Cuadrado-Gallego JJ, Rodríguez D, Sicilia MA *et al.* Software project effort estimation based on multiple parametric models generated through data clustering. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 22(3): 371~378 May 2007

Software Project Effort Estimation Based on Multiple Parametric Models Generated Through Data Clustering

Juan J. Cuadrado Gallego^1, Daniel Rodríguez^1, Miguel Ángel Sicilia^1, Miguel Garre Rubio^1
and Angel García Crespo^2

^{1}*Department of Computer Science, The University of Alcalá, Alcalá, Spain*

^{2}*Department of Computer Science, Carlos III University, Madrid, Spain*

E-mail: {jjcg, daniel.rodriguezg, msicila, miguel.garre}@uah.es; acrespo@ia.uc3m.es

Received March 15, 2006; revised February 15, 2007.

Abstract Parametric software effort estimation models usually consists of only a single mathematical relationship. With the advent of software repositories containing data from heterogeneous projects, these types of models suffer from poor adjustment and predictive accuracy. One possible way to alleviate this problem is the use of a set of mathematical equations obtained through dividing of the historical project datasets according to different parameters into subdatasets called partitions. In turn, partitions are divided into clusters that serve as a tool for more accurate models. In this paper, we describe the process, tool and results of such approach through a case study using a publicly available repository, ISBSG. Results suggest the adequacy of the technique as an extension of existing single-expression models without making the estimation process much more complex that uses a single estimation model. A tool to support the process is also presented.

Keywords software engineering, software measurement, effort estimation, clustering

1 Introduction

Parametric estimation techniques are nowadays widely used to measure and/or estimate the cost associated to software development[1]. The Parametric Estimating Handbook[2] defines parametric estimation as “a technique employing one or more cost estimating relationships and associated mathematical relationships and logic”. Parametric techniques are based on identifying variables that obtain numerical estimates from main input variables that are known to affect the effort or time spent in development.

One important aspect of the process of deriving models from databases is that of the heterogeneity of data. A measure of such heterogeneity is heteroscedasticity, i.e., non–uniform variance. It is well-known that heteroscedasticity is a problem affecting data sets that combine data from heterogeneous sources[3]. As a result, when using such software engineering databases, traditional application of regression equations to derive a single mathematical model results in poor adjustment to data and subsequent potential high deviations. This is due to the fact that a single model cannot capture the diversity of distribution of different segments of the database points. As an illustrative example, the straightforward application of a standard least squares regression algorithm to the points used in the reality tool of the ISBSG 8 database distribution results in measures of $MMRE = 2.8$ and $Pred(0.3) = 23\%$ (these measures are introduced later), which are poor figures of predictive quality.

The use of clustering techniques has been described as a solution to provide more realism to parametric models by decomposing the model in a number of sub–models, that are used for project estimation[4] with improved accuracy when compared with single models. The resulting predictive schemes have been called *segmented* models. One of the principal benefits of this kind of techniques is the fact that the search of segmented models satisfying some pre-established quality conditions can be automated through existing clustering methods.

The rest of this paper is structured as follows. Section 2 describes related work in segmented models for software estimation. The process for software project estimation using clustering techniques and associated tool is described in Section 3. Then, Section 4 reports on empirical work performed to validate the process using a publicly available repository. Finally, conclusions and future work are discussed in Section 5.

2 Related Work

Shepperd *et al.*[5] classify estimation and prediction techniques into three main categories: (i) expert judgement; (ii) algorithmic models; and (iii) machine learning. This work focuses on combining clustering as machine learning technique with classical regression models. The use of different clustering approaches has already been applied to several aspects of software management, including software estimation, software quality and software metrics.

Xu and Khoshgoftaar[6] use the fuzzy c-means algorithm for variable, the partitioning of the data into a number of clusters based on experiences.

Pedrycz and Succi[7] also use fuzzy c-means as a tool to derive prototypes related to software code measurements. Dick *et al.*[8] use the same algorithm for a simi-

Regular Paper
This work is supported by the Spanish Ministry of Science and Technology under Grant No. CICYT TIN2004-06689-C03.