What is Data Mining?

- Domain understanding
- Data selection
- Data cleaning, e.g., data duplication, missing data
- Preprocessing, e.g., integration of different files
- Pattern (knowledge) discovery
- Interpretation (e.g., visualization)
- Reporting

Data Mining “Architecture”

Illustrative Applications

- Prediction of equipment faults
- Determining a stock level
- Process control
- Fraud detection
- Genetics
- Disease staging and diagnosis
- Decision making

Pharmaceutical Industry

- Selection of “Patient suitable” medication
  - Adverse drug effects minimized
  - Drug effectiveness maximized
  - New markets for “seemingly ineffective” drugs
- “Medication bundle”
  - Life-time treatments
- Design and virtual testing of new drugs
What is Knowledge Discovery?

Data Mining is Not

- Data warehousing
- SQL / Ad hoc queries / reporting
- Software agents
- Online Analytical Processing (OLAP)
- Data visualization

Learning Systems (1/2)
- Classical statistical methods (e.g., discriminant analysis)
- Modern statistical techniques (e.g., k-nearest neighbor, Bayes theorem)
- Neural networks
- Support vector machines
- Decision tree algorithms
- Decision rule algorithms
- Learning classifier systems

Learning Systems (2/2)
- Association rule algorithms
- Text mining algorithms
- Meta-learning algorithms
- Inductive learning programming
- Sequence learning

Regression Models
- Simple linear regression = Linear combination of inputs
- Logistic regression = Logistic function of a linear combination of inputs
  - Classic “perceptron”

Neural Networks
- Based on biology
- Inputs transformed via a network of simple processors
- Processor combines (weighted) inputs and produces an output value
- Obvious questions: What transformation function do you use and how are the weights determined?
Neural Networks

- Feed-forward - Regression analogy
- Multi-layer NN - Nonlinear regression analogy

Types of Decision Trees

- CHAID: Chi-Square Automatic Interaction Detection
  - Kass (1980)
  - n-way splits
  - Categorical variables
- CART: Classification and Regression Trees
  - Binary splits
  - Continuous variables
- C4.5
  - Quinlan (1993)
  - Also used for rule induction

Text Mining

- Mining unstructured data (free-form text) is a challenge for data mining
- Usual solution is to impose structure on the data and then process using standard techniques, e.g.,
  - Simple heuristics (e.g., unusual words)
  - Domain expertise
  - Linguistic analysis
- Presentation is critical

Yet Another Classification

- Supervised
  - Regression models
  - k-Nearest-Neighbor
  - Neural networks
  - Rule induction
  - Decision trees
- Unsupervised
  - k-means clustering
  - Self organized maps

Supervised Learning Algorithms

- kNN
  - Quick and easy
  - Models tend to be very large
- Neural Networks
  - Difficult to interpret
  - Training can be time consuming
- Rule Induction
  - Understandable
  - Need to limit calculations
- Decision Trees
  - Understandable
  - Relatively fast
  - Easy to translate into SQL queries

Knowledge Representation Forms

- Decision rules
- Trees (graphs)
- Patterns (matrices)
## Drug Adverse Reaction Case Study

### Training data set

<table>
<thead>
<tr>
<th>Patient</th>
<th>Patient datum 1</th>
<th>Test_1</th>
<th>Patient datum 2</th>
<th>Test_2</th>
<th>Outcome</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.02</td>
<td>Red</td>
<td>2.98</td>
<td>High</td>
<td>No_Adv_Reaction</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.03</td>
<td>Black</td>
<td>1.04</td>
<td>Low</td>
<td>Adverse_Reaction</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>Blue</td>
<td>3.04</td>
<td>High</td>
<td>No_Adv_Reaction</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.02</td>
<td>Blue</td>
<td>3.11</td>
<td>High</td>
<td>No_Adv_Reaction</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
<td>Orange</td>
<td>0.96</td>
<td>Low</td>
<td>Adverse_Reaction</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.04</td>
<td>Blue</td>
<td>1.04</td>
<td>Medium</td>
<td>Adverse_Reaction</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.99</td>
<td>Orange</td>
<td>1.04</td>
<td>Medium</td>
<td>No_Adv_Reaction</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.02</td>
<td>Red</td>
<td>0.94</td>
<td>Low</td>
<td>Adverse_Reaction</td>
<td></td>
</tr>
</tbody>
</table>

### Decision rules

- **Rule 1.** IF (Patient_datum_1 < 0.515) THEN (D = Adverse_Reaction); [2, 2, 50.00%, 100.00%][2, 0][5, 6]
- **Rule 2.** IF (Test_2 = Low) THEN (D = Adverse_Reaction); [3, 3, 75.00%, 100.00%][3, 0][2, 5, 8]
- **Rule 3.** IF (Patient_datum_2 >= 2.01) THEN (D = No_Adv_Reaction); [3, 3, 75.00%, 100.00%][0, 3][1, 3, 4]
- **Rule 4.** IF (Patient_datum_1 >= 0.515) & (Test_1 = Orange) THEN (D = No_Adv_Reaction); [1, 1, 25.00%, 100.00%][0, 1][7]

