Overview

- Motivation
- Data mining models
- Visualization based on sensitivity analysis
- Regression problems
- Classification problems
- Definition of interesting plots
- Genetic search for 2D and 3D plots
Motivation

• Data mining – extracting new, potentially useful information from data
• DM Models are automatically generated
• Are models always credible?
• Are models comprehensible?
• How to extract information from models?

Visualization
Data mining models

- Often black-box models generated from data
- E.g. Neural networks
- What is inside?
Inductive model

- Estimates output from inputs
- Generated automatically
- Evolved by niching GA
- Grows from minimal form
- Contains hybrid units
- Several training methods
- Ensemble of models
Example: Housing data

Input variables

CRIM  ZN  INDUS  NOX  RM  AGE  DIS  RAD  TAX  PTRATIO  B  LSTA

- Per capita crime rate by town
- Proportion of owner-occupied units built prior to 1940
- Weighted distances to five Boston employment centres
- Median value of owner-occupied homes in $1000's

Output variable

MEDV
## Housing data – records

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PTRATIO</th>
<th>B</th>
<th>LSTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>0.00632</td>
<td>18</td>
<td>2.31</td>
<td>53.8</td>
<td>6.575</td>
<td>65.2</td>
<td>4.09</td>
<td>1</td>
<td>296</td>
<td>15.3</td>
<td>396.9</td>
</tr>
<tr>
<td>21.6</td>
<td>0.02731</td>
<td>0</td>
<td>7.07</td>
<td>46.9</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2</td>
<td>242</td>
<td>17.8</td>
<td>396.9</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>

4.98
9.14
MEDV

**Input variables**

**Output variable**
Housing data – inductive model

- Input variables:
  - CRIM
  - ZN
  - INDUS
  - NOX
  - RM
  - AGE
  - DIS
  - RAD
  - TAX
  - PTRATIO
  - B
  - LSTA

- Niching genetic algorithm evolves units in first layer
- Sigmoid function:
  - \( MEDV = \frac{1}{1 + \exp(-5.724 \times CRIM + 1.126)} \)
  - \( MEDV = \frac{1}{1 + \exp(-5.861 \times AGE + 2.111)} \)

- Output variable:
  - Error: 0.13
  - Error: 0.21
Housing data – inductive model

Input variables

CRIM  ZN  INDUS  NOX  RM  AGE  DIS  RAD  TAX  PTRATIO  B  LSTA

sigmoid

Error: 0.13

polynomial

sigmoid

Error: 0.21

sigmoid

Error: 0.24

linear

Error: 0.26

MEDV = 0.747*(1/(1-exp(-5.724*CRIM + 1.126)))
+ 0.582*(1/(1-exp(-5.861*AGE + 2.111)))^2 + 0.016

Niching genetic algorithm evolves units in second layer

Error: 0.10
Housing data – inductive model

Input variables

CRIM  ZN  INDUS  NOX  RM  AGE  DIS  RAD  TAX  PTRATIO  B  LSTA

sigmoid  sigmoid  sigmoid  linear  polynomial  polynomial  exponential

Error: 0.08

Output variable

MEDV

Constructed model has very low validation error!
### Housing data – inductive model

**Input variables**

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PTRATIO</th>
<th>B</th>
<th>LSTA</th>
</tr>
</thead>
</table>

\[
\text{MEDV} = \left( \exp\left( (0.038 \times 3.451 \times (1/(1-\exp(-5.724 \times \text{CRIM} + 1.126))) \times (1/(1-\exp(2.413 \times \text{DIS} - 2.581))) \times (1/(1-\exp(2.413 \times \text{DIS} - 2.581))) + 0.429 \times (1/(1-\exp(-5.861 \times \text{AGE} + 2.111))) + 0.024 \times (1/(1-\exp(2.413 \times \text{DIS} - 2.581))) + 0.036 + 0.038 \times 0.350 \times (1/(1-\exp(-3.613 \times \text{RAD} - 0.088))) + 0.999 \times (0.747 \times (1/(1-\exp(-5.724 \times \text{CRIM} + 1.126))) + 0.582 \times (1/(1-\exp(-5.861 \times \text{AGE} + 2.111))) \times (1/(1-\exp(-5.861 \times \text{AGE} + 2.111))) + 0.016) - 0.046 \times (1/(1-\exp(-5.724 \times \text{CRIM} + 1.126))) - 0.079 + 0.002 \times \text{INDUS} - 0.001 \times \text{LSTA} + 0.150) \times 0.860 \right) \times 13.072 - 14.874
\]

Math equation is not comprehensible any more – we have to treat it as a **black box model**!
Visualization based on sensitivity analysis

constant

moving

constant

\[ y_{k} \]

\[ x_{1} \quad x_{3} = \text{const} \]

GAME

min

\[ x_{2} \]

max
Sensitivity analysis of inductive model of MEDV

What will happen with the value of house when criminality in the area decreases/increases?

