Information and Software Technology 58 (2015) 206–230

Contents lists available at ScienceDirect

Information and Software Technology

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

Analogy-based software development effort estimation: A systematic mapping and review

Ali Idri [a,*], Fatima azzahra Amazal [a], Alain Abran [b]

[a] Software Projects Management Research Team, ENSIAS, Mohammed V Souissi University, Madinate Al Irfane, 10100 Rabat, Morocco
[b] Department of Software Engineering, Ecole de Technologie Supérieure, Montréal H3C IK3, Canada

ARTICLE INFO

Article history:
Received 30 July 2013
Received in revised form 28 July 2014
Accepted 29 July 2014
Available online 7 August 2014

Keywords:
Mapping study
Systematic literature review
Software development effort estimation
Analogy
Case-based reasoning

ABSTRACT

Context: Analogy-based Software development Effort Estimation (ASEE) techniques have gained considerable attention from the software engineering community. However, existing systematic map and review studies on software development effort prediction have not investigated in depth several issues of ASEE techniques, to the exception of comparisons with other types of estimation techniques.
Objective: The objective of this research is twofold: (1) to classify ASEE studies which primary goal is to propose new or modified ASEE techniques according to five criteria: research approach, contribution type, techniques used in combination with ASEE methods, and ASEE steps, as well as identifying publication channels and trends and (2) to analyze these studies from five perspectives: estimation accuracy, accuracy comparison, estimation context, impact of the techniques used in combination with ASEE methods, and ASEE tools.
Method: We performed a systematic mapping of studies for which the primary goal is to develop or to improve ASEE techniques published in the period 1990–2012, and reviewed them based on an automated search of four electronic databases.
Results: In total, we identified 65 studies published between 1990 and 2012, and classified them based on our predefined classification criteria. The mapping study revealed that most researchers focus on addressing problems related to the first step of an ASEE process, that is, feature and case subset selection. The results of our detailed analysis show that ASEE methods outperform the eight techniques with which they were compared, and tend to yield acceptable results especially when combining ASEE techniques with Fuzzy Logic (FL) or Genetic Algorithms (GA).
Conclusion: Based on the findings of this study, the use of other techniques such FL and GA in combination with an ASEE method is promising to generate more accurate estimates. However, the use of ASEE techniques by practitioners is still limited: developing more ASEE tools may facilitate the application of these techniques and then lead to increasing the use of ASEE techniques in industry.

Contents

* Corresponding author. Tel.: +212 661390943.
E-mail address: idri.ali123@gmail.com (A. Idri).

http://dx.doi.org/10.1016/j.infsof.2014.07.013
0950-5849/

1. Introduction

Estimating the cost of a software project in terms of effort is one of the most important activities in software project management. This is because rigorous planning, monitoring, and control of the project are not feasible if the estimates of software development cost are highly inaccurate. Unfortunately, the industry is plagued with unreliable estimates, and no effort estimation model has proven to be consistently successful at predicting software project effort in all situations [1]. Researchers in the software engineering community continue to propose new models to achieve effort prediction accuracy. Jørgensen and Shepperd [2] conducted a systematic review in which they identified up to 11 estimation approaches in 304 selected journal papers. These approaches fall into two major categories: parametric models, which are derived from the statistical and/or numerical analysis of historical project data, and machine learning (ML) models, which are based on a set of artificial intelligence techniques such as artificial neural networks (ANN), genetic algorithms (GA), analogy-based or case-based reasoning (CBR), decision trees, and genetic programming.

ML techniques are gaining increasing attention in software effort estimation research, as they can model the complex relationship between effort and software attributes (cost drivers), especially when this relationship is not linear and does not seem to have any predetermined form. Recently, Wen et al. [1] carried out a systematic literature review in which they identified eight types of ML techniques. ASEE and ANN-based effort estimation techniques are the most frequently used of these, 37% and 26% of the time respectively. Their SLR also showed that the CBR and ANN are more accurate in terms of the arithmetic mean of Preds(25) and arithmetic mean of MMREs, obtained from selected studies, than the other ML techniques (mPred(25) = 46% and mMMRE = 51% for CBR-based studies, and mPred(25) = 64% and mMMRE = 37% for ANN-based studies). This confirms the results of the study carried out in [2]: the use of ASEE techniques instead of other ML techniques (ANN, Classification and Regression Trees) is increasing over time (10% instead of 7% for ANN and 5% for classification and regression tress -CRT until the year 2004). Moreover, instead of ANNs which are often considered as black-box, ASEE techniques are claimed to be easily understood by users, as they are similar to human reasoning by analogy [1] (see Table C.13 of [1] in which more than 15 references are supporting this affirmation). Nevertheless, Section 4.3 discusses the numerous hard decisions and limitations that prevent ASEE techniques to be easily used in a given context.

In spite of these advantages, ASEE techniques are still limited by their inability to correctly handle categorical attributes (measured on a nominal or ordinal scale). Indeed, the commonly used way to assess the similarity between two software projects described by nominal attributes is to use the overlap measure which assigns a similarity of 1 if the values are identical and a similarity of 0 if the values are not identical [3–6]. For ordinal attributes, most studies map the ordinal values to their ranking numbers (or positions) and then assess the similarity using some arithmetic operations (addition, subtraction, etc.) that are not meaningful according to measurement theory [4–7]. Furthermore, inconsistent results have been reported regarding their accuracy, compared with other effort estimation techniques, both ML and non ML. For example, some studies [3,8–10] claim that ASEE techniques outperform regression models, while the results of others [11,12] indicate that regression

Table 1
Mapping study questions.

ID.	Mapping question	Main motivation
MQ1	Which (and how many) sources include papers on ASEE?	To provide effort estimation researchers with a list of relevant studies on ASEE
MQ2	What are the most frequently applied research approaches in the ASEE field, and how has this changed over time?	To identify research approaches and their trends over time in the ASEE literature
MQ3	What are the main types of contribution of ASEE studies?	To identify the different types of contribution of ASEE studies
MQ4	Which of the reported techniques are used the most frequently in combination with ASEE techniques?	To identify the techniques used in combination with analogy to improve the estimation accuracy of ASEE techniques
MQ5	Have the various steps of the analogy procedure received the same amount of attention on the part of researchers?	To classify the various steps of the analogy procedure based on the amount of attention they have received from researchers