Evaluation of Feature Extraction Methods on Software Cost Estimation

Burak Turhan, Onur Kutlubay, Ayse Bener
Department of Computer Engineering
Bogazici University
34342 Bebek, Istanbul, Turkey
turbanb@boun.edu.tr, kutlubay@cmpe.boun.edu.tr, bener@boun.edu.tr

Abstract

This research investigates the effects of linear and non-linear feature extraction methods on the cost estimation performance. We use Principal Component Analysis (PCA) and Isomap for extracting new features from observed ones and evaluate these methods with support vector regression (SVR) on publicly available datasets. Our results for these datasets indicate there is no significant difference between the performances of these linear and non-linear feature extraction methods.

1. Introduction

In this research we focus on carrying out an empirical study featuring both machine learning based regression analysis and feature extraction algorithms. We employ PCA, which is used in previous researches, and a non-linear method Isomap [2, 3]. We present the empirical evaluation of both methods combined with a standard machine learning algorithm, support vector regression.

2. Experiments and Results

We have used 2 public datasets, ‘Cocomo NASA’ and ‘SDR’, where the latter is compiled by the authors through several software houses in Turkey. All experiments are performed in a 10x10-fold cross-validation framework. The results are reported as the mean and standard deviation of PRED(30) values, which provides a more understandable perspective especially to the business users [1]. We use SVR in order to estimate the cost in the datasets. Linear kernel function is used because of its simplicity, accuracy and speed.

In NASA dataset, the best mean value of PRED(30) is obtained as 59.9%. Results in SDR dataset show that the best mean value of PRED(30) is obtained as 63.3%. Both results are obtained with only 1 feature either using Isomap or PCA. The reason is understood when the eigenvalues are examined. In both cases the first eigenvalue is significantly greater than the others. This means most of the information in data is explained by the first principle component. Performances on both datasets are given in Figure 1.

3. Conclusions

We evaluated a linear and a non-linear feature extraction method, PCA and Isomap respectively, for software cost estimation performance. We observe that there is no evidence in favor of linear or non-linear approaches for these datasets.

References

