Spring 2014

[Sabbatical Report]

Huanjing Wang

Sabbatical Leave Report

Dr. Huanjing Wang
Department of Computer Science
Ogden College of Science and Engineering
Western Kentucky University

My sabbatical leave was conducted during Spring semester 2014. The leave was successful because it strengthened my research in data mining and software engineering domains and resulted four full-paper publications in peer-reviewed international conferences and one journal paper (to be submitted to a peer-reviewed journal).

The purpose of my sabbatical was to complete two main projects: (1) Investigate the stability and defect prediction model performance of feature selection techniques together on real-world software metrics data and (2) Design a novel, robust, and efficient metric selection method for imbalanced data. The results of my sabbatical were summarized as follows:

1. **January**
   In this month, I worked on designing a novel, robust, and efficient metric selection method for imbalanced software metrics dataset. Identifying a small subset of metrics becomes an essential task before building defect prediction models. After preliminary study, I designed a wrapper-based feature (software metric) subset selection method which uses a classifier to discover which feature subsets are most useful. I then investigated the effect of performance metric within wrapper. To the best of my knowledge, no previous work has examined how the choice of performance metric within wrapper-based feature selection will affect classification performance. In this work, I used five wrapper-based feature selection methods to remove irrelevant and redundant features. These five wrappers vary based on the choice of performance metric used in the model evaluation process. The case study is based on software metrics and defect data (imbalanced) collected from a real world software project, Eclipse. The results demonstrate that Best Arithmetic Mean is the best performance metric used within the wrapper. Moreover, comparing to models built with full datasets, the performances of defect prediction models can be improved when metric subsets are selected through a wrapper subset selector.

   The work (6-page full paper) was published in the Twenty-Sixth International Conference on Software Engineering and Knowledge Engineering (SEKE 2014, July 1-3). This is a peer-reviewed international conference.

2. **February**
   In February, I started to study on the filter-based feature subset selection algorithms. In this work, I compare two filter-based feature subset selection techniques (correlation-based feature selection (CFS) and consistency) along with two search techniques (Best First (BF) and Greedy Stepwise (GS)) on four datasets from a real world software project. Six learners are used to build models with the selected software metrics. Each model is assessed using the area under the Receiver Operating Characteristic curve (AUC). I find that CFS-BF performed best and consistency-GS performed worst. In addition, the model built with the Logistic

Regression (LR) learner performs best in terms of the AUC performance metric. This leads us to recommend the use of CFS-BF to select software metric subsets and the LR learner for building software quality classification models.

The work (5-page full paper) is going to be published in the 20th ISSAT International Conference on Reliability & Quality in Design (RQD 2014, August 7-9). This is a peer-reviewed international conference.

3. **March**

I then focused my work on the similarity of wrapper-based feature (software metric) subset selection techniques. It is not clear how much the different parameters of wrapper-based feature selection (such as the choice of wrapper learner and wrapper performance metric) affect which features are chosen. To study how these two choices can affect the feature selection process, I test five different learners and five different performance metrics within the wrapper and then use our newly proposed Average Pairwise Tanimoto Index (APTI) to evaluate the similarity between techniques which share either a learner or a metric in common. Three software metric datasets from a real-world software project (Eclipse) are used in this study. Results demonstrate that Best Arithmetic Mean and Best Geometric Mean metrics exhibit most similarity regardless of learners; in addition, Overall Accuracy is least similar to each of the other metrics when considered individually with each. The five learners were also found to produce very low amounts of similarity. Thus, results demonstrate that the choice of both learner and performance metric has a major effect on which features are chosen by wrapper-based feature subset selection.

The work (5-page full paper) is also going to be published in the 20th ISSAT International Conference on Reliability & Quality in Design (RQD 2014, August 7-9). This is a peer-reviewed international conference.

4. **April**

In April, I am working on the stability of feature subset selection problem, which feature subset selection methods are stable in the face of changes to the data (here, the addition or removal of instances). I examine twenty-seven feature subset selection methods, including two filter-based techniques and twenty-five wrapper-based techniques (five choices of wrapper learner combined with five choices of wrapper performance metric). I used the Average Tanimoto Index (ATI) as our stability metric, because it is able to compare two feature subsets of different size. All experiments were conducted on three software metric datasets from a real-world software project. Our results show that the Correlation-Based Feature Selection (CFS) approach has the greatest stability overall. All wrapper-based techniques are less stable than CFS. Among the twenty-five wrappers, in general the Naive Bayes learner using either the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) or the Area Under the Precision-Recall Curve (PRC) performance metrics are the most stable wrapper-based approaches.

The work (6-page full paper) is going to be published in the 15th IEEE International Conference on Information Reuse and Integration (IRI 2014, August 13-15). This is a peer-reviewed international conference.

5. **May-June**

During this period, I investigated the stability and defect prediction model performance of feature ranking techniques together on 16 real-world software metrics datasets. In this project, I investigated twenty-five filter-based feature ranking techniques in terms of stability (robustness) and build classification models using five different classifiers. The models were evaluated using the AUC performance metrics. All experiments were conducted on 16 software metrics datasets from real-world software projects. The experimental results demonstrated that enough changes to the dataset can make any feature selection technique unstable, and that using more features increases the stability of most feature selection techniques. One ranker may outperform another in terms of stability, but results may be reversed in terms of classification. In this study, I have found Area Under the ROC Curve (ROC) performs better in both of stability and model performance. Overall, I conclude that the both stability and model performance must be taken in to account when selecting a feature selection technique. The extensive comparative study of feature selection techniques is very unique to this work, especially within the software engineering community. To my knowledge, no such a comprehensive empirical study exists in literature.

Based on the experimental results, I had completed the first draft of a 38 pages paper, titled "A Study on Feature Selection Stability Analysis and Defect Prediction Model Performance in Software Engineering Domain" and plan to submit to a peer-reviewed journal.

Activities:

I attended the 12th International Conference on Machine Learning and Applications which was held in Miami, Florida, in December 2013. In the conference, I met with my collaborator, Dr. Taghi Khoshgoftaar. We talked about the research plan in Spring semester. Due to his busy schedule, we decided to have a half-hour weekly phone meeting on every Monday in Spring instead of site visit. In addition, I also reviewed their research group member's work. The research group member includes PHD students, postdoc, faculty, and other collaborators.

I attended the Twenty-Sixth International Conference on Software Engineering and Knowledge Engineering which was held in July 1-3 in Vancouver, Canada. I gave a 20-minutes presentation in the conference. I chaired two sessions in the Twenty-Sixth International Conference on Software Engineering and Knowledge Engineering. I am also serving as one of the conference organizers for The 14th IEEE International Conference on BioInformatics and BioEngineering which will be held in November 10-12, 2014.

Outcome and Benefits:

- Four full-papers will be published in peer-reviewed international conferences.
- A 38-pages manuscript will be submitted to a peer-reviewed journal.
- As one of best conference papers, my SEKE paper was invited to submit an enhanced version to a IJSEKE (International Journal of Software Engineering and Knowledge. Engineering) Special Issue, Software Quality, Project Management and Data Modeling.
- Based on the preliminary results, I will submit two grant proposals, RCAP and Kentucky Science and Engineering Foundation (KSEF) RDE program.
- Benefit the future project classes (CS 396 and CS 496) and other classes (CS 443, CS 565, and CS 595).