Abstract

Feature subset selection is the process of choosing a subset of good features with respect to the target concept. A clustering based feature subset selection algorithm has been applied over software defect prediction data sets. Software defect prediction domain has been chosen due to the growing importance of maintaining high reliability and high quality for any software being developed. A software quality prediction model is built using software metrics and defect data collected from a previously developed system release or similar software projects. Upon validation of such a model, it could be used for predicting the fault-proneness of program modules that are currently under development. The proposed clustering based algorithm for feature selection uses minimum spanning tree based method to cluster features. And then the algorithm is applied over four different data sets and its impact is analyzed.

References

- Antonio Arauzo-Azofra, Jose Manuel Benitez and Juan Luis Castro, A Feature Set Measure Based On Relief, RASC2004.
- Roberto Battiti, Using Mutual Information For Selecting Features In Supervised Neural
- M Scherf, W Brauer, Improving RBF Networks By The Feature Selection Approach EUBAFES.
- Langley P., Selection of relevant features in machine learning, In Proceedings of the AAAI.
- Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, IEEE Transactions On Knowledge And Data Engineering Vol:25 No:1 Year 2013.

Index Terms
Computer Science Algorithms

Keywords
Relevant features Redundant Features Minimum spanning tree Tree partition graph based clustering
Software defect prediction
Naïve Bayes classifier
Decision tree classifier