Western Kentucky University

From the SelectedWorks of Dr. Huanjing Wang

December, 2009

Mining Data from Multiple Software Development Projects

Huanjing Wang, Western Kentucky University
Taghi M. Khoshgoftaar, Florida Atlantic University
Kehan Gao, Eastern Connecticut State University
Naeem Seliya

Available at: https://works.bepress.com/huanjing_wang/1/
Mining Data from Multiple Software Development Projects

Huanjing Wang, Taghi M. Khoshgoftaar, Kehan Gao, and Naeem Seliya
{huanjing.wang@wku.edu; taghi@cse.fau.edu; gaok@easternct.edu; nseliya@umich.edu}

Abstract

A large system often goes through multiple software project development cycles, in part due to changes in operation and development environments. For example, rapid turnover of the development team between releases can influence software quality, making it important to mine software project data over multiple system releases when building defect predictors. Data collection of software attributes are often conducted independent of the quality improvement goals, leading to the availability of a large number of attributes for analysis. Given the problems associated with variations in development process, data collection, and quality goals from one release to another emphasizes the importance of selecting a best-set of software attributes for software quality prediction. Moreover, it is intuitive to remove attributes that do not add to, or have an adverse effect on, the knowledge of the consequent model. Based on real-world software projects’ data, we present a large case study that compares wrapper-based feature ranking techniques (WRT) and our proposed hybrid feature selection (HFS) technique. The comparison is done using both three-fold cross-validation (CV) and three-fold cross-validation with risk impact (CVR). It is shown that HFS is better than WRT, while CV is superior to CVR.

Keywords: data preparation, attribute selection, data selection, software measurements, defect prediction.

1 Introduction

Developing high-quality software is a primary goal for any development team. Software practitioners strive to improve software quality by constructing software quality prediction models [9]. However, when a large system goes through multiple software releases, it is very likely that the characteristics and trends of software quality change from one project development to another. A given system release can have relatively different operation and development environments, which would be reflected in the software development process. Moreover, a rapid turnover of team members can also influence software quality characteristics. Thus, it is important to analyze software data from multiple sources, i.e., multiple development projects or releases, when building software quality prediction models.

A typical defect prediction model is built using software metrics and fault data of previously developed releases or similar projects. Then the quality of the under-development program modules is estimated, as fault-prone (fp) or not fault-prone (nfp), for example. As described earlier, characteristics of the software measurement data set play an important role in the efficacy of the defect prediction models. One such aspect is which software attributes are good quality indicators for the given system, especially when there is a likelihood of variation in development processes, data collection methods and objectives, and quality improvement goals from one release to another. We focus on feature selection in the context of software quality prediction and analysis when software measurement and defect data from multiple development sources/releases is available.

Feature selection is an essential task during the data preparation and data cleansing phases of a machine learning problem [12]. The aim of feature selection is to find a feature subset that characterizes the underlying data as good as, or even better, than the original data when all available features are considered. Feature ranking techniques rank the attributes based on their individual predictive power, while feature selection approaches select subsets of attributes that collectively have good predictive power. Feature subset selection can be categorized as filters, wrappers, or embedded methods. Filters are algorithms in which a feature subset is selected without involving any learning algorithm. Wrappers are algorithms that use feedback from a learning algorithm to determine which feature(s) to include for building a classifier. Embedded methods imply that the feature selection aspect is incorporated within a learning algorithm.

We evaluate two types of feature selection methods: wrapper-based feature ranking techniques (WRT) and our proposed hybrid feature selection method (HFS). In the case of WRT, classifiers with a single independent attribute (and the dependent class attribute) are built, and then the attributes are selected according to their individual predictive capability as measured by a performance metric. We used five different and commonly used learn-
The proposed HFS technique consists of a wrapper-based feature ranking technique and a consistency-based feature subset selection – the latter is our proposed automatic hybrid search (AHS) [10]. We assessed the feature selection methods (WRT and HFS) by building classification models with the selected attributes and evaluating them with two performance metrics [14]: Area Under ROC (AUC) and Best Geometric Mean (BGM). When building a defect prediction classifier, we considered two model-building strategies: three-fold cross-validation (CV) and three-fold cross-validation risk impact (CVR). For each strategy, features were selected through WRT and HFS, and we analyze which strategy provides better performance for a given classifier. Such a large comparative study of feature selection for defect prediction modeling is relatively unique to this paper. Moreover, all data sets analyzed in this study suffer from class imbalance, and investigating feature selection with such data sets poses additional challenges.

