Data Transformation and Attribute Subset Selection: Do they Help Make Differences in Software Failure Prediction?

Hao Jia1, Fengdi Shu1, Ye Yang1, Qi Li2

1 Institute of Software, Chinese Academy of Sciences, China.
2 University of Southern California, USA.

{jiahao, fdshu, ye}@itechs.iscas.ac.cn, qli1@usc.edu

Abstract

Data transformation and attribute subset selection have been adopted in improving software defect/failure prediction methods. However, little consensus was achieved on their effectiveness. This paper reports a comparative study on these two kinds of techniques combined with four classifier and datasets from two projects. The results indicate that data transformation displays unobvious influence on improving the performance, while attribute subset selection methods show distinguishably inconsistent output. Besides, consistency across releases and discrepancy between the open-source and in-house maintenance projects in the evaluation of these methods are discussed.

1. Introduction

Various code attributes have been widely used as the independent predictors. However, the distributions of many numeric attributes in defect datasets tend to be asymmetric (skewed), violating statistical criteria regarding distributional assumptions. Besides, in practice most data sets contain attributes either irrelevant or even harmful to a particular prediction goal. Hence, how to dispose of the data and attributes effectively comes into a challenging issue. Among all the methods used to remove the potential inconsistency, irrelevancy, and redundancy, data transformation (DT) and attribute subset selection (AttrSS) are the two commonly adopted strategies, which have been introduced into the defect prediction literature.

Our study aims at improving the convergence of the selection of DT and AttrSS methods to predict the binaries with post-release defects across releases for iteratively developed projects. Another goal is to explore the results in the evaluation of these methods between an open-source project (i.e. Eclipse) and a closed-sourced project (i.e. QMP). These DT methods rarely improve or debase the performance as shown in figure 2 a). In details, most of transforming methods behave rare superiority to raw classifier, represented by few wins. Specifically, for classifiers NB and IB1, most transformation methods result in degradations; as for classifiers J48 and RF, neither improvement nor degradation is discerned with few exceptions.

2. Study design

With respect to the selection criteria as reported satisfactory effects and ease to use, we study two kinds of DT methods as shown in table 1.

Table 1. Studied DT and AttrSS methods.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Method</th>
<th>Abbr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Transformation</td>
<td>Normalization</td>
<td></td>
</tr>
</tbody>
</table>

We conduct the study on datasets from two projects of different types, open-source Eclipse and closed-sourced QMP. To mitigate the potential bias caused by using single classifier, we select four classifiers - J48 decision tree (J48), Naïve Bayes (NB), IB1, and Random Forest (RF). Besides, to remedy problems in measuring the predictive performance caused by routinely adopted indicator-accuracy, area under the receiver operating characteristic (AUC) is calculated and averaged over ten 10-fold cross validation.

This study predicts failure-prone binaries with following procedures respectively:
1) using classifiers directly on original data (raw classifiers for short);
2) using DT methods; and if they improve the performance, then using AttrSS methods on the transformed data; otherwise, using them directly on original data.

3. Results and interpretation

- **Effects of data transformation methods**
  These DT methods rarely improve or debase the performance as shown in figure 2 a).
  
- **Effects of attribute subset selection methods**
  More that 80% cases record distinguishable performances between predictions with and without AttrSS methods as shown in figure 2 b). Given by pairwise comparisons, each classifier is improved by at least one AttrSS method, even though there is no consistency in their ranks across classifiers. Especially, the performance of J48 is improved by all methods in a significant level. There is a slender exception that, for RF, even though CBS and InfoGain improve it, the differences are not evident enough to bear out their excellence.

- **Consistency analysis across releases and granularities**
  In evaluation of data transformation methods, Eclipse dataset can be treated insensitively to different releases and granularities. For QMP datasets, much more evident consistency across releases according to NB and IB1 are found. With J48 and RF, little discrepancy emerges.
  
In the evaluation of 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.

Figure 2. Mean AUC values obtained by a) DT methods and b) AttrSS methods.

4. Discussion

- **Evaluation of data transformation 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 attribute subset selection 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.

- **Threats to validity**
  The main threat of this study to external validity is that we cannot assume a priori that the results of a study on one project generalize all the projects with the same type, even though we take actions to reduce the potential bias by adopting subject projects with different types and developing methodologies, various classifiers, and candidate methods. Possible threats to internal validity include issues such as the choice of AUC as performance indicator, selection of statistical test methods, and the 4 classifiers used in our study.

5. Conclusions

This paper studied the applicability and efficiency of data transformation and attribute subset selection methods in failure prediction. It offers various beneficial recommendations in selecting attribute subset selection and data transformation methods for improving classifiers’ predictive performance, which is helpful in maintaining iteratively developed projects.

Future work includes investigation and extension of current study to extend the scope of current analysis to cover more projects to improve the applicability of our results and search for more concrete and conclusive recommendations in practice.