Applied Soft Computing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Evaluation of estimation models using the Minimum Interval of Equivalence☆

José Javier Dolado[a,*], Daniel Rodriguez[b], Mark Harman[c], William B. Langdon[c], Federica Sarro[c]

[a] Facultad de Informática, UPV/EHU, University of the Basque Country, Spain
[b] Dept. of Computer Science, University of Alcalá, 28871, Spain
[c] CREST, University College London, WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form 21 January 2016
Accepted 28 March 2016
Available online xxx

Keywords:
Software estimations
Soft computing
Equivalence Hypothesis Testing
Credible intervals
Bootstrap

ABSTRACT

This article proposes a new measure to compare soft computing methods for software estimation. This new measure is based on the concepts of Equivalence Hypothesis Testing (EHT). Using the ideas of EHT, a dimensionless measure is defined using the Minimum Interval of Equivalence and a random estimation. The dimensionless nature of the metric allows us to compare methods independently of the data samples used.

The motivation of the current proposal comes from the biases that other criteria show when applied to the comparison of software estimation methods. In this work, the level of error for comparing the equivalence of methods is set using EHT. Several soft computing methods are compared, including genetic programming, neural networks, regression and model trees, linear regression (ordinary and least mean squares) and instance-based methods. The experimental work has been performed on several publicly available datasets.

Given a dataset and an estimation method we compute the upper point of Minimum Interval of Equivalence, MIEu, on the confidence intervals of the errors. Afterwards, the new measure, MIEratio, is calculated as the relative distance of the MIEu to the random estimation.

Finally, the data distributions of the MIEratios are analysed by means of probability intervals, showing the viability of this approach. In this experimental work, it can be observed that there is an advantage for the genetic programming and linear regression methods by comparing the values of the intervals.

1. Introduction

The search for the best model to estimate software development effort or the code size is a recurring theme in software engineering research. The evaluation and comparison of various estimation models is usually performed using classical hypothesis tests [1,2] and other tools [3,4]. Although statistical testing methods have been considered as very powerful techniques in showing that two models are different, the estimates so obtained may not be within a range of any interest. There is a controversy related to the use of the *p*-values, which have been one of the most used criteria when assessing experimental results [5]. The ban on *p*-values established by a journal [6] implies that additional criteria must be used when comparing experimental data and methods. One of the most used criterion for comparing software estimation methods is the Mean Magnitude of the Relative Error (MMRE). Despite the fact that it has been proved as inadequate and inconsistent [7,8], it is still one of the most frequently reported evaluation criterion in the literature. The MMRE is a biased measure that should not be used for comparing models [9].

In this paper, a measure based on the approach of Equivalence Hypothesis Testing (EHT) is proposed. Using the upper point of the Minimum Interval of Equivalence (MIEu) for the absolute error and a random estimation as a reference point, we propose the MIEratio as the relative distance of the MIEu with respect to the random estimation. In this way, those measures will be computed on several publicly available datasets using a variety of estimation methods. At the end of the process, we construct

☆ Replication package available at https://github.com/danrodgar/mieratio.

* Corresponding author at: Facultad de Informática, UPV/EHU, University of the Basque Country, Spain. Tel.: +34 943018053.
E-mail addresses: javier.dolado@ehu.eus (J.J. Dolado), daniel.rodriguezg@uah.es (D. Rodriguez), m.harman@cs.ucl.ac.uk (M. Harman), W.Langdon@cs.ucl.ac.uk (W.B. Langdon), f.sarro@ucl.ac.uk (F. Sarro).

http://dx.doi.org/10.1016/j.asoc.2016.03.026
1568-4946/

several probability intervals that will allow us the comparison of the methods.

The following steps summarise the evaluation method:

1. Different estimations for each dataset are generated with different estimation methods, varying parameters. A bootstrapped confidence interval of the absolute error of the geometric mean is computed for each dataset, for each estimation method and for each set of parameters.
2. From the confidence intervals generated in the previous step, the one with the upper limit closest to 0 is selected and we take that upper limit point as the "Minimum Interval of Equivalence" (MIEu).
3. A random estimation is computed for each dataset. We assume this is the worst estimation an analyst can make.
4. For each dataset, the values obtained in steps 2 and 3 are used to compute the MIEratio as the measure for assessing the precision of the method.
5. Finally, the MIEratios are grouped by method. The distributions are analysed and plotted using credible intervals and highest posterior density intervals, taking a Bayesian point of view.

The rest of the article is organised as follows. Section 2 describes the approach followed in step 2, which takes its roots in the *bioequivalence analysis* method used in the medical and pharmacological fields. The elements described form the basis for the rest of the work. Section 3 describes the concepts used in steps 3 and 4 and defines a new measure for classifying methods, the MIEratio (see Section 3.2). Section 4.1 describes the estimation methods and Section 4.2 shows the datasets used. Section 4.3 describes in detail the data analysis procedures and Section 5 presents our results. Next, Section 6 analyses the data distributions of the MIEratios obtained. Threats to the validity are discussed in Section 7. Finally, Section 8 concludes the paper and highlights future research directions.