The paper is organized as follows. Section 2 provides information on the cross-validation strategies, feature selection techniques, and learners used. Section 3 provides a description of the software projects’ data, and Section 4 details the empirical results. Relevant literature on feature selection is presented in Section 5. Section 6 concludes the paper and provides suggestions for future work.

2 Methodology

2.1 Cross-Validation Methods

We use two strategies for the WRT approaches. First, a three-fold cross-validation (CV) method is used. For each attribute to be ranked, a reduced dataset consisting solely of that attribute and the class attribute is created. For each of the three folds (partitions), a classifier is trained on the other two folds (partitions), then tested on that fold. For each learner, all five performance metrics are calculated based on cross-validation. Performance scores are obtained in this manner for all attributes in the data set, one at a time. After the five performance metrics are obtained for all five learners for every attribute, features are ranked based on a given performance metric for each classifier.

The second method, three-fold cross-validation with risk impact (CVR), is similar to CV except that each instance of the training data set contains \( n - 1 \) independent features and the class feature. For each learner, we build a model with all \( n \) independent features and the class feature and recording all performance measures, denoted as \( PM_n \), where \( PM \) represents one of the five performance metrics. For each independent feature \( i \), we eliminate it, and build a model with one less independent feature to get the associated performance measure, denoted as \( PM_{n-1}^i \), where \( 1 \leq i \leq n \). The risk impact of each feature \( i \) for a particular learner is defined as \( PM_n - PM_{n-1}^i \). Each feature has a risk impact score and these features are ranked in a descending order according to risk impact scores. CVR is more time consuming than CV since it builds models using \( n - 1 \) independent features while CV uses only one independent feature.

2.2 Wrapper-Based Ranking Techniques

The WRT feature selection approach was implemented by our team within the WEKA framework [14]. The approach consists of two parts, a learner (classifier) and a performance metric, that collectively form a ranker (Figure 1). In the feature ranking technique, every feature in a given fit (training) data set is used individually in the model building process. We used the CV and CVR strategies to build the classification models and then assessed their performances based on a specific performance metric. We then ranked the features and selected the top \( \lceil \log_2 n \rceil \) features according to their respective scores, where \( n \) is the number of independent features for a given data set. In this study, we selected six features from the attribute set that contained 42 features.

The reasons for selecting the top \( \lceil \log_2 n \rceil \) features are: (1) no general guidance is available on the number of features that should be selected when using a feature ranking technique; (2) in a recent empirical study [8], we showed that it was appropriate to use \( \lceil \log_2 n \rceil \) as the number of features when using WEKA [14] to build a classifier for binary classification in general, and for imbalanced data sets in particular. Thus, we assume \( \lceil \log_2 n \rceil \) is a good choice for this research; and (3) a software engineering expert with more than 20 years experience recommends selecting \( \lceil \log_2 n \rceil \) number of metrics for high-assurance systems such as our case study system.

In a two-group classification, e.g., \( fp \) and \( nfp \), there are four possible outcomes: true positive (TP), false positive (FP), true negative (TN) and false negative (FN), where positive represents \( fp \) and negative represents \( nfp \). The respective frequency of the four outcomes form the basis for several other performance measures that are commonly used for classifier evaluation. We use five of them for the wrapper-based feature ranking process, and they are:

1. Overall Accuracy (OA) provides a single value range
from 0 to 1. It can be obtained by $\frac{|TP| + |FN|}{N}$, where $N$ is the total number of instances in the data set.

2. Default Geometric Mean (DGM) is a single-value measure that ranges from 0 to 1, and a perfect classifier provides a value of 1. Geometric Mean (GM) is defined as the square root of the product of true positive rate (defined as $\frac{|TP|}{|TP| + |FN|}$) and true negative rate (defined as $\frac{|TN|}{|TN| + |FP|}$). The threshold, $t = 0.5$, is used for DGM, while the Best Geometric Mean (BGM) is the maximum GM value that is obtained when varying the threshold between 0 and 1.