### Decision Rule Metrics

<table>
<thead>
<tr>
<th>Rule 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF (Flow = 6) AND (Pressure = 7) THEN (Efficiency = 81); [13, 8, 4.19%, 61.54%][1, 8, 4]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Support</th>
<th>Strength</th>
<th>Relative strength</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ 524 ], { 527, 528, 529, 530, 531, 533, 535, 536 }, { 525, 526, 532, 534 } - Supporting objects</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Definitions

- **Support** = Number of objects satisfying conditions of the rule
- **Strength** = Number of objects satisfying conditions and the decision of the rule
- **Relative strength** = Number of objects satisfying conditions and decision of the rule/The number of objects in the class
- **Confidence** = Strength/Support

### Classification Accuracy

<table>
<thead>
<tr>
<th>Test: Leaving-one-out Conflict Matrix</th>
<th>ADVERSE_R</th>
<th>NO_ADV_R</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADVERSE_R</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NO_ADV_R</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average Accuracy [%]</th>
<th>Correct</th>
<th>Incorrect</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>75.00</td>
<td>25.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ADVERSE_R</td>
<td>75.00</td>
<td>25.00</td>
<td>0.00</td>
</tr>
<tr>
<td>NO_ADV_R</td>
<td>75.00</td>
<td>25.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### Decision rules

- **Rule 113**
  - IF (B_Master >= 1634.26) AND (B_Temp in (1601.2, 1660.22)) AND (B_Pressure in [17.05, 18.45]) AND (A_point = 0.255) AND (Average_O2 = 77) THEN (Eff = 87) OR (Eff = 88); [6, 6, 23.08%, 100.00%][0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0] [[2164, 2167, 2168], [2163, 2165, 2166]]
Decision rules

Rule 12
IF (Ave_Middle_Bed = 0) AND (PA_Fan_Flow = 18) THEN (Efficiency = 71);
[16, 10, 10.31%, 62.50%] [1, 2, 10, 2]
[682], [681], [933, 936],
{875, 876, 877, 878, 879, 880, 934, 935, 1000, 1001],
{881, 882}

Decision Rule vs Decision Tree

Decision Tree

F1 F2 F3 F4 D
0 0 1 1 Two
0 1 1 1 Three
1 1 1 1 Four

Decision Tree vs Rule Tree

Decision Tree

Rule 1. (F3 = 0) THEN (D = One);
[1, 100.00%, 100.00%][1]
Rule 2. (F2 = 0) AND (F3 = 1) THEN (D = Two);
[1, 100.00%, 100.00%][2]
Rule 3. (F1 = 0) AND (F2 = 1) THEN (D = Three);
[1, 100.00%, 100.00%][3]
Rule 4. (F1 = 1) THEN (D = Four);
[1, 100.00%, 100.00%][4]

F1 F2 F3 F4 D
0 0 1 1 Two
0 1 1 1 Three
1 1 1 1 Four

Rule 1. (F3 = 0) THEN (D = One);
[1, 100.00%, 100.00%][1]
Rule 2. (F2 = 0) AND (F3 = 1) THEN (D = Two);
[1, 100.00%, 100.00%][2]
Rule 3. (F1 = 0) AND (F2 = 1) THEN (D = Three);
[1, 100.00%, 100.00%][3]
Rule 4. (F1 = 1) THEN (D = Four);
[1, 100.00%, 100.00%][4]
Decision Rule Algorithms

- Identify unique features of an object rather than commonality among all objects

Use of Extracted Knowledge

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>One</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>Two</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Three</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Four</td>
</tr>
</tbody>
</table>

Traditional Modeling

- Regression analysis
- Neural network

Data Mining

- Rules
- Decision trees
- Patterns

Evolution in Data Mining

- Data Farming
- Cultivating data rather than assuming that it is available

Data Life Cycle

- Data farming
- Result evaluation
- Knowledge extraction
- Decision-making
Data Farming

Pull data approach

VS

Push data approach in classical data mining

Data Farming

New Science!

Define features that

• Maximize classification accuracy and

• Minimize the data collection cost

Data Mining Standards

• Predictive Model Markup Language (PMML)
  - The Data Mining Group (www.dmg.org)
  - XML based (DTD)
• Java Data Mining API spec request (JSR-000073)
  - Oracle, Sun, IBM, …
  - Support for data mining APIs on J2EE platforms
  - Build, manage, and score models programmatically
• OLE DB for Data Mining
  - Microsoft
  - Table based
  - Incorporates PMML

Summary

• Data mining algorithms support a new paradigm:
  Identify what is unique about an object
• DM tools to enter new areas of information analysis

References (1/2)


References (2/2)
