The Journal of Systems and Software 99 (2015) 120–134

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Integrating non-parametric models with linear components for producing software cost estimations

Nikolaos Mittas [a], Efi Papatheocharous [b,c], Lefteris Angelis [a,*], Andreas S. Andreou [d]

[a] Department of Informatics, Aristotle University of Thessaloniki, Greece
[b] Department of Computer Science, University of Cyprus, Cyprus
[c] Swedish Institute of Computer Science (SICS), SE-16429 Kista, Sweden
[d] Department of Electrical Engineering/Computer Engineering and Informatics, Cyprus University of Technology, Cyprus

ARTICLE INFO

Article history:
Received 18 February 2014
Received in revised form 29 August 2014
Accepted 19 September 2014
Available online 30 September 2014

Keywords:
Software cost estimation
Semi-parametric models

ABSTRACT

A long-lasting endeavor in the area of software project management is minimizing the risks caused by under- or over-estimations of the overall effort required to build new software systems. Deciding which method to use for achieving accurate cost estimations among the many methods proposed in the relevant literature is a significant issue for project managers. This paper investigates whether it is possible to improve the accuracy of estimations produced by popular non-parametric techniques by coupling them with a linear component, thus producing a new set of techniques called semi-parametric models (SPMs). The non-parametric models examined in this work include estimation by analogy (EbA), artificial neural networks (ANN), support vector machines (SVM) and locally weighted regression (LOESS). Our experimentation shows that the estimation ability of SPMs is superior to their non-parametric counterparts, especially in cases where both a linear and non-linear relationship exists between software effort and the related cost drivers. The proposed approach is empirically validated through a statistical framework which uses multiple comparisons to rank and cluster the models examined in non-overlapping groups performing significantly different.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Software cost estimation (SCE) entails the process to create and apply an appropriate model to estimate the resources required to build a fully functional software system. One of the main resources estimated is the human effort. A category of widely used estimation models uses a formula to predict the person-months required to develop the system. However the complex, multifaceted and intangible nature of software, as well as the facts that each system is unique and that the development process followed is usually project- and organization-specific, make the process of SCE a tough endeavor. The way to predict software effort is usually empirically oriented and is mostly based on variables expressing the size of software, measured with metrics such as function points (FP), which commonly describe software functions in terms of inputs, outputs, inquiries, files and external interfaces, complexity and technologies utilized throughout the development process (Pressman, 2000).

In empirical effort estimation models, the relationship between effort and the functionality of a software system (expressed for example in FP) is frequently characterized as non-linear (Pressman, 2000). Typical lines of code- or FP-oriented models in the literature, such as the Walston–Felix (Walston and Felix, 1977), Bailey–Basili (Bailey and Basili, 1981), Boehm (Boehm, 1981; Boehm et al., 2000), Doty (Doty et al., 1977a,b), Kemerer (Kemerer, 1987) and COPLIMO (In et al., 2006), make use of non-linear formulas. Moreover, it is often met in projects with high diversification in terms of quality, architectural design or complexity of implementation and validation, that effort exhibits relatively high values even for small to medium sized projects. This is an indication that the combined effect of other variables – project characteristics – on the effort is complicated and that pure linear methods for effort estimation may not be always optimal, even though they have been extensively used by researchers (Jørgensen and Shepperd, 2007).

A significant amount of research in SCE during the past years has been concentrated on comparing and evaluating different techniques for eventually improving project planning, resource allocation and product quality (Jørgensen and Shepperd, 2007). Moreover, researchers are often required to systematically gather

* Corresponding author. Tel.: +30 2310998230; fax: +30 2310998230.
E-mail addresses: nmittas@csd.auth.gr (N. Mittas), efi.papatheocharous@cs.ucy.ac.cy (E. Papatheocharous), lef@csd.auth.gr (L. Angelis), andreas.andreou@cut.ac.cy (A.S. Andreou).

http://dx.doi.org/10.1016/j.jss.2014.09.025
0164-1212/© 2014 Elsevier Inc. All rights reserved.

