A Weight Analysis-based Wrapper Approach to Neural Nets Feature Subset Selection

Dietrich Hans Schuschel
Longbow Apache Software
Boeing Company, AZ, USA

Chun-Nan Hsu 許鈞南
Institute of Information Science
Academia Sinica, Taipei, TAIWAN

Prepared for the presentation in the 10th IEEE International Conference on Tools with Artificial Intelligence, November 11, 1998

Overview

• The need for feature selection
• The wrapper model of feature selection
• The ANNIGMA heuristic
• Integrating ANNIGMA into the wrapper
• Comparing ANNIGMA-wrapper with previous work
• Application in the pulse-echo signal classification domain
• Application in the helicopter blade strain prediction domain
• Conclusion
Feature Subset Selection

• Features (or attributes) affect performance and training time
  – relevance
  – noise
  – overfitting
• The need for feature subset selection
  – In general neural net-based application development
  – In novel and advanced applications of neural nets
  – In assisting human understanding of data relationships

The Wrapper Model of Feature Subset Selection

• From previous work by (John, Kohavi, Pfleger, 1994), (Caruana, Freitag, 1994), and (Kohavi, Sommerfield, 1995)
  – Strong and weak attribute relevance
  – The detrimental effect of excess attributes
  – Prediction accuracy as the attribute subset metric
  – The optimal attribute subset
• The wrapper model addresses the above issues
The Filter Model Vs. The Wrapper Model

- The wrapper model is a framework for attribute subset selection
  - The induction algorithm is used to evaluate the attribute subset
  - The wrapper searches for the optimal attribute subset

The Feature Subset Search and The ANNIGMA Heuristic

- Previous work in wrapper model
  - Use AI techniques like backward elimination, forward selection, etc.
  - Accelerated through compound operators, bi-directional search, or randomized algorithm
- The ANNIGMA heuristic is a powerful feature subset search heuristic for neural nets
  - Ranks the attributes by relevance to the neural net’s output
  - Provides direction to the optimal attribute subset search
  - Reduces the number of training cycles of the wrapper model from $O(2^m)$ to $O(m)$ for $m$ attributes
The ANNIGMA Heuristic

• Artificial Neural Net Gain Measurement Approximation
  – Normalize the inputs
  – Train the neural net
  – Examine the weights of the trained net
• Calculation:
  \[ \text{ANNIGMA}_k = \frac{LG_{ik}}{\max(LG_k)} \times 100 \] \hspace{2cm} (1)
  \[ LG_{ik} = \frac{\Delta O_i}{\Delta A_k} \] \hspace{2cm} (2)
  \[ LG_a = \sum_{j} W_j \times W_k \] \hspace{2cm} (3)

Integrating ANNIGMA into the Wrapper Algorithm
Integrating ANNIGMA into the Wrapper Algorithm

- Multiple neural nets are trained in each cycle
  - Perform cross-validation
  - Mitigate neural net training variability
- Use minimal training parameters
  - Avoid overtraining problems
  - Accelerate the wrapper cycles
- There is no need to test attribute subsets
  - The ANNIGMA heuristic ranks the attribute’s relevance
- The ANNIGMA heuristic is weighted by neural net error

Comparing ANNIGMA-wrapper with previous work

- (Setiono, Liu, 1997), SL-NNFS Neural Network Feature Selector
- (Yang, Hanovar, 1997), YH-DistAL Genetic neural net generation
- (John, Kohavi, Pfleger 1994), JKP, using C4.5 decision tree induction algorithm (Quinlan),
  - FW - Forward Selection,
  - BW - Backwards Selection
- (Liu, Setiono, 1996a), LSa - LVF Las Vegas Filter
- (Liu, Setiono, 1996b), LSb - LVW Las Vegas Wrapper
- Neural net results provide a baseline comparison for neural nets without ANNIGMA-wrapper
Standard Artificial Datasets -
CorrAL

–Dataset has 6 attributes, 1 correlated 75%, 1 irrelevant, and 4 in a
  target binary combining function.
–ANNIGMA-wrapper used tangent-sigmoidal transfer function with 10
  hidden nodes and 8 training cycles per wrapper cycle.

<table>
<thead>
<tr>
<th>Paper/method</th>
<th>Attributes (STD)</th>
<th>Error rate (STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNIGMA - wrapper</td>
<td>4 (0.2)</td>
<td>0 (0.2)</td>
</tr>
<tr>
<td>JKP - C4.5 - FW</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>JKP - C4.5 – BW</td>
<td>5</td>
<td>W</td>
</tr>
<tr>
<td>L.Sa – C4.5 – LVF</td>
<td>4</td>
<td>W</td>
</tr>
<tr>
<td>L.Sh – C4.5 – LVW</td>
<td>4</td>
<td>6.2</td>
</tr>
<tr>
<td>Neural net</td>
<td>6</td>
<td>W</td>
</tr>
</tbody>
</table>

Standard Artificial Datasets -
Monk3 (W/O preprocessing)

–Dataset has 6 attributes, target concept needs 3 attributes,
  (a5 = 3 and a4 = 1) or (a5 ≠ 4 and a2 ≠ 3)
–5% class noise in training set. 2.8% (12 / 432) of total records noisy.
–ANNIGMA-wrapper used tangent-sigmoidal transfer function with 8
  hidden nodes and 15 training cycles per wrapper cycle.

