Using Association Rules to Make Rule-based Classifiers Robust

Hong Hu and Jiuyong Li

Department of Mathematics and Computing
University of Southern Queensland
Australia

huhong@usq.edu.au, jiuyong@usq.edu.au
Outline

1. Motivation.
2. Problem.
3. Some previous theoretical results.
4. Experimental validation.
5. A robust classifier — OAC.
6. Conclusions.
## Motivation

<table>
<thead>
<tr>
<th>NO.</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
A decision tree

What prediction will this tree make when a coming record containing no Outlook information?
Robust prediction problem

Making prediction on the test data that is less complete than the training data.

Practical implication:
Training data, typically some selective history data, more controllable.
Test data, future coming data, less controllable.

Another robustness problem:
Handling missing data in training data.
Different from other existing methods

- General methods for handling missing values are to pre-process data by substituting missing values by estimations by some approaches, e.g. the nearest neighbours (Batista & Monard 2003).
- This method does not estimate and substitute any missing values, but to make use of larger rule sets to make classifiers be “immune” from the missing test data.
Rules

• Simple and expressive presentation of knowledge.

• Good understandability and interpretability.

• A major goal in machine learning and data mining.

• Many types of rules, classification rules, association rules and first order rules.

• Only consider classification rules.
Rule discovery methods

- The covering algorithm based methods, e.g. AQ15 (Michalski, Mozetic, Hong & Lavrac 1986) and CN2 (Clark & Niblett 1989, Clark & Boswell 1991).

- Decision tree based methods (simultaneous covering algorithm), e.g. C4.5rules (Quinlan 1993).

- Association rule based methods, e.g. Apriori (Agrawal & Srikant 1994) and FP-growth (Han, Pei & Yin 2000).

- Optimal class association rule based methods, e.g. constraint association rule mining method (Bayardo, Agrawal & Gunopulos 2000) and optimal class association rule mining (Li, Shen & Topor 2002).
Rule-based Prediction — Classifiers

- Ordered rule based classifiers: rules are organised as a sequence usually in the descending accuracy order, and only the first matching rule makes a prediction. For example, C4.5rules (Quinlan 1993) and CBA (Liu, Hsu & Ma 1998).

- Unordered rule based classifiers: rules are not organised in a sequence, and all (or most) matching rules participate in the prediction. For example, improved CN2 (Clark & Boswell 1991) and CMAR (Li, Han & Pei 2001).
## Comments

<table>
<thead>
<tr>
<th>Ordered classifiers</th>
<th>Unordered classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple and easy to understand</td>
<td>May be too complex to understand</td>
</tr>
<tr>
<td>How to sort rules?</td>
<td>How to resolve dependency?</td>
</tr>
<tr>
<td>The default prediction.</td>
<td>How to vote?</td>
</tr>
</tbody>
</table>

The ordered rule based classifier is still a simple and effective option.

The covering algorithm is used to sort rules by reducing false positives.
Robust rule based prediction

RuleSet1, from decision tree
1. If outlook is sunny and humidity is high, then do not play tennis.
2. If outlook is sunny and humidity is normal, then play tennis.
3. If outlook is overcast, then play tennis.
4. If outlook is rain and wind is strong, then do not play tennis.
5. If outlook is rain and wind is weak, then play tennis.

RuleSet2, some complementary rules for missing outlook information
6. If humidity is normal and wind is weak, then play tennis.
7. If temperature is cool and wind is weak, then play tennis.
8. If temperature is mild and humidity is normal, then play tennis.
9. If humidity is normal, then play tennis.

(RuleSet1 + RuleSet2) is more robust than RuleSet1.
**k-optimal rule sets — an illustration**

<table>
<thead>
<tr>
<th>Record: $a, b, c, d, e$ class $z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (or min)-optimal rule set</td>
</tr>
<tr>
<td>$a \rightarrow z$</td>
</tr>
</tbody>
</table>

Precise definitions please refer (Li, Topor & Shen 2002).
$k$-optimal rule sets — some conclusions in informal way

1. The optimal rule set is the most robust rule set.

2. A $(k + 1)$-optimal rule set is at least the same robust as a $k$-optimal rule set.

Precise theorems please refer (Li, Topor & Shen 2002).
Robust rule based prediction — experimental design

- use 10 cross validation.
- Randomly add missing values to test data controlled by parameter $l$.
  On average each record has $l$ missing values.
- Repeat $10 \times 10$ times for one data set.
- Experiment on 26 data sets form UCML.
- Compare with some benchmarks: C4.5rules and CBA.
Figure 1: The average size of different classifiers on 28 data sets
Figure 2: The average accuracy of four classifiers with no default on 28 data sets
Comments

- A traditional brief: large classifiers cause overfitting problems.
- They do not necessarily overfit data sets, but are evidently more robust.
**OAC — a 1-Optimal Association rule based Classifier**

- Mining the optimal rule set (Li, Shen & Topor 2002).
- Select 1-optimal rule set (Li, Topor & Shen 2002). Laplace accuracy is used.
- Sort rules in 1-optimal rule set by the covering algorithm to minimize misclassification rate.
- (Optional) The majority class in the remaining un-classification data is set as the default class.
Default — good or bad?

What prediction should a classifier make on a test record when there is no rule matching it?

- With the default, predicted to be the default class.
  
  Problem: a medial data set containing 99% healthy cases and 1% disease cases.
  
  Setting default as health gives 99% accuracy!?

- Without the default, predicted to be non-class and counted as an error.
  
  Accuracy is more realistic.
Figure 3: The average accuracy of three classifiers with the default on 28 data sets
The number of missing attribute values per row
Accuracy (no default) (%)
The average of 28 data sets (\(l_{sup} = 0.01\) conf = 0.5 len = 6)

OAC (noDefault)
C4.5rules (noDefault)

Figure 4: The average accuracy of two classifiers without default on 28 data sets
Conclusions

• Experimentally verify that \((k + 1)\)-optimal rule sets are more robust than \(k\)-optimal rule sets.

• A 1-optimal rule based classifier, OAC, is more robust than two benchmark rule based classifiers, C4.5rules and CBA.

• OAC is as accurate as C4.5rules and CBA when the test data is as complete as the training data set, and OAC relies less on the default prediction than C4.5rules.
References


Han, J., Pei, J. & Yin, Y. (2000), Mining frequent patterns without candidate generation, in ‘Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD’00)’, May, pp. 1–12.


Liu, B., Hsu, W. & Ma, Y. (1998), Integrating classification and
association rule mining, in ‘Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)’, pp. 27–31.


Thank you.

Questions.