Jestr

Journal of Engineering Science and Technology Review 4 (3) (2011) 266 - 270
Special Issue on Econophysics

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Research Article

Evaluating the Performances of Software Cost Estimation Models through Prediction Intervals

N. Mittas*

Dep. of Electrical Engineering, TEI Kavalas, 65 404 St.Lucas, Kavala, Greece.

Abstract

The task of predicting accurately the cost required for the completion of a new software project is a challenging issue in the Software Cost Estimation area, since it is closely related with the activities of project management and the wise decision-making of organizations in order to bid, plan and budget a forthcoming system. However, the accurate prediction of the cost is often obtained with great uncertainty and for this reason there has been noted a lack of convergence in experimental studies. The main reason for the discrepancy can be derived from the inherent characteristic of prediction methodologies, since they produce point estimates without taking into account the risk covering the whole process. In this study, we propose a statistical framework, so as to focus on the construction of Prediction Intervals which provide an "optimistic" and a "pessimistic" guess for the true magnitude of the cost. The proposed framework that incorporates different accuracy indicators, formal hypothesis testing and graphical inspection of the predictive performance is applied on a dataset with real software projects.

Keywords: Software Cost Estimation, Prediction Interval, ARPI, Estimation by Analogy, Ordinary Least Squares Regression.

1. Introduction

Judging by reports from everyday practice and findings in the literature, software has become one of the most important parts of our lives, whereas on the same time it has also turned into the most expensive component of computer systems. Due to the abovementioned facts, *Software Cost Estimation* (SCE) is considered as one of the most critical phases in planning, scheduling and risk management of software projects and has attracted the interest of many researchers during the last decades [1]. Generally, SCE is the process of assessing the overall expense of software in terms of money, effort and time.

The economies of software projects play a significant role to both developers and customers, since they are the basis of generating request for proposals, contracts negotiations, scheduling, monitoring and controlling the whole process of development [2]. The underestimation of a project may affect the earnings of the development organization leading to wrong managerial decision-making with systems that exceed their budgets and delays of the deliverables. On the other hand, overestimating the costs can lead to the cancelling and loss of a contract, since too many resources result in not winning the contract.

Despite the evolutionary introduction of many prediction methodologies ranging from expert judgement techniques to algorithmic and machine learning models, the findings are associated with inconsistencies regarding the superiority of a technique over another (see for example [3]). Although the researchers strive to indentify the factors for the lack of convergence in experimental studies, it seems that they do not take into account an inherent limitation of prediction systems that produce estimates which are expressed as single numbers (*point estimates*) without considering the uncertainty and risk when estimating a single value of cost [4].

Hence, the estimation of a *Prediction Interval* (PI) consisting of a lower and upper limit between which the future value expected to lie with a predefined probability, is a more realistic approach, especially from a project manager's point of view. Forecast practitioners in other applied areas often face a similar quandary and so most managers do realize the importance of providing PIs instead of a single value of estimate. Although there is an imperative need for the construction of reliable and accurate PIs in SCE ([4], [5] and [6]), the topic of the comparison of PIs has not attracted much of the interest of the research community, yet.

Our aim is therefore to deal with a critical research issue in SCE concerning the simultaneous comparison of PIs derived from alternative prediction models. More precisely, we examined the predictive power of four models over a public domain dataset. Usually interval estimates are created during the point estimation process by computing a PI for the prediction. Another alternative is to predefine the intervals and then to use a model that predicts in which of the intervals the cost will fall. The comparison of PIs for the alternative models is based on the well-known *hit-rate*, the measure of *Actual effort Relative to PI* (ARPI) that inspects how the actual cost values are distributed relative to the cost

* E-mail address: nmittas@csd.auth.gr

PIs and analyzes the bias in the estimated uncertainty distribution [5] and a width analysis of PIs through formal statistical hypothesis testing. Finally, we recently proposed a measure that generalizes the hit-rate taking into account the similarity and width of PIs [6].

The rest of the paper is organized as follows: In Section 2, we present the experimental setup of the study concerning the alternative prediction models, the evaluation of PIs and the indicators of the predictive power. In Section 3, we present the experimental results and finally, in Section 4 we conclude by discussing the results.

2. Experimental Setup

This section provides information concerning the comparison framework and the experimental setup of the study. More precisely, we give certain descriptions about the alternative prediction models, the construction of PIs and the measures evaluated for the comparison purposes.

2.1 Comparative Prediction Methods

In this study, we decide to select different methods covering a part of the distribution of the proposed methodologies appeared so far in the literature of SCE [1]. The selected methods can be grouped into two main categories that are *i*) methods that produce point estimates accompanied by prediction intervals and *ii*) methods that estimate the cost within predefined intervals. Due to the space limitation of the study, we do not give a detailed description of the prediction models, since they are well-established methods and have been already applied in SCE. On the other hand, in Table 1, we present a brief and a more abstract portrayal of the general idea for each method.