3. Area Under the ROC Curve (AUC) has been widely used to measure classification model performance [2]. The ROC (Receiver Operating Characteristic) curve characterizes the trade-off between the true positive rate and the false positive rate. A perfect classifier provides an AUC that equals 1.

4. Area Under the Precision-Recall Curve (PRC) is a single-value measure that depicts the trade-off between Recall and Precision [14]. A classifier that is optimal in AUC space may not be optimal in PRC space.

5. Best Arithmetic Mean (BAM) is based on the arithmetic mean of the true positive rate and true negative rate values. The Best Arithmetic Mean (BAM) uses the maximum arithmetic mean that is obtained when varying the threshold between 0 and 1.

2.3 Proposed Hybrid Feature Selection Method

Feature subset selection techniques search the set of possible features as a group and evaluate their collective suitability. However, the traditional approaches to feature selection with a single evaluation criterion have shown limited capability in terms of knowledge discovery and decision support. We proposed the hybrid feature selection (HFS) method which combines wrapper ranking and the Automatic Hybrid Search (AHS), our recently proposed feature subset selection method [10]. AHS uses the consistency rate properties, relying on its monotonic property of consistency rate. The AHS algorithm was developed and implemented by our team in Java.

The AHS algorithm works as follows. The consistency rate of the full feature set is computed first, and then starting with size 1 of any feature, the feature subsets that have the locally highest consistency rate are selected. These selected feature subsets will be used to generate supersets. The process is repeated until finding the attribute subsets that have the same consistency rate or the specified number of features is reached. The newly proposed hybrid feature selection (HFS) method consists of two steps (Figure 2). First, use a wrapper-based feature ranking technique to rank the features and the top 30% of features are selected from the ranked list. Twelve features are selected in this study; Second, apply AHS to select a subset of $k$ features with the highest local consistency rate. The size of feature subset depends on the application domain and project characteristics. In this study, $k$ is set to six.

2.4 Classifiers

Classification is a form of data analysis that can be used to extract a model that minimizes the number of classification errors on a training data set. The first step of classification is to build a classification model that can describe the predetermined set of data classes, whereas in the second step, the classification model is evaluated using an independent test data set. The software quality prediction models are built with five different learners available in WEKA [14], including: Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors, Support Vector Machines, and Logistic Regression. These learners are used in data mining serve as both aids to the ranking process and an inductive learner for the classification performance process. Unless stated otherwise, we use default parameter settings for the different learners as specified in WEKA [14]. Parameter settings are changed only when a significant improvement in performance is obtained.

3 Software Measurement Data Description

The software metrics and defect data used in our case study were collected from four consecutive releases of a very large telecommunications software system (denoted as LLTS). Each of the software measurement data sets consist of 42 software metrics: 24 product, 14 process, and four execution metrics [5]. Details on these metrics are avoided due to paper size consideration. A program module with one or more faults is considered $fp$, and $nfp$ otherwise. Labeled chronologically as SP1, SP2, SP3 and SP4, the four data sets consist of 3649, 3981, 3541 and 3978 modules, respectively. These data sets suffer from the class imbalance problem, i.e., the proportion of $fp$ modules is much lower than that of the $nfp$ modules. The proportions of $nfp$ modules of four releases are 93.72%, 95.25%, 98.67%, and 97.69%, respectively. The SP1 data set is used as fit (training) data, while the SP2, SP3 and SP4 data sets are used as independent test (evaluation) data.