2. Equivalence Hypothesis Testing and confidence intervals

When making inferences about a population represented by a parameter w, the usual way to proceed is to state a null hypothesis H_0 about the population mean μ_w, $H_0 : \mu_w = \mu_0$, with μ_0 a specified value, and usually $\mu_0 = 0$ when analysing differences. Classical hypothesis testing proceeds by computing a statistic test and examining whether the null hypothesis $H_0 : \mu_w = 0$ can be rejected or not in favour of the alternative hypothesis $H_1 : \mu_w \neq 0$. The statistical tests try to disprove the null hypothesis.

Although classic "Null Hypothesis Significance Test" (NHST) is the standard approach in the software data analysis area, there is an equally valid alternative for the comparison of methods. Under the name of "Equivalence Hypothesis Testing" the null hypothesis is that of "inequality" between the things that we want to compare. This difference is assumed to be larger than a limit Δ. Therefore, the burden of the proof is on the alternative hypothesis of equivalence within the interval $(-\Delta, +\Delta)$. This interval has different names such as "equivalence margin", "irrelevant difference", "margin of interest", "equivalence range", "equivalence limit", "minimal meaningful distance", etc. [10].

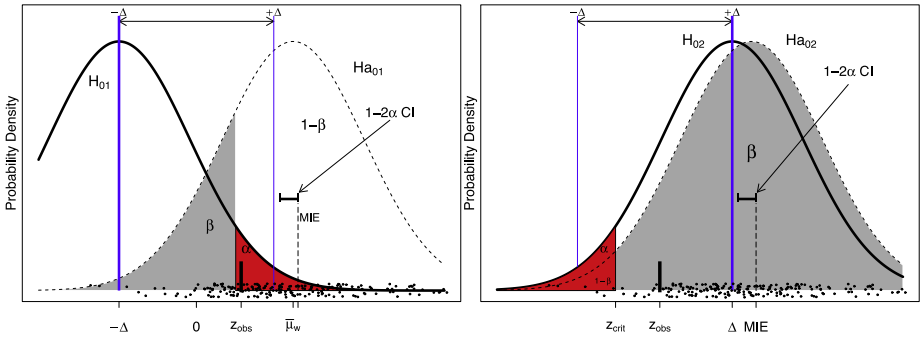
In EHT, the statistical tests and the confidence intervals are computed to check whether the null hypothesis of inequivalence can be rejected. The main benefit of this approach is that the statistical Type I Error when the null hypothesis is true, commonly named α, is controlled by the analyst, because it has to be predetermined in the null hypothesis. This α is the risk that the analyst is willing to take by wrongly accepting the equivalence of the things compared (i.e., rejecting the assumption of inequivalence). Note that in the NHST the error α has a different interpretation from EHT, i.e., it is the probability of wrongly accepting the difference of the things (rejecting the null difference). Here, the α, or Type I Error, is interpreted in the sense of EHT, i.e., the probability of concluding that the estimates and actual values differ (in absolute terms of the mean) by less than the MIEu when in fact they differ by a value of the MIEu or more. A review of the basic concepts used in EHT can be found in [10–13].

2.1. Confidence intervals and Two One-Sided Tests

There are two common approaches used to carry out the equivalence testing in frequentist statistics: Two One-Sided Tests and confidence interval methods (see for example, [11, Chapter 4; 14, Chapter 3]). In the following, both approaches are outlined.

2.1.1. Two One-Sided Tests

Let us assume that the parameter w has a normal distribution and $\bar{\mu}_w$ is its sample mean. The interval $(-\Delta, +\Delta)$ can be considered as *acceptable* for μ_w, which is also termed as the *irrelevant difference* for μ_w. The rationale for the Two One-Sided Tests (TOST) [15] is based on the fact that an irrelevant difference (or equivalence) within a range $(-\Delta, +\Delta)$ can be established on w by rejecting the two null hypotheses $H_{01} : \mu_w \leq -\Delta$ and $H_{02} : \mu_w \geq \Delta$. If both H_{01} and H_{02} are rejected then the conclusion is that $-\Delta < \mu_w < \Delta$. Fig. 1 shows a hypothetical distribution of values represented by the parameter w, $\bar{\mu}_w$ as the sample mean. For the sake of simplicity, let us assume normal distributions. In Fig. 1(a), we observe that H_{01} is rejected when the one-sided test is performed at $-\Delta$ (with the risk α, Type I Error, set at 0.05) because the observed value from the data, z_{obs}, is within the critical region. Therefore, it can be concluded that the value represented by μ_w is of no practical importance. However, in Fig. 1(b), when performing a t-test at $+\Delta$, it can be observed that H_{02} is not rejected, therefore



(a) H_{01} at $-\Delta$. (b) H_{02} at Δ.

Fig. 1. Visualisation of the TOST approach. The figure also shows the confidence interval on the mean μ_w outside the interval $(-\Delta, \Delta)$.

Please cite this article in press as: J.J. Dolado, et al., Evaluation of estimation models using the Minimum Interval of Equivalence, Appl. Soft Comput. J. (2016), http://dx.doi.org/10.1016/j.asoc.2016.03.026