The Journal of Systems and Software 103 (2015) 36–52

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

An empirical evaluation of ensemble adjustment methods for analogy-based effort estimation

Mohammad Azzeh [a,*], Ali Bou Nassif [b], Leandro L. Minku [c]

[a] Department of Software Engineering, Applied Science University, Amman 166, Jordan
[b] Department of Electrical and Computer Engineering, University of Western Ontario, 1151 Richmond St, London, ON, N6A 3K7, Canada
[c] School of Computer Science, The University of Birmingham, Office 244, Edgbaston, Birmingham B15 2TT, UK

ARTICLE INFO

Article history:
Received 16 July 2014
Revised 14 January 2015
Accepted 16 January 2015
Available online 22 January 2015

Keywords:
Ensemble learning
Analogy based estimation
Adjustment methods

ABSTRACT

Context: Effort adjustment is an essential part of analogy-based effort estimation, used to tune and adapt nearest analogies in order to produce more accurate estimations. Currently, there are plenty of adjustment methods proposed in literature, but there is no consensus on which method produces more accurate estimates and under which settings.
Objective: This paper investigates the potential of ensemble learning for variants of adjustment methods used in analogy-based effort estimation. The number *k* of analogies to be used is also investigated.
Method: We perform a large scale comparison study where many ensembles constructed from *n* out of 40 possible valid variants of adjustment methods are applied to eight datasets. The performance of each method was evaluated based on standardized accuracy and effect size.
Results: The results have been subjected to statistical significance testing, and show reasonable significant improvements on the predictive performance where ensemble methods are applied.
Conclusion: Our conclusions suggest that ensembles of adjustment methods can work well and achieve good performance, even though they are not always superior to single methods. We also recommend constructing ensembles from only linear adjustment methods, as they have shown better performance and were frequently ranked higher.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Analogy-based effort estimation (EBA) is a commonly used method for predicting the most likely software development effort (Angelis and Stamelos, 2000; Auer et al., 2006). It is based on the assumption that software projects with similar characteristics have similar effort values (Keung et al., 2008; Kocaguneli et al., 2012; Shepperd and Kadoda, 2001; Mittas et al., 2008). Reusing efforts of the selected analogies directly without considering revision is less accurate (Azzeh, 2012; Kirsopp et al., 2003). Therefore, an adjustment technique should be applied to calibrate and tune the generated estimate based on the characteristics of both source and target projects. The goal of using adjustment is to minimize differences between a new project and its nearest analogies, and therefore increase EBA's accuracy.

Many adjustment methods have been proposed in the past 20 years (Azzeh, 2012), but as of yet, there is no univocal conclusion as to which adjustment method integrated with EBA produces the most accurate predictions, and under which settings. However, Azzeh's (2012) replication study reported an important insight. He showed that, even though no particular method is significantly superior to others, guidelines can be given to explain how and under what conditions to use each of the existing methods. It has been concluded that each method favors: (1) different feature set, (2) different number of nearest analogies (*k*) and (3) specific type of features (i.e. continuous or categorical). Moreover, the results from that study showed that some adjustment methods cannot outperform conventional EBA over some datasets. For these reasons it was difficult to recommend a particular method against others over a particular dataset. We believe that it would be more promising to combine existing methods in order to benefit from their individual advantages (and consequently improve the accuracy of adjusted EBA) rather than to create a new adjustment method.

The literature on predictive methods for software effort estimation has shown that combining several predictive models into an ensemble can produce more accurate results than single models (Kocaguneli et al., 2012). Prior work on ensemble methods in the area of data mining also reports that ensembles can produce accurate results in comparison to single models, if not superior

* Corresponding author. Tel.: +962799930089.
E-mail addresses: m.y.azzeh@asu.edu.jo (M. Azzeh), abounas@uwo.ca (A.B. Nassif), L.L.Minku@cs.bham.ac.uk (L.L. Minku).

http://dx.doi.org/10.1016/j.jss.2015.01.028
0164-1212/© 2015 Elsevier Inc. All rights reserved.

(Seni and Elder, 2010; Hastie et al., 2008; Kohavi, 1995). The idea behind the success of ensembles is that the accurate predictions given by some of its models to a given example can patch the mistakes given by others to this example (Kocaguneli et al., 2012). In this way, the overall accuracy of the ensemble can be better than the individual accuracies of its base models. In order to achieve that, it is well accepted that the base models composing the ensemble should be diverse, i.e., they should make different mistakes on the same data points (Minku and Xin, 2013; Chandra and Yao, 2006). If they make the same mistakes, then the ensemble will also make the same mistakes as the individual models, and its performance will be no better than the individual performances. In other words, ensembles of non-diverse models are unsuccessful in improving the accuracy of these models.

Even though ensembles of software effort estimation models have been increasingly studied in software engineering, this is the first study that attempts to combine adjustment methods into ensembles. It is not known whether ensembles of adjustment methods would be successful in improving the accuracy of the calibration of EBA, and consequently the accuracy of EBA itself. In particular, it is not known whether different adjustment techniques behave diversely enough, i.e., if their amount of diversity is enough to lead to improvements in performance. If they do not, then combining these different techniques into an ensemble may not really improve performance. The main objective of this study is thus to investigate the potential of ensembles of adjustment methods for EBA.