4 Empirical Case Study and Results

The experiments were conducted to discover the impact of (1) standard cross-validation vs. cross-validation risk impact; (2) wrapper based feature ranking vs. hybrid feature
for each cross-validation method (CV and CVR)
for each learner (NB, MLP, KNN, SVM, and LR)
for each ranking technique (OA, DGM, AUC, PRC, and BAM)
rank features using fit data SP1
select top 6 features from ranked list
build classification model using reduced data set on SP1 with same learner
validate the model and collect the performance metrics AUC and BGM using test data SP2, SP3 and SP4

end

end

end

for each cross-validation method (CV and CVR)
for each learner (NB, MLP, KNN, SVM, and LR)
for each ranking technique (OA, DGM, AUC, PRC, and BAM)
rank features using fit data SP1
select top 30% features (12 features) from ranked list
select size 6 of feature subset using AHS
build classification model using reduced data set on SP1 with same learner
validate the model and collect the performance metrics AUC and BGM using test data SP2, SP3 and SP4

end

end

end

for each cross-validation method (WRT and HFS)
for each learner (NB, MLP, KNN, SVM and LR)
for each ranking technique (WRT vs. HFS)
select top 5 features (25 features) from ranked list
select size 5 of feature subset using AHS
build classification model using reduced data set on SP1 with same learner
validate the model and collect the performance metrics AUC and BGM using test data SP2, SP3 and SP4

end

end
Table 5. Learner – LR

<table>
<thead>
<tr>
<th>Learner</th>
<th>AUC CV</th>
<th>BGM CV</th>
<th>AUC CR</th>
<th>BGM CR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WRT</td>
<td>HFS</td>
<td>WRT</td>
<td>HFS</td>
</tr>
<tr>
<td></td>
<td>WRT</td>
<td>HFS</td>
<td>WRT</td>
<td>HFS</td>
</tr>
<tr>
<td>OS</td>
<td>0.8258</td>
<td>0.8132</td>
<td>0.7685</td>
<td>0.7685</td>
</tr>
<tr>
<td>SP2 AUC</td>
<td>0.8331</td>
<td>0.8131</td>
<td>0.7655</td>
<td>0.7649</td>
</tr>
<tr>
<td>SP2 IRC</td>
<td>0.8331</td>
<td>0.8131</td>
<td>0.7655</td>
<td>0.7649</td>
</tr>
<tr>
<td>SP2 BAM</td>
<td>0.8224</td>
<td>0.8124</td>
<td>0.7528</td>
<td>0.7526</td>
</tr>
<tr>
<td>SP3 AUC</td>
<td>0.8471</td>
<td>0.8282</td>
<td>0.7756</td>
<td>0.7747</td>
</tr>
<tr>
<td>SP3 IRC</td>
<td>0.8471</td>
<td>0.8282</td>
<td>0.7756</td>
<td>0.7747</td>
</tr>
<tr>
<td>SP3 BAM</td>
<td>0.8471</td>
<td>0.8282</td>
<td>0.7756</td>
<td>0.7747</td>
</tr>
<tr>
<td>SP4 AUC</td>
<td>0.8341</td>
<td>0.8134</td>
<td>0.7593</td>
<td>0.7584</td>
</tr>
<tr>
<td>SP4 IRC</td>
<td>0.8341</td>
<td>0.8134</td>
<td>0.7593</td>
<td>0.7584</td>
</tr>
<tr>
<td>SP4 BAM</td>
<td>0.8341</td>
<td>0.8134</td>
<td>0.7593</td>
<td>0.7584</td>
</tr>
</tbody>
</table>

Table 6. ANOVA in terms of AUC

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum Sq.</th>
<th>d.f.</th>
<th>Mean Sq.</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>123.8</td>
<td>4</td>
<td>30.9</td>
<td>3.02</td>
<td>0.035</td>
</tr>
<tr>
<td>B</td>
<td>0.03</td>
<td>1</td>
<td>0.003</td>
<td>0.02</td>
<td>0.888</td>
</tr>
<tr>
<td>C</td>
<td>124.1</td>
<td>4</td>
<td>31.0</td>
<td>3.03</td>
<td>0.036</td>
</tr>
<tr>
<td>D</td>
<td>0.004</td>
<td>1</td>
<td>0.004</td>
<td>0.02</td>
<td>0.888</td>
</tr>
<tr>
<td>A×B</td>
<td>0.152</td>
<td>4</td>
<td>0.038</td>
<td>0.33</td>
<td>0.782</td>
</tr>
<tr>
<td>A×C</td>
<td>0.191</td>
<td>16</td>
<td>0.012</td>
<td>0.26</td>
<td>0.735</td>
</tr>
<tr>
<td>A×D</td>
<td>0.007</td>
<td>4</td>
<td>0.002</td>
<td>0.02</td>
<td>0.888</td>
</tr>
<tr>
<td>B×C</td>
<td>0.023</td>
<td>1</td>
<td>0.023</td>
<td>0.26</td>
<td>0.735</td>
</tr>
<tr>
<td>B×D</td>
<td>0.003</td>
<td>1</td>
<td>0.003</td>
<td>0.02</td>
<td>0.888</td>
</tr>
<tr>
<td>C×D</td>
<td>0.007</td>
<td>4</td>
<td>0.002</td>
<td>0.02</td>
<td>0.888</td>
</tr>
<tr>
<td>Error</td>
<td>0.945</td>
<td>256</td>
<td>0.003</td>
<td>0.02</td>
<td>0.888</td>
</tr>
<tr>
<td>Total</td>
<td>2586</td>
<td>299</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Multiple comparisons for AUC