Credible output?
Ensemble of inductive models

- Random initialization
- Developing on the same training set
- Training affect just well defined areas of input space
- Each model - unique architecture, similar complexity, similar transfer functions
- Similar behavior for well defined areas
- Different behavior – under-defined areas
Credibility of models: Artificial data set

Advantages:

- No need of the training data set,
- Modeling method success considered,
- Inputs importance considered.

Credibility: the criterion is a dispersion of models’ responses.
Example: Models of hot water consumption
Cold water consumption, increasing humidity
Models on Housing data

- Single model
- Ensemble of 10 models
Classification problems

- Data:
  - Setoza class
  - Virginica class
  - Versicolor class

- Blue GAME model
  (Iris Setoza class)
  - output = 1
  - decision boundary
  - output = 0
Credibility of classifiers

GAME model 1  GAME model 2  GAME model 3  GAME models (1*2*3)

Iris Setosa

petal_width  petal_length

Iris Virginica

petal_width  petal_length

Iris Versicolor

petal_width  petal_length

petal_width  petal_length

petal_width  petal_length
Overlapping models X ensemble
Random behavior filtered out

Before

After
Problem – how to find information in n-dim space?

- Multidimensional space of input variables
- What we are looking for?
  - Interesting relationship of IO variables
  - Regions of high sensitivity
  - Credible models (compromise response)
- Can we automate the search?
When a plot is interesting for us?
Definition of interesting plot

- Minimal volume of the envelope \( p \Rightarrow \min \)

\[
p = \sum_{j=x_{\text{start}}}^{x_{\text{start}}+x_{\text{size}}} \left( \arg \max_{0<i\leq m} \left( y_i(j) \right) - \arg \min_{0<i\leq m} \left( y_i(j) \right) \right)
\]

- Maximal sensitivity of the output to the change of \( x_i \) input variable – \( y_{\text{size}} \Rightarrow \max \)

\[
y_{\text{size}} = \arg \max (\bar{y}(t)) - \arg \min (\bar{y}(t)), t \in (x_{\text{start}}, x_{\text{size}})
\]

- Maximal size of the area – \( x_{\text{isize}} \Rightarrow \max \)
Multiobjective optimization

- Interestingness:

\[ \text{fitness} = y_{\text{size}} \cdot \frac{1}{p} \cdot x_{\text{size}} \]

- Unknown variables:
  \[ x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, x_{\text{istart}}, x_{\text{isize}} \]

- We will use “Niching” genetic algorithm

Chromosome:

\[
\begin{array}{cccccccc}
  x_1 & x_2 & \ldots & x_{i-1} & x_{i+1} & \ldots & x_n & x_{\text{istart}} & x_{\text{isize}} \\
\end{array}
\]
Niching GA on simple data

**Chromosome:**

\[ x_{\text{start}} \quad x_{\text{size}} \]

**Search space is 2D, can be visualized**

Very simple problem

**Fitness function:**

\[ \text{fitness} = x_{\text{size}} \times \frac{1}{p} \times y_{\text{size}} \]
Niching GA locates also local optima

- Three subpopulations (niches) of individuals survived
Automated retrieval of plots showing interesting behavior

Genetic algorithm with special fitness function is used to adjust all other inputs (dimensions)

Best so far individual found (generation 0 – 17)
Housing data – interesting plot retrieved

Low fitness

High fitness

Before

After
Conclusion

- Credible regression
- Credible classification
- Automated retrieval
Future work I

- Automated knowledge extraction from data
Future work II

• FAKE GAME framework
Future work III

• Just released as open source project
  – Automated data preprocessing
  – Automated model building, validation
  – Optimization methods
  – Visualization

see and join us:
http://www.sourceforge.net/projects/fakegame