<table>
<thead>
<tr>
<th>Paper/method</th>
<th>Attributes (STD)</th>
<th>Error rate (STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNIGMA - wrapper</td>
<td>2 3 (0.7)</td>
<td>2.9 (0.8)</td>
</tr>
<tr>
<td>JKP - C4.5 - FW</td>
<td>3</td>
<td>W</td>
</tr>
<tr>
<td>JKP - C4.5 – BW</td>
<td>2</td>
<td>I</td>
</tr>
<tr>
<td>L.Sa – C4.5 – LVF</td>
<td>3</td>
<td>W</td>
</tr>
<tr>
<td>L.Sh – C4.5 – LVW</td>
<td>3</td>
<td>W</td>
</tr>
<tr>
<td>Neural net</td>
<td>6</td>
<td>W</td>
</tr>
</tbody>
</table>
Standard Artificial Datasets - Monk3 (With preprocessing)

– Dataset has 15 attributes, target concept is MONK3 preprocessed into binary attributes. 5% class noise in training set.
– ANNIGMA-wrapper used linear transfer function with 3 hidden nodes and 8 training cycles per wrapper cycle.

<table>
<thead>
<tr>
<th>Paper/method</th>
<th>Attributes (STD)</th>
<th>Error rate (STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNIGMA - wrapper</td>
<td>2.2 (0.4)</td>
<td>2.8 (0)</td>
</tr>
<tr>
<td>SL – NNFS</td>
<td>3.9 (1.8) W</td>
<td>1.6 (1.7) I2.7</td>
</tr>
<tr>
<td>Neural net</td>
<td>15 W</td>
<td>2.8 (0) -</td>
</tr>
</tbody>
</table>

Standard Real-World Datasets - Vote

– Dataset has 16 attributes, target concept is predicting affiliation for congressmen based on voting records.
– ANNIGMA-wrapper used linear transfer function with 3 hidden nodes and 8 training cycles per wrapper cycle.

<table>
<thead>
<tr>
<th>Paper/method</th>
<th>Attributes (STD)</th>
<th>Error rate (STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNIGMA - wrapper</td>
<td>3.3 (1.9) L</td>
<td>3.1 (0.2) L</td>
</tr>
<tr>
<td>SL – NNFS</td>
<td>2 (0.2) W</td>
<td>3.2 (1.6) W</td>
</tr>
<tr>
<td>YH – DistAL</td>
<td>8.9 (1.8) W</td>
<td>1.2 (1.2) L</td>
</tr>
<tr>
<td>JKP - C4.5 - FW</td>
<td>3 L</td>
<td>-</td>
</tr>
<tr>
<td>JKP - C4.5 – BW</td>
<td>15 W</td>
<td>3 L</td>
</tr>
<tr>
<td>LSh – C4.5 – LVF</td>
<td>8 W</td>
<td>3.5 W</td>
</tr>
<tr>
<td>LSh – C4.5 – LVW</td>
<td>3 W</td>
<td>4 W</td>
</tr>
<tr>
<td>Neural net</td>
<td>16 W</td>
<td>3.2 (0) W</td>
</tr>
</tbody>
</table>
Standard Real-World Datasets - Credit

– Dataset has 15 attributes, target concept is predicting approval for credit based on credit application data.
– ANNIGMA-wrapper used tangent-sigmoidal transfer function with 10 hidden nodes and 10 training cycles per wrapper cycle.

<table>
<thead>
<tr>
<th>Paper/method</th>
<th>Attributes (STD)</th>
<th>Error rate (STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNIGMA - wrapper</td>
<td>6.8 (2.8)</td>
<td>12 (1.1)</td>
</tr>
<tr>
<td>YH – DistAL</td>
<td>8 (2.1)</td>
<td>W 8.5 (2.8)</td>
</tr>
<tr>
<td>JKP - C4.5 – FW</td>
<td>3</td>
<td>L 19</td>
</tr>
<tr>
<td>JKP - C4.5 – BW</td>
<td>14</td>
<td>W 21</td>
</tr>
<tr>
<td>LSa – C4.5 –LVF</td>
<td>5</td>
<td>- 15.2</td>
</tr>
<tr>
<td>LSb – C4.5 –LVW</td>
<td>6</td>
<td>- 15</td>
</tr>
<tr>
<td>Neural net</td>
<td>15</td>
<td>W 14 (21)</td>
</tr>
</tbody>
</table>

Pulse-Echo Signal Classification

• Classification of fixed-length records into 2 categories
• Wavelet compression: selects 25 of 192 coefficients, achieves 97% accuracy
• ANNIGMA-wrapper: selects 10 coefficients (only 2 in wavelet results), achieves 99% accuracy
Helicopter Rotor Blade Strain Signal Prediction

• Identify a subset of helicopter instrument signals that can predict the rotor blade strain signal
• 41 signals form the initial attribute set
• The Signals are sampled at 8 bits, 1000 samples/second
• Resulting data set size approximately 1 Gigabyte
• Noise, data dropouts, and unknown data ranges are additional challenges
• One ANNIGMA-wrapper cycle takes 200 hours of exclusive Pentium-pro PC run-time

Strain Signal Prediction - Results

<table>
<thead>
<tr>
<th>Wrapper Cycle</th>
<th>Number of Attributes</th>
<th>Average Absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41</td>
<td>0.173</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>0.169</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>0.163</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>0.774</td>
</tr>
</tbody>
</table>

• At wrapper cycle 4, the error increases, therefore the attribute subset associated with wrapper cycle 3 will be used
• Other signals and signal transforms will be used to increase accuracy in subsequent ANNIGMA-wrapper cycles.
Strain Signal Prediction - Results
Wrapper Cycle 3 Neural Net

Conclusion - Contributions

• The ANNIGMA heuristic can rank attribute relevance from a trained neural net’s weights
• Wrapper model attribute subset selection becomes practical for real-world neural net-based applications development when accelerated by the ANNIGMA heuristic
• Some limitations of the ANNIGMA-wrapper algorithm
  – Requirement for normalization
  – Requires more computer resources than wavelet compression