(6) test data set (SP2, SP3 or SP4). For example, the first value in Table 1, 0.6977, refers to the predictive accuracy in terms of AUC for the NB classifier on the first test data set (SP2), where the NB classifier was built with the six features that were selected using the ranker, which was formed by the NB learner in conjunction with the OA performance metric and using the CV approach on the fit data.

We perform a four-way ANalysis Of Variance (ANOVA) F-test with respect to the AUC and BGM metrics (separately), to statistically examine the various effects on the models’ performances. The four main factors in the ANOVA test are: Factor A, for the five performance metrics used for the rankings (OA, DGM, AUC, PRC and BAM); Factor B, for the two cross-validation strategies (CV and CVR); Factor C, for the five learners (NB, MLP, KNN, SVM and LR); and Factor D, for the two feature selection methods (WRT and HFS). The interaction effects of any two factors were also considered in the ANOVA test.

The ANOVA model can be used to test the hypothesis

Table 7. Classification on 42 attributes

<table>
<thead>
<tr>
<th>Learner</th>
<th>AUC SP1</th>
<th>BGM SP1</th>
<th>AUC SP2</th>
<th>BGM SP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.8149</td>
<td>0.7451</td>
<td>0.7985</td>
<td>0.7248</td>
</tr>
<tr>
<td>MLP</td>
<td>0.8014</td>
<td>0.7399</td>
<td>0.7222</td>
<td>0.7068</td>
</tr>
<tr>
<td>KNN</td>
<td>0.7649</td>
<td>0.7235</td>
<td>0.8054</td>
<td>0.7346</td>
</tr>
<tr>
<td>SVM</td>
<td>0.6862</td>
<td>0.7019</td>
<td>0.6619</td>
<td>0.6612</td>
</tr>
<tr>
<td>LR</td>
<td>0.8287</td>
<td>0.7486</td>
<td>0.7989</td>
<td>0.7357</td>
</tr>
</tbody>
</table>

Figure 3. Multiple comparisons for AUC
that the AUC (or BGM) for the main factors A, B, C and D and/or for the interaction effects (terms) A × B, A × C, A × D, B × C, B × D and C × D are equal against the alternative hypothesis that at least one mean is different. If the alternative hypothesis is accepted, multiple comparisons can be used to determine which of the means are significantly different from the others. We performed the multiple comparison tests using Tukey’s honestly significant difference (HSD) criterion. A significance level of \( \alpha = 5\% \) is used for all statistical tests. The ANOVA results based on AUC are presented in Table 6. From the table, we can see that the \( p \)-values (last column of Table 6) for the main factors A and C and the interaction terms A × B and A × C are less than 0.05, indicating the AUC values are not similar for all groups in each of these factors or interaction terms.

Additional multiple comparisons for the main factors and interaction terms were performed to investigate the difference among the respective groups (levels). Although the ANOVA tests showed that all the group means are equal (similar) for Factor B and Factor D, we also performed the multiple comparisons for both factors to identify which group means are higher. The test results are shown in Figure 3, where each sub-figure displays graphs with each group mean represented by a symbol (◦) and 95% confidence interval. The summarized results reveal the following:

1. For Factor A (Figure 3(a)), the rankers based on AUC or PRC as performance metrics significantly outperformed those using OA or DGM as performance metrics. The rankers based on the BAM performance metric come in within the two above groups.

