Cost-Sensitive Decision Forest: CSForest
Michael J. Siers and Md Zahidul Islam
School of Computing and Mathematics, Charles Sturt University

Our CSForest Publications

The technique CSForest was a product of Michael’s Honours in 2014.
» CORE Rank B Conference AI 2014 [2].

Introduction

» CSForest is a cost-sensitive classification algorithm. It uses the methodology of the decision forest algorithm SysFor [3] with the splitting criteria used in CSTree [4].

In our two related publications, we applied CSForest to the software defect prediction problem. This is an application domain in which classification is used to determine which sections of a program’s source code are likely to contain defects (bugs).

Software Defect Prediction Datasets

» We use the publicly available NASA-MDP software defect datasets. Features of the dataset include:
- Various line of code counts (logical, whitespace etc.)
- Halstead’s Complexity Metrics
- Cyclomatic Complexity Metrics

» The number of records of the datasets range from roughly 500 to 1500 records.

Software Defect Prediction Datasets

The number of attributes/features of the datasets are roughly 30.

» Each record in the dataset describes a software module (typically a function/method).

CSForest

» CSForest first finds a set of “good” splitting points for a dataset based on the total expected cost criterion. The good splitting points are constrained by two user-defined settings:
- Separation: If it is a numerical split, the splitting value is not too close to another good split.
- Goodness: The split is close enough to the best split for minimizing the splitting cost criterion.

» Building the trees
- Each good splitting point is used a root node in a CSTree decision tree.
- If the user requested more trees, then more trees are built using the good splitting points at the next level in the tree.

» Classifying new records (CSVoting)
- The set of leaves \( L \) that the record \( r \) falls in is found.
- \( \forall l \in L \), the total expected cost is found for each possible classification.
- Finally, \( r \) is classified as the cheapest classification.

The splitting criterion (From CSTree)

Consider a distribution \( d \) with records which belong to either the positive class or the negative class. Given that we know the cost of making true positive \( C_{TP} \), true negative \( C_{TN} \), false negative \( C_{FN} \), and false positive \( C_{FP} \) predictions, we can calculate the cost of labelling every record in the distribution as negative \( C_N \) or positive \( C_P \):

\[
C_N = N_{TN} \times C_{TN} + N_{FN} \times C_{FN}
\]

\[
C_P = N_{TP} \times C_{TP} + N_{FP} \times C_{FP}
\]

Where \( N_{TN}, N_{TP}, N_{FN}, \) and \( N_{FP} \) are the number of true negative, true positive, false negative, and false positive predictions respectively.

The total expected cost is then calculated as:

\[
E = \frac{2 \times C_P \times C_N}{C_P + C_N}
\]

Favoured splits reduce the total cost the most after splitting compared to without splitting.

Results: Table

**Performance Comparison (lower cost, the better)**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CSForest</th>
<th>CSTree</th>
<th>SysFor (Voting 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC2′</td>
<td>165</td>
<td>164</td>
<td>129</td>
</tr>
<tr>
<td>PC1′</td>
<td>290</td>
<td>255</td>
<td>276</td>
</tr>
<tr>
<td>KC1′</td>
<td>1187</td>
<td>1404</td>
<td>1168</td>
</tr>
<tr>
<td>PC3′</td>
<td>586</td>
<td>664</td>
<td>521</td>
</tr>
<tr>
<td>MC1′</td>
<td>291</td>
<td>275</td>
<td>261</td>
</tr>
<tr>
<td>PC2′</td>
<td>84</td>
<td>81</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 1: Cost comparison

We used the following cost-matrix in our comparison.

\[
C_{TP} = 1 \quad C_{TN} = 0
\]

\[
C_{FP} = 1 \quad C_{FN} = 5
\]

Table 2: Cost-Matrix

Results: Figure

Figure 1: Example module

Figure 2: Average CSForest Performance Increase from Existing Techniques

References


Acknowledgments

Thank you to the various reviewers for the two CSForest related publications.

Contact Information

» Email: msiers@csu.edu.au, zislam@csu.edu.au
» Phone: Michael: +61435608521