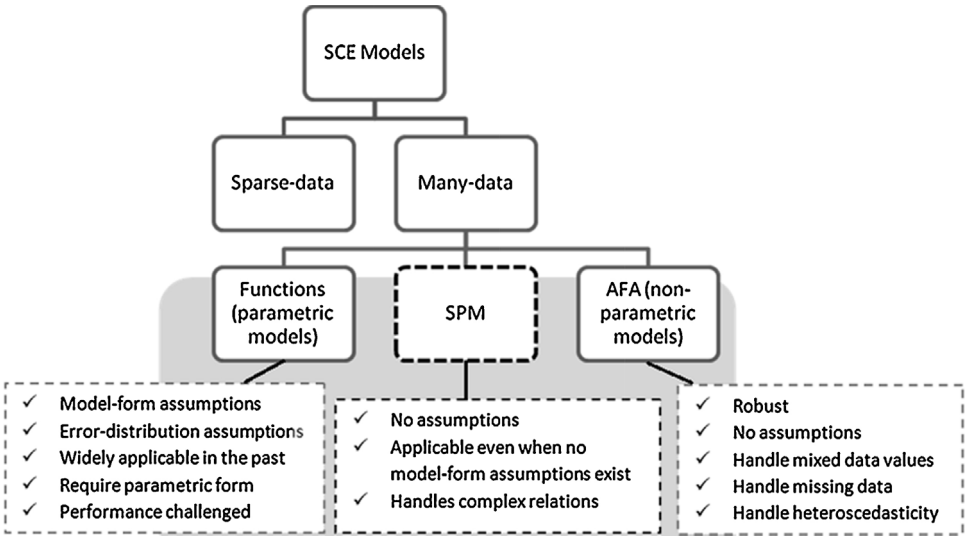

Fig. 1. SCE models categorization schema and models comparison.

and "purify" software data before any analysis action is performed by resolving problems with outliers, dispersed or missing values, implicit definitions, erroneous, tacit or latent software measurements, as well as coping with a data-starved and noisy domain (Song and Shepperd, 2011). The main focus of previous work on the topic has been to propose and methodologically apply new approaches, and present their merits over other methods. Despite the increasing competitive landscape no method or technique may be considered yet as the "*silver bullet*". The following research question posed long ago still remains unanswered today: "*Which prediction method is considered the best under specific circumstances?*" (Shepperd and Kadoda, 2001).

Having in mind the line of research stemming from the previous statement, our work investigates the following argument: "*Selecting the best performing prediction method in SCE among various alternatives is not of primal importance. Rather, any method performing similarly to the rest alternative methods may be considered as a "good enough" SCE approach*". Applying to two representative datasets a variety of prediction methods, including a new type of models, the semi-parametric models (SPMs), we show that selecting a single "best" method in a particular situation may not be of primal importance, but offering a range of viable alternatives that perform equally well is more beneficial to project managers. The validation of the arguments made in support of the above is based on carrying out systematic experiments and performing statistical hypothesis testing. Furthermore, the aim is to show that in certain cases, depending on the nature of data and the relationships between cost and independent variables, SPMs exhibit better or at least comparable performance than their non-parametric counterpart, or the pure linear model. The approach followed in this work is to fully exploit the benefits offered by SPMs by first handling the non-linear part of the observed relationship using a non-parametric model and then estimating the linear part with a parametric model.

Since the objective of many studies in SCE was the comparison between parametric and non-parametric methods, this general form of SPMs which combines them in a hybrid model has not been yet thoroughly investigated or used in a systematic manner. Mittas and Angelis (2008b, 2010) experimented with a combination of least squares (LS) regression and estimation by analogy (EbA) into a single model, namely LSEbA. The primary difference from the previous work of Mittas and Angelis is that in the present study our goal is to expand and demonstrate fully the potentials of SPMs in the context of SCE through the utilization of alternative non-parametric models. Additionally, we use the notion of multiple comparisons through a statistical framework (Mittas and Angelis, 2013) to identify non-overlapping groups of models with significantly different performances. In this way, we utilize a systematic way to validate our results by investigating whether the competing models differ significantly in terms of prediction accuracy.

The remainder of the paper is organized as follows: Section 2 presents a short review of SCE related categorizations of models and describes the research questions and contribution of the present work. Section 3 concentrates on the specific non-parametric techniques employed in the experimentation part, i.e., *estimation by analogy* (EbA), *locally weighted regression* (LOESS), *artificial neural networks* (ANN) and *support vector machines* (SVM). Section 4 describes the proposed SPMs and explains how the alternative non-parametric techniques mentioned above were incorporated. Section 5 provides the details of the experimental setup, the validation procedure followed, the accuracy measures used and the statistical tests performed. Section 6 summarizes the experimental results and finally, Section 7 provides the discussion, presents possible threats to validity and outlines future work.

2. SCE models categorization and contribution

2.1. SCE models categorization

Having in mind the typical categorization schema of Myrtveit et al. (2005), SCE models may be distinguished into two general classes, the *sparse-data* and the *many-data* methods. The former methods require few or no historical data, such as expert judgment techniques, whereas the *many-data* methods are based on available datasets and may be subdivided into *functions* with a predefined mathematical formula describing explicitly the relationship between the cost and independent variables (for example ordinary least squares regression) and *arbitrary function approximators* making no assumption on the underlying relationship (for example estimation by analogy and machine learning techniques). The semi-parametric models utilized in our work may be considered an extension to the models described in Myrtveit et al. (2005). A comparison among the models is provided in Fig. 1.

Data-driven models (either *sparse-data* or *many-data* models according to the categorization schema of Myrtveit et al. (2005)) have been considered the most popular approaches in the topic of SCE. The majority of the journal papers found in the comprehensive study of Jørgensen and Shepperd (2007) evaluate such

Download English Version:

https://daneshyari.com/en/article/461041

Download Persian Version:

https://daneshyari.com/article/461041

Daneshyari.com