2. For Factor B (Figure 3(b)), CV performed generally better than CVR (at significance level of 0.336). However, since CV has lower relative complexity, we recommend CV in this research.

3. For Factor C (Figure 3(c)), the NB, MLP and LR classifiers performed significantly better than the SVM and KNN classifiers. While MLP and LR performed the best, SVM performed the worst.

4. For Factor D (Figure 3(d)), HFS outperformed WRT (at significance level of 0.252). HFS searches a subset of features that collectively has good predictive power, while WRT only considers each feature’s predictive capability; therefore, may miss a feature that only has better predictive power when combined with another feature(s). We recommend HFS in this research.

5. For the interaction term A × B (Figure 3(e)), 10 groups (levels) are presented – five performance metrics used in conjunction with two cross-validation methods. The rankers based on AUC and PRC performed better than the rankers based on OA and DGM for both cross-validation methods, CV and CVR. The rankers based on BAM outperformed the rankers based on all other performance metrics for the CV method. In contrast, rankers based on BAM performed worst compared to all other performance metrics for the CVR method.

6. For the interaction term A × C (Figure 3(f)), 25 groups (rankers) are presented. The rankers that use the NB, MLP and LR learners and based on the AUC, PRC and BAM performance metrics outperformed the other rankers. This conclusion is consistent with the results obtained from the main factors A and C. Subsequent to an ANOVA test and multiple comparisons analysis with respect to BGM, the conclusions drawn were very similar to those when AUC is the response variable.

We also compared the performances (with respect to AUC and BGM) of the defect prediction models built with the subsets of features to those built with the complete set of features – Table 7. A \( t \)-test was used for the comparisons at the \( \alpha = 0.05 \) significance level. For each learner, we have two groups of classification models: one group was created by using the smaller subsets of features (six attributes in this study as explained earlier), while the other group was created by using the original dataset with 42 software attributes. The \( t \)-test results are presented in Table 8. For all the learners, except KNN, the classification performances are very similar (\( p \)-values are less than 0.05) regardless of whether the complete set of attributes were used to build the prediction models or when the selected subsets of features were used to build the prediction models. In the case of KNN, the models constructed using the complete set of attributes performed better than the models constructed using the smaller subsets of features.

### 5 Related Work

We provide a brief overview of the most relevant works on feature selection, primarily in the software engineering field. The reader is referred to the cited references for further details on the respective works – due to the paper size.

The application of feature selection to problems in the software quality and reliability engineering is rather limited. Chen et al. [1] study wrapper-based techniques for software cost estimation, concluding that the data reduction improved estimations. Rodríguez et al. [12] evaluate three filter- and three wrapper-based models for software metrics and defect data sets, with the conclusion that wrappers were better than filters but only at a high computational cost. Compared to our work, Rodríguez et al. [12] base their conclusions on cross-validation training performance alone. We use three independent test data sets for model evaluation, in addition to using cross-validation for model training. Using independent test data sets during feature selection for defect prediction is relatively unique to our study.

6 Conclusion

A system with a long operational life is associated with practical issues related to software defect prediction, such as likely variations in the development process, data collection strategy, software quality goals, etc. from one development cycle to another. The software attributes and defect data collected from each source (a given release) provide unique perspectives into quality characteristics of that development project. Thus, mining project data from multiple system releases becomes vital to building a useful defect predictor.

This paper investigates feature selection for software quality modeling. A large case study of software metrics and defect data compares the attribute selection performances of wrapper-based feature ranking techniques (WRT) and our proposed hybrid feature selection (HFS) technique. Four large software measurement data sets obtained from a real-world system are used in the study. Our results reveal that among the two attribute selection techniques, HFS outperforms WRT. A comparison of the two cross-validation strategies indicates that cross-validation by itself is better than cross-validation with risk impact consideration. For the case study, even the removal of over 85% of software metrics did not have a negative effect on the consequent model’s performance.

Further work will include empirical analysis with additional feature selection techniques in the context of data from other software development projects. A comparison of wrapper-based ranking and filter-based ranking would also benefit software quality practitioners.

References