Information and Software Technology 55 (2013) 1512–1528

Contents lists available at SciVerse ScienceDirect

Information and Software Technology

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

Ensembles and locality: Insight on improving software effort estimation

Leandro L. Minku *, Xin Yao

CERCIA, School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

ARTICLE INFO

Article history:
Available online 13 October 2012

Keywords:
Software effort estimation
Ensembles of learning machines
Locality
Empirical validation

ABSTRACT

Context: Ensembles of learning machines and locality are considered two important topics for the next research frontier on Software Effort Estimation (SEE).
Objectives: We aim at (1) evaluating whether existing automated ensembles of learning machines generally improve SEEs given by single learning machines and which of them would be more useful; (2) analysing the adequacy of different locality approaches; and getting insight on (3) how to improve SEE and (4) how to evaluate/choose machine learning (ML) models for SEE.
Method: A principled experimental framework is used for the analysis and to provide insights that are not based simply on intuition or speculation. A comprehensive experimental study of several automated ensembles, single learning machines and locality approaches, which present features potentially beneficial for SEE, is performed. Additionally, an analysis of feature selection and regression trees (RTs), and an investigation of two tailored forms of combining ensembles and locality are performed to provide further insight on improving SEE.
Results: Bagging ensembles of RTs show to perform well, being highly ranked in terms of performance across different data sets, being frequently among the best approaches for each data set and rarely performing considerably worse than the best approach for any data set. They are recommended over other learning machines should an organisation have no resources to perform experiments to chose a model. Even though RTs have been shown to be more reliable locality approaches, other approaches such as *k*-Means and *k*-Nearest Neighbours can also perform well, in particular for more heterogeneous data sets.
Conclusion: Combining the power of automated ensembles and locality can lead to competitive results in SEE. By analysing such approaches, we provide several insights that can be used by future research in the area.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Estimating the cost of a software project is a task of strategic importance in project management. Both over and underestimations of cost can cause serious problems to a company. For instance, overestimations may result in a company loosing contracts or wasting resources, whereas underestimations may result in poor quality, delayed or unfinished software systems. The major contributing factor for software cost is effort [1]. So, models for estimating software cost/effort can be used as decision support tools, allowing investigation of the impact of certain requirements and development team features on the cost/effort of a project to be developed.

Several different software cost or software effort estimation (SEE) approaches have been proposed [2]. Among them, effort estimators based on machine learning (ML) approaches such as multi-layer perceptrons (MLPs), radial basis function (RBF) networks and regression trees (RTs) [3–9] have been receiving increased attention [2]. The motivation behind the use of such approaches is that they make no or minimal assumptions about the function being modelled and the data used for training. For instance, Tronto et al. [7] showed that MLPs improve SEE over conventional linear models because they are not restricted to linear functions, being able to model observations that lie far from the best straight line.

More recently, ensembles of learning machines have attracted attention of the SEE community for building SEE models [9,8,10,11]. However, existing work on automated[1] ensembles of learning machines for SEE presents contradictory conclusions regarding whether ensembles improve or not performance for SEE. Section 2.1 presents more details on these works. In the current work, we perform a principled and extensive analysis of existing automated ensembles of learning machines to determine whether they generally improve effort estimations given by single learning

* Corresponding author.
E-mail addresses: l.l.minku@cs.bham.ac.uk (L.L. Minku), x.yao@cs.bham.ac.uk (X. Yao).

[1] We refer to an approach as automated when, given the project data, it does not require human intervention and decision-making in order to be used. More details can be found in Section 2.1.

0950-5849/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.infsof.2012.09.012

machines. We build upon previous work and improve on their weaknesses by following a principled framework and doing an extensive analysis.

The methodology used in our work has the following advantages in comparison to previous work using existing automated ensembles:

- Use of principled experimentation, considering both parameter choice, statistical tests and magnitude of improvements.
- Use of a more reliable non-asymmetric performance measure (Mean Absolute Error – MAE) and a measure that facilitates investigation of the magnitude of the differences in performance (Standardised Accuracy – SA), rather than using only measures based on the Magnitude of the Relative Error (MRE) such as Mean MRE (MMRE) and the Percentage of Estimates within N% of the actual values PRED(N).
- Comparison using three different ensembles of learning machines (Bagging [12], Random [13] and Negative Correlation Learning [14]) which present features potentially beneficial to SEE.
- Use of a larger number of data sets (thirteen against five, the highest number of data sets previously used in studies involving automated ensembles), including both PROMISE [15] and ISBSG [16] data sets, rather than just PROMISE data sets.
- Experimental analysis of the behaviour of promising approaches, gaining insight on how to improve SEE.

