When Data mining brings its force to modelling!

Dalila HAMAMI
Prof. Carron SHANKLAND
Prof. Baghdad ATMANI

21/04/2015
PhD Student Day
Computing Science and Mathematics
Outline

- Our approach
- Tuberculosis modelling
- Tuberculosis data mining
- Optimisation
- Conclusion
Data mining brings its force to modelling!

Optimising simulated model

Combining data mining with modelling
Our approach

• Modelling and simulation
• Optimising model
  ➢ Reviewing existing knowledge
  ➢ using data mining approach
Our approach

A: Modelling and Simulation phase

1: Problem definition
2: Design the study
3: Design the conceptual model
4: Process definition: Inputs, Outputs, Rules Assumptions
5: Developing Bio-PEPA model
6: Simulation and analysis

B: Optimizing the model

1: Data cleaning and transformation
2: Features selection using a range of algorithms
3: Classification using a range of algorithms
4: Set of rules or pertinent attributes

C: Data Mining phase

Resulting model

Database

Expert knowledge
Our approach

Modelling and Simulation phase

1: Problem definition
2: Design the study
3: Design the conceptual model
4: Process definition: Inputs, Outputs, Rules Assumptions
5: Developing Bio-PEPA model
6: Simulation and analysis

Data Mining phase

1: Data cleaning and transformation
2: Features selection using a range of algorithms
3: Classification using a range of algorithms
4: Set of rules or pertinent attributes

Resulting model

Optimizing the model

Expert knowledge
Modelling and simulation phase

Epidemiological compartmental model
Tuberculosis model

Susceptible (S) → Infected TP (I) → Treated TP (R)

Susceptible (S) → Recovered

Infected TP (I) → Died
Infected TP (I) → Lost
Infected TP (I) → Failed
Infected TP (I) → Transferred

Recovered → Diagnosed infected TP
Died
Lost
Failed
Transferred
Bio-PEPA Tuberculosis model for 2009
Bio-PEPA tuberculosis model for 2010

![Graph showing data comparison between simulated and observed data for treated cases.](image)
How do we identify model validity?

![Graph showing the number of lost simulated data and observed data over the years 2009 to 2012. The graph compares simulated data (blue line) and observed data (red line).](image-url)
How do we identify validity?
Why data mining?

- Can’t find the data needed
Why data mining?

• Can’t find the data needed

• Can’t get the data needed
Why data mining?

- Can’t find the data needed
- Can’t get the data needed
- Can’t understand/explain the data
Data mining

1: Data cleaning and transformation

2: Features selection using a range of algorithms

3: Classification, Association rules, Clustering… etc

Resulting model

4: Set of rules or pertinent attributes
Database

1: Data cleaning and transformation

2: Features selection using a range of algorithms

3: Classification, Association rules, Clustering...etc

Resulting model

4: Set of rules or pertinent attributes

C: Data Mining phase
## Data mining/data

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Data mining
Decision tree induction

```
IF (Outlook = Sunny) \land (Humidity = High)
THEN PlayTennis = No

IF (Outlook = Sunny) \land (Humidity = Normal)
THEN PlayTennis = Yes
```
# Tuberculosis dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>Trimester</th>
<th>Year</th>
<th>Address 2</th>
<th>Address 1</th>
<th>Type</th>
<th>Smear 1</th>
<th>Smear 2</th>
<th>Smear 3</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>F</td>
<td>1</td>
<td>2009</td>
<td>C11</td>
<td>M1</td>
<td>N</td>
<td>P+</td>
<td>NA</td>
<td>NA</td>
<td>LOST</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>F</td>
<td>2</td>
<td>2011</td>
<td>C22</td>
<td>M2</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>RECOVERED</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>M</td>
<td>1</td>
<td>2012</td>
<td>C13</td>
<td>M1</td>
<td>R</td>
<td>P+</td>
<td>P</td>
<td>P+</td>
<td>TRANSFERED</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>F</td>
<td>4</td>
<td>2010</td>
<td>C14</td>
<td>M1</td>
<td>F</td>
<td>P</td>
<td>NA</td>
<td>P</td>
<td>FAILED</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>200</td>
<td>54</td>
<td>M</td>
<td>3</td>
<td>2011</td>
<td>C22</td>
<td>M2</td>
<td>N</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>LOST</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Tuberculosis decision tree
Tuberculosis decision tree

ADDRESS 1

- =M1
  - Failed

- =M2
  - Lost

SMEAR 1

- =P
  - Died
  - Recovered

- =P+
  - Transferred

Recovered

SMEAR 2

- =P+
  - Transferred

- =NA
  - Recovered

Recovered

SMEAR 3

- =P
  - Failed

- =N
  - Recovered

Failed

Recovered

Lost
Optimising Bio-PEPA model
Optimising Bio-PEPA model

Tuberculosis model 2010

-M1-

Simulated data

- 3
- 2
- 4
- 3

Observed data

- 27
- 5
- 3
- 3
- 0

-M2-

Simulated data

- 0
- 3
- 18
- 5

Observed data

- 0
- 4
- 16
- 1

Legend:
- Died
- Failed
- Recovered
- Lost
- Transferred
Optimising Bio-PEPA model

M1 - Tuberculosis model 2012

M2 -
CONCLUSION

USING DATA MINING HELPS:

• Expert to understand where he should investigate more research to understand the epidemic process.
• Developer to optimise the initial simulated model
What is going on?

How other data mining techniques can help?
Thanks