Name: Romana Ishrat

Supervisor's name: Dr. Rafat Parveen

Title: Machine Learning Approach to Object-Oriented Software Quality Estimation -A Decision Tree Perspective

Department: Department of Computer Science
Jamia Millia Islamia, New Delhi-25

Abstract: Findings

The following are the outcomes of my research work:

1. The proposed solution requires the metrics from the software project as soon as they release in the early stage of software development. This metric data along with the fault data is used to build the decision tree based software quality estimation models. For implementation and comprehensive evaluation of decision tree algorithms for quality estimation, datasets collected from NASA's Metric Data Program repository.
2. Prior to classification modeling, the datasets were pre-processed according to the requirements of the algorithms. As software quality estimation model were built with single classifiers and ensemble classifiers, it has been found that the ensembles techniques performed well. The single classifiers have shown weakness in producing good quality results as compared to ensembles. The prediction accuracy of ensembles is found to be higher than single classifiers.
3. Among various ensemble techniques applied for quality estimation, Random Forest appeared as the best technique in terms of prediction accuracy for this case study. Though ensemble techniques produced good classification accuracy but these techniques generate highly complex models. These models are difficult to analyze, as they comprise of dozens of individual models it is not easy to analyze what factors are contributing to the improved decision. These classical decision tree techniques are not robust to imprecise and inconsistent data.

4. To achieve more accurate and robust (to inconsistent data) model, fuzzy decision tree and pattern tree techniques have been implemented for quality estimation. The first technique fuzzy decision tree is based on fuzzy logic and fuzzy set theory incorporated in classical decision trees. FID3.4- a fuzzy decision tree based modeling technique for quality estimation has been proposed. The experimental results demonstrate the effectiveness of the proposed method. The tree models build using small numbers of fuzzy terms are compact. Another important feature of these technique is that the results produced using the partitions i.e. the fuzzy set generated by the FID3.4 pre-processors are more accurate than compared to pre-partitioned sets.
5. This fuzzy decision tree technique is compared with the classical decision tree for the same datasets with small number of metrics as well as for the large number of metrics. It was found that fuzzy decision tree produced consistently higher prediction accuracy than classical decision trees. The FDT are more robust and lead to more accurate model for fault prediction. As the number of input variable increased i.e. for large number of metrics, FDT generate complex model. However pruning can be applied to reduce the complexity. Overall FDT performed in consistently.
6. To achieve less complex model, a new technique- Pattern trees have been used in this thesis, decision tree techniques used different set of metrics from same datasets for model generation. The main reason for that is the development environment and the processes involved in datasets or projects were different from each other. It has been observed that particular metrics cannot be stated as good fault predictor for all the projects. Thus decision tree based software quality estimation models should be adjusted on the individual projects.

7. In this thesis for quality estimation. Pattern tree uses fuzzy aggregations to aggregate from the bottom level to the top. The pattern trees generated using RMSE similarity measure produced best results. The pattern trees produced higher prediction accuracy as compared to all other techniques used. The pattern trees found to be less structural complex. Thus, pattern trees can be used as quality estimation model in object-oriented software. These models can be applied to datasets with small as well as large number of attributes. These trees will definitely help in boosting the quality of any object oriented software system. Fuzzy decision trees can be equally effective in fault prediction.