With that in mind, this study aims at answering the following research questions:

RQ1. Is there evidence that ensembles improve the accuracy of adjusted EBA?
RQ2. Which approach is better for adjustment, linear or non-linear methods?
RQ3. Is there evidence that using different *k* analogies makes adjustment methods behave diversely?

The main contributions of this paper are the following:

(1) An evaluation of each adjusted EBA variant over all datasets to identify the ones that are actual prediction methods based on *standardized accuracy* (SA) measure and effect size.
(2) Ranking and clustering of actual prediction methods using Scott–Knott to identify the best methods with smallest *mean absolute error*.
(3) A new approach to build ensembles of adjustment methods based on Scott–Knott test method and Borda count procedure. This method can work well when all best methods identified by Scott–Knott are statistically similar. Existing methods such as win-tie-loss (Kocaguneli et al., 2012) cannot work well in this case because their ranking mechanism depends on the significance test between different methods.
(4) An evaluation of ensembles of adjustment methods against single adjustment methods using SA, effect size and other ranking methods, to determine whether ensembles are successful in improving performance of single adjustment methods.

In summary, this study is the first work to investigate ensembles of adjustment methods and the first work to create ensembles using Scott–Knott test and Borda count procedure.The remainder of the paper is structured as follows: Section 2 presents an overview of ensemble methods, as well as, the related work on adjustment methods and ensembles in software effort estimation. Section 3 describes the methodology conducted in this research. Section 4 shows the obtained results, which are discussed in Section 5. Section 6 presents threats to validity of our study. Finally, Section 7 presents our conclusions.

2. Background and related work

2.1. Ensembles in software effort estimation

Ensembles are learning methods that combine single (aka base) predictive models through a particular aggregation mechanism. The prediction given by the ensemble is a combination of the predictions given by each of its base models, e.g., weighted average (Seni and Elder, 2010). The principal idea of ensembles is that if their models are accurate and diverse, then their performance will be better than the one of its base models. Two models are said to be diverse if they make different errors on the same examples (Chandra and Yao, 2006). It is expected that diverse base models will give poor predictions to different examples. So, the poor predictions of a few models can be compensated by the good predictions of others, and the ensemble as a whole can achieve better performance than its base models (Song et al., 2013). On the other hand, if the ensemble is composed of non-diverse base models, its performance will not be better than its base models' individual performances (Kuncheva and Whitaker, 2003; Zhao and Ram, 2004; Brown et al., 2005).

The majority of studies in software effort estimation attempt to develop a new estimation method, and then compare the performance of that method against some well-known historical methods under certain conditions (Menzies et al., 2006). The area of software estimation appears now saturated with many predictive methods. Therefore, rather than developing new methods, there is a trend to replicate previous studies and investigate how we can benefit from their strengths. In practice, measuring accuracies of a particular method against some historical methods under certain settings cannot remain valid when changes on experimental conditions are made (Menzies et al., 2010). Thus, the method that is being considered superior over dataset *X* may not remain superior over other datasets or under different parameters (Mittas and Angelis, 2013). These facts are also true for EBA adjustment methods since most of them use learning methods that need parameters configuration for each training dataset. So, rather than proposing a new adjustment method we aim to benefit from the existing ones by using *ensembles*. Ensembles have been increasingly used in software engineering to solve regression and classification problems. In the software effort estimation area, Jorgensen recommends that when generating better estimates in expert judgment, it is necessary to use multiple decisions rather than a single one (Jorgensen, 2004).

Kocaguneli et al. (2012) distinguish between two main categories of prediction methods: *learner* method and *solo* method. Learner is a single method without supplement of pre or post-processing stages. The *solo* method is a method supplied with a pre-processing stage such as normalization and/or feature selection. Accordingly, the term *mutli-method* is used to indicate a collection of two or more solo methods (Kocaguneli et al., 2012). Different solo methods can be used to construct ensemble methods because they present different biases and assumptions. The importance of ranking stability and ensemble methods was studied over 90 solo methods and 20 datasets. The results obtained concluded that the ensemble methods were consistently superior, trustworthy and had a smaller error rate. However, their ensemble method is not guaranteed to work well in other contexts, because it concentrates on selecting the most accurate and stable solo methods, and there is no guarantee that these methods will behave diversely. Using different solo-methods does not guarantee that the corresponding base methods will behave diversely enough, i.e., it does not ensure that they are adequate for composing ensembles. In particular, it is known that there is a trade-off between diversity and accuracy of base models (Chandra and Xin, 2006). So, if solo-methods are chosen based on their accuracy only, the ensemble may lack diversity. Therefore, additional studies are necessary when combining other types of solo methods, in order to check whether they would lead to well performing ensembles.

Download English Version:

https://daneshyari.com/en/article/6885642

Download Persian Version:

https://daneshyari.com/article/6885642

Daneshyari.com