Abstract

Demand for producing quality software has rapidly increased during the last few years. This is leading to an increase in the development of machine learning methods for exploring data sets which can be used in constructing models for predicting quality attributes such as fault proneness, maintenance effort, testing effort, productivity, and reliability. This paper examines and compares logistic regression and six machine learning methods (Artificial neural network, decision tree, support vector machine, cascade correlation network, group method of data handling polynomial method, gene expression programming). These methods are explored empirically to find the effect of static code metrics on the fault proneness of software modules. We use publicly available data set AR1 to analyze and compare the regression and machine learning methods in this study. The performance of the methods is compared by computing the area under the curve using Receiver Operating Characteristic (ROC) analysis. The results show that the area under the curve (measured from the ROC analysis) of model predicted using decision tree modeling is 0.865 and is a better model than the model predicted using regression.
and other machine learning methods. The study shows that the machine learning methods are useful in constructing software quality models.

Reference


[22] Duman, E. Comparison of decision tree algorithms in identifying bank customers who are likely to buy credit cards. Seventh International Baltic Conference on Databases and Information Systems, Kaunas, Lithuania, July 3-6, 2006.


Prediction of Fault-Prone Software Modules using Statistical and Machine Learning Methods


Index Terms

Computer Science
Software Engineering

Keywords

Software quality
Metrics

Logistic regression

Receiver Operating Characteristics (ROC) curve

machine learning