DATA TRANSFORMATION
AND ATTRIBUTE SUBSET SELECTION:
DO THEY HELP
MAKE DIFFERENCES IN
SOFTWARE FAILURE PREDICTION?

Hao Jia, Fengdi Shu, Ye Yang, Qi Li
Institute of Software, Chinese Academy of Sciences
Sep 23, 2009
OUTLINES

- Introduction
- Study Design & Result Interpretation
- Discussion
- Conclusion
OUTLINES

- Introduction
- Methodology
- Experiment
- Conclusion
INTRODUCTION

• Background
  • Challenging issue - how to dispose of the asymmetric data and irrelevant attributes effectively in software engineering
  • Commonly adopted strategies
    ➢ data transformation (DT)
    ➢ attribute subset selection (AttrSS)
• Motivation

• Decision making - improve the convergence of the selection of DT and AttrSS methods to predict the binaries across releases with post-release defects for iteratively developed projects.

• Context - explore the evaluation of these methods between an open-source project and an in-house project.
OUTLINES

- Introduction
- Study Design & Result Interpretation
- Discussion
- Conclusion
**Selection criteria for DT and AtrSS methods**

- Reported satisfactory effects
- Ease to use

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Method</th>
<th>Abbr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Transformation (DT)</td>
<td><em>Normalization</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>min-max</td>
<td>min-max</td>
</tr>
<tr>
<td></td>
<td>z-score</td>
<td>z-score</td>
</tr>
<tr>
<td>Log Transformation</td>
<td></td>
<td>log</td>
</tr>
<tr>
<td>Attribute Subset Selection (AttrSS)</td>
<td><em>Statistical Methods</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Principal component analysis</td>
<td>PCA</td>
</tr>
<tr>
<td><em>Search-based Methods</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>information gain attribute ranking</td>
<td>InfoGain</td>
</tr>
<tr>
<td></td>
<td>correlation-based feature selection</td>
<td>CFS</td>
</tr>
<tr>
<td></td>
<td>consistency-based subset evaluation</td>
<td>CBS</td>
</tr>
</tbody>
</table>
## Datasets

### Open-sourced Eclipse

<table>
<thead>
<tr>
<th>Release</th>
<th># Files</th>
<th>% Failure-prone</th>
<th># Packages</th>
<th>% Failure-prone</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>6729</td>
<td>14.49%</td>
<td>377</td>
<td>50.40%</td>
</tr>
<tr>
<td>2.1</td>
<td>7888</td>
<td>10.83%</td>
<td>434</td>
<td>44.70%</td>
</tr>
<tr>
<td>3.0</td>
<td>10593</td>
<td>14.80%</td>
<td>661</td>
<td>47.35%</td>
</tr>
</tbody>
</table>

### Closed-sourced QMP

<table>
<thead>
<tr>
<th>Release</th>
<th># Modules</th>
<th>% Failure-prone Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>30.00%</td>
</tr>
<tr>
<td>B</td>
<td>31</td>
<td>25.81%</td>
</tr>
<tr>
<td>C</td>
<td>37</td>
<td>59.50%</td>
</tr>
</tbody>
</table>
Classifiers
- J48 decision tree (J48), Naïve Bayes (NB), IB1, and Random Forest (RF)

Performance measures
- Area under the receiver operating characteristic (AUC) over ten 10-fold cross validation

Steps
- Using classifiers directly on original data (raw classifiers for short); then
- Using DT methods; and if they improve the performance, then using AttrSS methods on the transformed data; otherwise, using them directly on original data.
Effects of data transformation

These DT methods rarely improve or debase the performance.
**RESULTS AND INTERPRETATION**

- **Effects of attribute subset selection**
  - More than 80% cases record distinguishable performances between predictions with and without AttrSS methods.
RESULTS AND INTERPRETATION

- Consistency analysis across releases and granularities
- Data transformation methods
  - For Eclipse dataset - insensitive to different releases and granularities
  - For QMP datasets - much more evidently consistent across releases with NB and IB1; with J48 and RF, little discrepancy emerges.
Consistency analysis across releases and granularities

Attribute subset selection methods

- For Eclipse dataset - the discrepancies among methods grow obscure when granularity goes coarse.
- For QMP datasets - similar conditions occur.
- In conclusion, unless InfoGain is available with fittest threshold, the choice of AttrSS methods displays little consistency in preference across releases and granularities.
OUTLINES

- Introduction
- Study Design & Result Interpretation
- Discussion
- Conclusion
EVALUATION OF DT METHODS

- Z-score normalization works better than min-max normalization.
- Log transformation outputs unacceptable results with NB and IB1.

Recommendations. Data transformation display unobvious influence on improving classifiers’ effects for failure prediction, while log transformation is of high possibility to degrade Naïve Bayes.
Evaluation of AttrSS methods

- Information gain behaves the best with fittest threshold.
- Predictive performance is generally degraded by PCA.
- CBS and CFS perform generally well across classifiers.

Recommendations. Attribute subset selection is a useful way to enhance the predictive performance. InfoGain is optimal, but when the fittest subset is not at hand (since the computation is effort-consuming), its performance goes unsatisfactory; PCA seems useless or even harmful; CBS and CFS may be a tradeoff.
• Discrepancy in the evaluation between Eclipse and QMP
  • Eclipse datasets at file and package level vs. QMP datasets at module level.
  • Recommendations. Concerning the distinctive properties of these two types of projects, separate strategies should be planned for them.
**DISCUSSION**

- **Threats to validity**
  - Main threat to external validity
    - generalization
  - Main threat to internal validity
    - choice of AUC as performance indicator,
    - selection of statistical test methods, and
    - the 4 classifiers used in our study
OUTLINES

- Introduction
- Study Design & Result Interpretation
- Discussion
- Conclusion
CONCLUSION

• What we studied
  • Applicability and efficiency of data transformation and attribute subset selection methods in failure prediction.

• What we offers
  • Recommendations in selecting attribute subset selection and data transformation methods for improving classifiers’ predictive performance.

• Our future work
  • Extend the scope of current analysis
  • Search for more concrete and conclusive recommendations in practice.
Thank You!