  - yes: 24.579%
  - no: 75.421%

- Has bronchitis
  - yes: 56.521%
  - no: 43.479%

- Tuberculosis or cancer
  - yes: 49.885%
  - no: 50.114%

- Positive X-ray?
  - yes: 100%
  - no: Scenario 1: yes

- Dyspnoea?
  - yes: Scenario 1: yes
  - no: 100%
..but recent visit to Asia
The power of BNs

Explicitly model causal factors
Reason from effect to cause and vice versa
‘Explaining away’
Overturn previous beliefs
Make predictions with incomplete data
Combine diverse types of evidence
Visible auditable reasoning
Why Bayesian networks are needed for risk assessment
Assessing Risk of Road Fatalities: Classic Statistical Approach

- Temperature
  - Colder months
- Number of Fatalities
  - Fewer fatalities
Classic (but wrong) approach to risk

Static factors

Factor 1 — Factor 2 — Factor n-1 — Factor n

Fail (y/n)
What we really need

Static factors

Factor 1  Factor 2  Factor n-1  Factor n

Propensity to fail (y/n)

Intensive support (y/n)

Quit course (y/n)

Fail (y/n)
Risk = Probability * Impact?
Bayesian Net with causal view of risk

- **Trigger:** Meteor on collision course with Earth
- **Risk event:** Meteor strikes Earth
- **Consequence:** Loss of Life
- **Control:** Blow up Meteor
- **Mitigant:** Build Underground cities
The challenges
Bayesian Networks: Barriers and Challenges

Resistance to subjective probabilities
Building realistic models
Handling continuous variables properly
Resistance to Bayes

• OK – but even if I accept the calculations are ‘correct’ I don’t accept subjective priors

There is no such thing as a truly objective frequentist approach
A Real World Bayesian Network
How to build big BNs?
Options for Building BNs

Structure and probability tables all learnt from data only (‘machine learning’)

Structure informed by experts, probability tables learnt from data

Structure and tables built by experts
### Machine Learning Option

<table>
<thead>
<tr>
<th>Visit</th>
<th>Malaria</th>
<th>Tuberculosis</th>
<th>Smoking</th>
<th>Cancer</th>
<th>TB X-ray</th>
<th>XRay</th>
<th>Bronchitis</th>
<th>Dyspnea</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>2</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Abnormal</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>3</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>4</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>5</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>6</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>7</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>8</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>9</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>10</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>11</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>12</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>13</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>14</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>15</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>16</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>17</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>18</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>19</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>20</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>21</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>22</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>23</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>24</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>25</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>26</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>27</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>28</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>29</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>30</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>31</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>32</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>33</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>34</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>35</td>
<td>Absent</td>
<td>Absent</td>
<td>True</td>
<td>Absent</td>
<td>False</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
</tbody>
</table>

### ML Algorithm

- **Asia**
- **TB**
- **Cancer**
- **Bronch**
- **Pos X-ray**
- **Dyspnea**

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Has bronchitis</th>
<th>Has lung cancer</th>
<th>Has tuberculosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>No</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>No</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>No</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
Structure informed by experts

<table>
<thead>
<tr>
<th>Has bronchitis</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has lung cancer</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Has tuberculosis</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Has bronchitis</td>
<td>yes</td>
<td>0.2539334</td>
</tr>
<tr>
<td>Has lung cancer</td>
<td>yes</td>
<td>0.74628682</td>
</tr>
</tbody>
</table>

Asia
Smoker
Bronch
TB
Cancer
Pos X-ray
Dyspnea
Structure and tables by experts

ML Algorithm

<table>
<thead>
<tr>
<th>Has bronchitis</th>
<th>yes</th>
<th>no</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuberculosis...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>0.875</td>
<td>0.7619048</td>
<td>0.68085104</td>
<td>0.1122449</td>
</tr>
<tr>
<td>no</td>
<td>0.125</td>
<td>0.23809524</td>
<td>0.31914896</td>
<td>0.8877551</td>
</tr>
</tbody>
</table>
Handling continuous nodes

Static discretisation: inefficient and devastatingly inaccurate

Developments in dynamic discretisation will have revolutionary effect
Dynamic discretization
Typical Applications

Predicting reliability of critical systems
Typical Applications

Software defect prediction
Typical Applications

Aircraft accident traffic risk
Typical Applications

Warranty return rates of electronic parts
Typical Applications

Operational risk in financial institutions
Typical Applications

Hazards in petrochemical industry
Typical Applications

- Predicting reliability of critical systems
- Software defect prediction
- Aircraft accident traffic risk
- Warranty return rates of electronic parts
- Operational risk in financial institutions

Probabilistic and risk based legal arguments

R vs Levi Bellfield
Conclusions

Genuine risk assessment requires causal Bayesian networks

Bayesian networks have been used effectively in a range of real world problems.

Major remaining barrier to widespread use is conceptual/presentational
“... although there have been several excellent books dedicated to Bayesian Networks and related methods, these books tend to be aimed at readers who already have a high-level of mathematical sophistication .... As such they are not really accessible to readers who are not already proficient in those subjects. This book is an exciting development because it addresses this problem”.

Judea Pearl, winner 2011 Turing Award for work on AI reasoning