Fit Data Selection for Software Effort Estimation Models

Koji Toda  
NARA Institute of Science and Technology  
Kansai Science City, 630-0192 Japan  
koji-to@is.naist.jp

Akito Monden  
NARA Institute of Science and Technology  
Kansai Science City, 630-0192 Japan  
akito-m@is.naist.jp

Ken-ichi Matsumoto  
NARA Institute of Science and Technology  
Kansai Science City, 630-0192 Japan  
matumoto@is.naist.jp

ABSTRACT
To construct a better multivariate regression model for software effort estimation, this paper proposes a method to select projects as a fit data from a given project data set based on estimation target's features. While regression models were often constructed from all available project data, this paper showed the necessity of fit data selection, and showed that the proposed method is one of the effective and systematic means to do the selection.

Categories and Subject Descriptors
D.2.9 [Software Engineering]: Management – Cost estimation, K.6.1 [Management of Computing and Information]: Project and People Management – Strategic information system planning

General Terms
Management, measurement, economics

Keywords
Effort estimation, Multivariate regression.

1. BACKGROUND
Multivariate regression modeling is a simple but widely used method for software effort estimation[1]. It is because various tools for model construction are available, various variable selection methods such as forward-backward stepwise selection are available, and it also gives reasonable estimation performance. However, few researches have been made on how to prepare a proper fit dataset for the model construction. Indeed, regression model’s estimation performance greatly depends on the fit data. Generally speaking, to construct a better model, fit data should contain similar projects whose development environments, processes or application domains are the same.

2. PROPOSED METHOD
We propose a systematic method to select fit data for regression models. Our basic idea is to select candidates of fit data sets as many as possible each having at least one similar feature with the target project, build regression models using each candidate, and pickup one model (and its fit data candidate) having the best fit to the data. The procedure is shown in Figure 1.

Copyright is held by the author/owner(s).
ESEM '08, October 9–10, 2008, Kaiserslautern, Germany.

[STEP 1] Candidates of fit data sets are selected from a given project data set (containing all past projects). Each candidate contains projects that have one of features of the target project. If the target project has three features (COBOL, Stand Alone and Re-development), each fit data candidate contains COBOL projects, Stand Alone projects and Re-development projects. If the size of a candidate (i.e. the number projects involved) is smaller than the minimum fit data size, we remove this candidate from the candidates.

[STEP 2] Construct regression models using each fit data candidate and calculate the goodness of fit (e.g., residual mean square or adjusted R-square) for each model.

[STEP 3] Select the best model and its fit data candidate based on the goodness of fit. If “COBOL language model” was the best of three candidates, this model and its fit data candidate is selected.
These steps are simplified ones where all project features are given as categorical variables. However, project data sets usually contain quantitative variables (e.g. Function Points and Project Elapsed Time), therefore, we translate the quantitative variables into categorical ones by partitioning each quantitative variable by a given threshold in the fit data candidate selection step.

3. EXPERIMENT
We evaluated the proposed method using a subset of the ISBSG dataset collected by International Software Benchmarking Standard Group (ISBSG)[2]. The data subset we used contains 109 projects and 10 variables. When building regression models, we used the “Summary Work Effort” as an objective variable and we used nine predictor variables, Function Points, Count Approach, Summary Work Effort, Effort Plan, Effort Specify, Development Type, Architecture, Primary Programming Language, Recording Method and Resource Level. Details of these variables are shown in [3]. All these predictor variables can be measured by the design phase. In the experiment, we decided the minimum fit data size to be 10. As the goodness of fit, we tested two criteria: (1) residual mean squared (RMS) and (2) adjusted R-squared (adj.R^2). We used the stepwise multivariate regression analysis as a modeling technique. We split the initial dataset into fit and test randomly each having the same size. We repeated this operation 10 times.

The result is shown in Table 1. “Conventional” means that a regression model was built using all available projects (i.e. without fit data selection.) The proposed methods showed better accuracy than the conventional in both criteria. In the proposed method, using RMS as criteria was better than using adj.R^2.

4. SUMMARY
This paper proposed a method to select projects as a fit data from a given project data set based on estimation target's features. The main advantage of the proposed method is that it can be systematically (automatically) used without detailed knowledge about the given project dataset.

5. References

Acknowledgment
This work was conducted as a part of the StagE Project, the Development of Next Generation IT Infrastructure, supported by Ministry of Education, Culture, Sports, Science and Technology.