The Journal of Systems and Software 82 (2009) 241–252

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

A study of project selection and feature weighting for analogy based software cost estimation

Y.F. Li *, M. Xie, T.N. Goh

Department of Industrial and Systems Engineering, National University of Singapore, Singapore 119 260, Singapore

ARTICLE INFO

Article history:
Received 14 May 2007
Received in revised form 4 June 2008
Accepted 4 June 2008
Available online 17 June 2008

Keywords:
Software cost estimation
Analogy based estimation
Feature weighting
Project selection
Genetic algorithm
Artificial datasets

ABSTRACT

A number of software cost estimation methods have been presented in literature over the past decades. Analogy based estimation (ABE), which is essentially a case based reasoning (CBR) approach, is one of the most popular techniques. In order to improve the performance of ABE, many previous studies proposed effective approaches to optimize the weights of the project features (feature weighting) in its similarity function. However, ABE is still criticized for the low prediction accuracy, the large memory requirement, and the expensive computation cost. To alleviate these drawbacks, in this paper we propose the project selection technique for ABE (PSABE) which reduces the whole project base into a small subset that consist only of representative projects. Moreover, PSABE is combined with the feature weighting to form FWPS-ABE for a further improvement of ABE. The proposed methods are validated on four datasets (two real-world sets and two artificial sets) and compared with conventional ABE, feature weighted ABE (FWABE), and machine learning methods. The promising results indicate that project selection technique could significantly improve analogy based models for software cost estimation.

1. Introduction

Software cost estimation is critical for the success of software project management. It affects almost management activities including resource allocation, project bidding, and project planning (Pendharkar et al., 2005; Auer et al., 2006; Jorgensen and Shepperd, 2007). The importance of accurate estimation has led to extensive research efforts to software cost estimation methods. From a comprehensive review (Boehm et al., 2000), these methods could be classified into the following six categories: *parametric models* including COCOMO (Boehm, 1981; Huang et al., 2007), SLIM (Putnam and Myers, 1992), and SEER-SEM (Jensen, 1983); *expert judgment* including Delphi technique (Helmer, 1966) and work breakdown structure based methods (Tausworthe, 1980; Jorgensen, 2004); *learning oriented techniques* including machine learning methods (Heiat, 2002; Shin and Goel, 2000; Oliveira, 2006) and analogy based estimation (Shepperd and Schofield, 1997; Auer et al., 2006; Huang and Chiu, 2006); *regression based methods* including ordinary least square regression (Mendes et al., 2005; Costagliola et al., 2005) and robust regression (Miyazaki et al., 1994); *dynamics based models* (Madachy, 1994); *composite methods* (Chulani et al., 1999; MacDonell and Shepperd, 2003).

The analogy based estimation (ABE) which is essentially a case-based reasoning (CBR) approach (Shepperd and Schofield, 1997) was first proposed by Sternberg (1977). Due to its conceptual simplicity and empirical competitiveness, ABE has been extensively studied and applied (Shepperd and Schofield, 1997; Walkerden and Jeffery, 1999; Angelis and Stamelos, 2000; Mendes et al., 2003; Auer et al., 2006; Huang and Chiu, 2006; Chiu and Huang, 2007). The basic idea of ABE is simple: when provided a new project for estimation, compare it with historical projects to retrieve the most similar projects which are then used to predict the cost of new project. Generally, the ABE (or CBR) consists of four parts: a historical project dataset, a similarity function, a solution function and the associated retrieval rules (Kolodner, 1993). One of the associated central parts in ABE is the similarity function, which measures the level of similarity between two different projects. Since each project feature (or cost driver) has one position in the similarity function and therefore largely determines which historical projects should be retrieved for final prediction, there are several approaches focusing on searching the appropriate weight of each feature, such as Shepperd and Schofield (1997), Walkerden and Jeffery (1999), Angelis and Stamelos (2000), Mendes et al. (2003), Auer et al. (2006), Huang and Chiu (2006).

However, some difficulties are still confronted by ABE methods. Such as the non-normal characteristics (includes skewness, heteroscedasticity and excessive outliers) of the software engineering datasets (Pickard et al., 2001) and the increasing sizes of the datasets (Shepperd and Kadoda, 2001). The large and non-normal datasets always lead ABE methods to low prediction accuracy and high computational expense (Huang et al., 2002). To alleviate these drawbacks, many research works in the CBR literature (Lipowezky,

* Corresponding author. Tel.: +65 83442816.
E-mail address: liyanfu@nus.edu.sg (Y.F. Li).

0164-1212/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2008.06.001

1998; Babu and Murty, 2001; Huang et al., 2002) have been devoted to the case selection technique. The objective of case selection (CS) is to identify and remove redundant and noisy projects. By reducing the whole project base into a smaller subset that consist only of representative projects, the CS could save the computing time searching for most similar projects and produce quality prediction results. Moreover, the simultaneous optimization of feature weighting and case selection in CBR has been investigated in several studies (Kuncheva and Jain, 1999; Rozsypal and Kubat, 2003; Ahn et al., 2006) and significant improvements are reported from these studies.

From the discussion above, it is worthwhile to investigate case selection technique in the context of analogy based software cost estimation. In this study, we propose genetic algorithm for project selection for ABE (PSABE) and the simultaneous optimization of feature weights and project selection for ABE (FWPSABE). The proposed two techniques are compared against the feature weighting ABE (ABE), the conventional ABE and other popular cost estimation methods including ANN, RBF, SVM and CART. For the consistency of terminology, in rest of this paper we refer the case selection as project selection for ABE.