Another area of research considered as promising in software project estimation is locality [17]. Approaches that perform estimations considering mainly training examples that are similar to the project being estimated can be referred to as based on locality. As SEE data sets tend to be relatively small and very heterogeneous, such approaches are likely to be more adequate. Examples of works considering locality are Cuadrado Gallego et al. [18] and Kocaguneli et al. [19]. Section 2.2 explains locality further. Even though locality is a promising area, it is not yet clear what locality approach would be more adequate for SEE. For instance, RTs are promising due to the hierarchy of features that they create, but it is not known whether this simply provides the same benefit as other locality approaches or a feature selector. Our work investigates different locality approaches for SEE, providing insight for future approaches on improving SEE.

As an additional contribution, our paper builds upon previous work and proposes an experimental framework for evaluation of SEE approaches. The framework joins (1) the power of statistical tests for comparison of multiple learning machines over multiple data sets as recommended in the general ML literature [20], to (2) an analysis of the approaches among the best, and to (3) the use of a standardised measure proposed by Shepperd and Mc Donell [21] for evaluating prediction systems in software project estimation.

In short, our paper addresses the following research questions:

- RQ1: Do existing automated ensembles of learning machines generally improve effort estimations given by single learning machines, including potentially adequate locality learning machines such as RTs? Which of them would be more useful?
- RQ2: What locality approach is more adequate for SEE tasks? In particular, how well does RT locality do in comparison to other locality approaches? On what type of data sets?
- RQ3: What insight on how to further improve SEE can we gain by analysing competitive ensemble and locality approaches?
- RQ4: How to evaluate/choose ML models for SEE?

Our key contribution is not in a new algorithm, but a better understanding/insight. Furthermore, such better understanding/insight is based on experimental studies, not just an intuition or speculation. We show that combining the power of automated ensembles and locality can lead to competitive results in SEE. For instance, when considering the symmetric performance measure MAE, bagging ensembles of RTs perform well. They are highly ranked in terms of performance across different data sets, are frequently among the best approaches for each data set and rarely perform considerably worse than the best approach for any data set. So, they are recommended over other learning machines should an organisation have no resources to perform experiments to chose a model. Moreover, tailored approaches using ensembles at a higher level and locality at a lower level may be particular useful for improving performance on smaller data sets, whereas approaches using locality at a higher level may be particularly useful for improving on larger data sets. Future work on SEE may benefit from exploiting that further. In terms of locality approach, RTs have been shown to be more reliable than other approaches due to their ability to create hierarchies of features. Nevertheless, k-means and k-nearest neighbours can also perform well, in particular for more heterogeneous data sets.

The rest of this paper is organised as follows: Section 2 presents related work on ensembles, locality and evaluation of models. Section 3 describes the data sets used in our study. Section 4 explains the experimental framework, which represents part of the answer to RQ4. Section 5 presents the evaluation of existing automated ensembles against single learning machines. It mainly aims at answering RQ1, but also partly addresses RQ2 by considering a promising locality approach namely RTs. This section also gives some insights on how to improve SEE (RQ3) by revealing the success of an approach joining the power of locality and ensembles and by showing that bagging ensembles still have room for improvement. As the analysis singles out a comparatively well performing approach, Section 5 also complements the answer to RQ4. Section 6 performs an analysis of locality approaches, answering RQ2 and part of RQ3. Section 7 presents an analysis of RTs and tailored approaches joining the power of ensembles and locality, mainly addressing RQ3. Section 8 explains threats to validity. Section 9 presents conclusions and future work.

2. Related work

2.1. Ensembles of learning machines for SEE

Ensembles of learning machines are sets of learning machines[2] trained to perform the same task and combined with the aim of improving predictive performance [22]. It is commonly agreed that the base learning machines should behave differently from each other. Otherwise, the overall prediction will not be better than the individual predictions [23–25]. So, different ensemble learning approaches can be seen as different ways to generate diversity among the base learning machines.

Ensembles of learning machines have recently attracted the attention of the SEE community, as they can frequently improve performance over single learning machines. For example, Bootstrap Aggregating (bagging) [12], a well known ensemble approach with solid theoretical background, is able to turn weak learning machines into stronger ones. This can be particularly useful for SEE, as the data sets are usually small, leading to typically less accurate learning machines than in other applications of ML.

In this section, we briefly describe previous work on ensembles for SEE. The works presented by Braga et al. [9], Kultur et al. [8] and Kocaguneli et al. [10] represent the starting points of the research

[2] The learning machines used to compose an ensemble are frequently called base learning machines.

Download English Version:

https://daneshyari.com/en/article/10367095

Download Persian Version:

https://daneshyari.com/article/10367095

Daneshyari.com