Minimal Cost Complexity Pruning of Meta-Classifiers

Andreas L. Prodromidis
Salvatore J. Stolfo

Department of Computer Science
Columbia University
Combining multiple models

Training data set

Learning Algorithm
Classifier-1

Learning Algorithm
Classifier-2

Learning Algorithm
Classifier-3

Meta-learning

Meta-Classifier
**Meta-learning**

- **Training Data** $D_1$
- **Learning Algorithm** $L_1$
- **Classifier** $C_1 = L_1(D_1)$
- **Validation Data** $D$
- **Classifier** $C_2 = L_2(D_2)$
- **Predictions** $C_1(D)$
- **Predictions** $C_2(D)$

**Meta Classifier**

- **Meta-learning Algorithm** $ML$
- **Meta-level Training Data**
A meta-learning training set example

Validation set

<table>
<thead>
<tr>
<th>CID</th>
<th>InputType</th>
<th>Amt</th>
<th>...</th>
<th>True Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>54341</td>
<td>Swipe</td>
<td>19.72</td>
<td></td>
<td>Legitimate</td>
</tr>
<tr>
<td>54432</td>
<td>KeyIn</td>
<td>88.19</td>
<td></td>
<td>Fraudulent</td>
</tr>
<tr>
<td>54101</td>
<td>Phone</td>
<td>11.99</td>
<td></td>
<td>Legitimate</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Meta-level training set - Stacking (Wolpert-92)

<table>
<thead>
<tr>
<th>Classifier-1</th>
<th>Classifier-2</th>
<th>Classifier-3</th>
<th>...</th>
<th>True Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legitimate</td>
<td>Legitimate</td>
<td>Legitimate</td>
<td>...</td>
<td>Legitimate</td>
</tr>
<tr>
<td>Legitimate</td>
<td>Fraudulent</td>
<td>Legitimate</td>
<td>...</td>
<td>Fraudulent</td>
</tr>
<tr>
<td>Fraudulent</td>
<td>Fraudulent</td>
<td>Legitimate</td>
<td>...</td>
<td>Legitimate</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Meta-Classifying

Meta-Classifying data

Testing (unclassified) Data

Classifier C₁

Classifier C₂

Classifier C₃

Meta Classifier MC

Predictions

C₁(x)

C₂(x)

C₃(x)

MC(C₁(x), C₂(x), C₃(x))
Decision tree meta-classifier

- Ripper classifier
- Fraud? Legitimate?
- Bayes classifier
- Fraud? Legitimate?
- CART classifier
- ID3 classifier
- CART classifier
Efficiency

- Compute base classifiers in parallel
- Compute “small” meta-classifiers
  - to reduce memory requirements
  - to produce fast classifications
- Pre-training pruning
  - filter before meta-learning (NIT’98, KDD’98-DDM)
- Post-training pruning
  - discard after meta-learning (Prodromidis-et-al-98)
A graphical description

Map arbitrary meta-classifier to a decision tree representation

Prune the decision tree model

Map pruned decision tree to original meta-classifier representation

(Mapping via modeling of the meta-classifier’s behavior)
Post-training pruning

- Minimal cost complexity pruning (Breiman-et-al-84)
  - \( R(T) \): misclassification cost of a decision tree \( T \)
  - \( C(T) \): complexity of tree (= number of terminal nodes)
  - \( \alpha \): complexity parameter
- Seek to minimize \( R_\alpha(T) \), \( R_\alpha(T) = R(T) + \alpha \cdot C(T) \)
Decision tree model (unpruned)

\[ R_\alpha(T) = R(T) + \alpha \cdot C(T) \]
Decision tree model (pruned)

$R_\alpha(T) = R(T) + \alpha \cdot C(T)$
Decision tree modeling of meta-classifiers

- Meta-Classificr
  - Classifiers

- Decision Tree Meta-Classificr
  - Classifiers

- Meta-level Training Data
- Decision Tree Learning Algorithm (e.g. CART)
- Decision Tree Training Data
- Predictions
Final pruned meta-classifier
Credit Card Fraud detection

- Chase Credit Card data
  - 500,000 transaction records
  - 30 attributes (numerical, categorical) in 137 bytes per record
  - 20% fraud, 80% non fraud

- First Union Credit Card data
  - 500,000 transaction records
  - 28 attributes (numerical categorical) in 137 bytes per record
  - 15% fraud, 85% non fraud

- Attributes
  - Hashed credit card account number, date, time, type of entry of transaction, type of merchant, amount, validity codes, past payment information, account information, confidential fields, etc.
  - The fraud label
Experimental setting

• Divide data sets in 12 subsets at 6 sites
• Five learning algorithms
  – Naïve Bayes, C4.5, CART, ID3, Ripper
• Exchange classifiers
  – 10 local, 50 remote per site
• Meta-learn only the remote classifiers
## Meta-learning results

<table>
<thead>
<tr>
<th>Type of Classification model</th>
<th>Size</th>
<th>Accuracy</th>
<th>TP-FP</th>
<th>Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best over a single subset</td>
<td>1</td>
<td>88.5%</td>
<td>0.551</td>
<td>$812K</td>
</tr>
<tr>
<td>Best over largest possible subset</td>
<td>1</td>
<td>88.8%</td>
<td>0.568</td>
<td>$840K</td>
</tr>
<tr>
<td>Meta-classifier</td>
<td>50</td>
<td>89.6%</td>
<td>0.621</td>
<td>$818K</td>
</tr>
<tr>
<td>Chase's COTS system</td>
<td>--</td>
<td>85.7%</td>
<td>0.523</td>
<td>$682K</td>
</tr>
</tbody>
</table>

Chase data
Maximum savings: $1,470K

<table>
<thead>
<tr>
<th>Type of Classification model</th>
<th>Size</th>
<th>Accuracy</th>
<th>TP-FP</th>
<th>Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best over a single subset</td>
<td>1</td>
<td>95.2%</td>
<td>0.749</td>
<td>$806K</td>
</tr>
<tr>
<td>Best over largest possible subset</td>
<td>1</td>
<td>95.3%</td>
<td>0.787</td>
<td>$828K</td>
</tr>
<tr>
<td>Meta-classifier</td>
<td>50</td>
<td>96.5%</td>
<td>0.831</td>
<td>$944K</td>
</tr>
</tbody>
</table>

First Union data
Maximum savings: $1,085K
Pruning results
More information

- About the paper

- About the JAM project

- E-mail contact
  - andreas@cs.columbia.edu