To compare different estimation methods, the empirical validation is very crucial. This has led to many studies use various real datasets to conduct comparisons of different cost estimation methods. However most published real datasets are relatively small (Mair et al., 2005) and the small real dataset could be problematic if we would like to show the significant differences between the estimation methods. Another drawback of the real world datasets is that the true properties of them may not be fully known. The artificially generated datasets (Pickard et al., 2001; Shepperd and Kadoda, 2001; Foss et al., 2003; Myrtveit et al., 2005) with known characteristics provide a feasible way to the above problems. Thus, we generate two artificial datasets and select two well known real-world datasets for controlled experiments.

The rest of this paper is organized as follows: Section 2 presents a brief overview on the conventional ABE method. In Section 3, the general framework of feature weight and project selection system for ABE is described. Section 4 presents the real world datasets and the experiments design. In Section 5, the results on two real world data sets are summarized and analyzed. In Section 6, two artificial datasets are generated, experiments are conducted on these two datasets, and results are summarized and analyzed. The final section presents the conclusion, and future works.

2. Overview on analogy based cost estimation

Analogy based method is a pure form of case based reasoning (CBR) with no expert used. Generally, ABE model comprises of four components: a historical dataset, a similarity function, a solution function and the associated retrieval rules (Kolodner, 1993). The ABE system process also consists of four stages:

1. Collect the past projects' information and prepare the historical dataset.
2. Select new project's relevant features such as function points (FP) and lines of source code (LOC), which are also collected for past projects.
3. Retrieval the past projects, estimate the similarities between new project and the past projects, and find the most similar past projects. The commonly used similarities are functions of weighted Euclidean distance and the weighted Manhattan distance.
4. Predict the cost of the new project from the chosen analogues by the solution function. Generally the un-weighted average is used as solution function.

The historical dataset which keeps all information of past projects is a key component in ABE system. However, it often contains noisy or redundant projects. By reducing the whole historical dataset into a smaller but more representative subset, the project selection technique positively affects the conventional ABE systems. First, it reduces the search space, thus more computing resources searching for most similar projects are saved. Secondly, it also produces quality predictions because it may eliminate noise in the historical dataset.

In the following sections, other components of ABE system including similar function, the number of most similar projects, and solution function are presented.

2.1. Similarity function

The *similarity function* measures the level of similarity between projects. Among different types of similarity functions, euclidean similarity (ES) and manhattan similarity (MS) based similarities are widely accepted (ES: Shepperd and Schofield, 1997. MS: Chiu and Huang, 2007). The Euclidean similarity is based on the Euclidean distance between two projects:

$$\mathrm{Sim}(p,p\prime) = 1\Big/\left[\sqrt{\sum_{i=1}^{n} w_i \mathrm{Dis}(f_i, f\prime_i)} + \delta\right] \qquad \delta = 0.0001$$
$$\mathrm{Dis}(f_i,f_i') = \begin{cases} (f_i - f_i')^2, & \text{if } f_i \text{ and } f_i' \text{ are numeric or ordinal} \\ 1 & \text{if } f_i \text{ and } f_i' \text{ are nominal and } f_i = f_i' \\ 0 & \text{if } f_i \text{ and } f_i' \text{ are nominal and } f_i \neq f_i' \end{cases} \tag{1}$$

where p and p' denote the projects, f_i and f_i' denote the ith feature value of their corresponding projects, $w_i = [0,1]$ is the weight of the ith feature, $\delta = 0.0001$ is a small constant to prevent the situation the denominator equals 0, and n is the total number of features.

The Manhattan similarity is based on the Manhattan distance which is the sum of the absolute distances for each pair of features

$$\mathrm{Sim}(p,p\prime) = 1\Big/\left[\sum_{i=1}^{n} w_i \mathrm{Dis}(f_i, f_i') + \delta\right] \qquad \delta = 0.0001$$
$$\mathrm{Dis}(f_i,f_i') = \begin{cases} |f_i - f_i'| & \text{if } f_i \text{ and } f_i' \text{ are numeric or ordinal} \\ 1 & \text{if } f_i \text{ and } f_i' \text{ are nominal and } f_i = f_i' \\ 0 & \text{if } f_i \text{ and } f_i' \text{ are nominal and } f_i \neq f_i' \end{cases} \tag{2}$$

An important issue in the similarity functions is how to assign appropriate weight w_i to each feature pair, because each feature may have different relevance to the project cost. In the literature, several approaches were focusing on this topic: Shepperd and Schofield (1997) set each weight to be either 1 or 0 then apply a brute-force approach choosing optimal weights; Auer et al. (2006) extent Shepperd and Schofield's approach to the flexible extensive search method. Walkerden and Jeffery (1999) use human judgment to determine the feature weights; Angelis and Stamelos (2000) choose a value generated from statistical analysis as the feature weights. More recently, Huang and Chiu (2006) propose the genetic algorithm to optimize feature weights.

2.2. K number of similar projects

This parameter refers to the K number of most similar projects that is close to the project being estimated. Some studies suggested $K = 1$ (Walkerden and Jeffery, 1999; Auer et al., 2006; Chiu and Huang, 2007). However, we sets $K = \{1,2,3,4,5\}$ since many studies recommend K equals to two or three (Shepperd and Schofield, 1997; Mendes et al., 2003; Jorgensen et al., 2003; Huang and Chiu,