2.2 Evaluation of Prediction Intervals

From what we have already mentioned, it is clear that two of the comparative prediction models (CART and NB) classify a project in a predefined interval. In order to apply these techniques, we have to compute a new variable that categorizes the dependent variable's effort, measured on an ordinal scale. For this reason, we compute the quartiles of the empirical distribution of effort. Finally, four interval categories are generated implying that all categories have almost the same probability to contain the actual effort of a new project. On the other hand, the rest methods result in a point estimate, so there is a need to briefly describe how we can derive a PI from point estimates in regression-based techniques.

The technique used is known as the *leave-one-out cross-validation* (LOOCV) [6], i.e. after removing a project from the dataset, a model for each prediction technique is generated using all the remaining projects and this in turn is utilized to provide estimation, along with a PI for the cost value of the removed project. As far as the parametric OLS concerns, it is known from the theory that a PI can easily be evaluated by explicit formulae [12]. In contrast, there is no such way to evaluate PIs for the case of the non-parametric EbA model, and therefore a simulation technique, namely non-parametric bootstrap, is adopted [5]. The method is based entirely on the empirical distribution of the dataset without any assumption on the population distribution, whereas the rationale behind the procedure is the generation of a large number of independent samples drawn with replacement from the original sample.

Table 1. Description of prediction methodologies

Methodology	**General Idea**
Methods that produce point estimates accompanied by prediction intervals	
Ordinary Least Squares Regression (OLS)	Explains the relation between several independent variables (cost factors) and a dependent variable (effort) in the form of a parametric linear relationship (see indicatively, [7]). Since the variables are usually non-normally distributed, they need some transformation (logarithmic) in order to obtain a valid linear model. In addition, in order to handle mixed data with categorical and continuous variables, we replace the categorical variables by binary (or dummy) variables.
Estimation by Analogy (EbA)	Mimics the human instinctive decision-making by comparing with similar cases. A type of non-parametric regression procedure, where the unknown values of the dependent variable are estimated by the known values, of the same variable, corresponding to neighbours (analogies) of the estimated case [10]. Analogies are found through the evaluation of a prefixed similarity (or dissimilarity) criterion of cases, based on the independent variables.
Methods that estimate the cost within predefined intervals	
Classification and Regression Tree (CART)	A statistical and machine-learning procedure widely used in predictive modelling for building classification models with a tree-based structure [9]. The CART model consists of a hierarchy of decisions, whereas the algorithm used operates by choosing the best variable for splitting data into two groups at the root node. It can use any one of several different splitting criteria, all producing the effect of partitioning the data at an internal node into disjoint subsets in such way that the class labels are as homogeneous as possible. This splitting procedure is then applied recursively to the data in each of the child nodes. A greedy local search method to identify good candidate tree structures is used.
Naïve Bayes Classifier (NB)	A simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions and despite this fact, in practice NB often competes well with more sophisticated classifiers [10].

2.3 Comparison of Prediction Intervals

As the main scope of the study is the comparison of PIs, in this section, we present the way that the predictive performance is evaluated in the experiment. Firstly, the performance is computed via the *hit-rate* that is the most known accuracy indicator defined as the percentage of the

total number of correct predictions to the total number of projects in the dataset. Despite the wide applicability of this measure, we have to point out that hit-rate does not take into account the width of the PIs and the overlapping part of PIs obtained by two comparative techniques.

Jorgensen et al. [5] proposed the utilization of the *Actual effort Relative to PI* (ARPI), a measure that is able to reveal the potential bias in the estimated uncertainty distribution (Eq. 1). The ARPI provides values close to -0.5 and 0.5 when the actual effort is close to the estimated lower and upper bounds, respectively, and equals 0 when the actual effort corresponds to the midpoint of a PI. An ARPI value outside the interval [-0.5, 0.5] means that the actual effort is outside the effort PI.

$$ARPI_i = \frac{Act_i - PI\_midpo\text{int}_i}{upper_i - lower_i} \quad (1)$$

where

$$PI\_midpo\text{int} = \frac{upper_i + lower_i}{2}$$

Recently, we proposed a new measure, namely *adjusted hit-rate* that can be considered as the generalization of hit-rate taking into account not only the inclusion of the actual value in an interval, but also the similarity and width of the comparative intervals [6]. Supposing that we have to compare two intervals $A=[a_1,b_1]$ and $B=[a_2,b_2]$, there are four possible cases (Table 2) concerning their overlapping part: *i*) the intervals do not have any overlapping parts, i.e. their intersection is the empty set ($A\cap B=\varnothing$) (Table 2a), *ii*) the intervals have an overlapping part, i.e. their intersection is not an empty set ($A\cap B\neq\varnothing=[a_2,b_1]$) (Table 2b), *iii*) one of the intervals is contained within the other interval, i.e. their intersection is the entire "smaller" interval ($A\cap B=B=[a_2,b_2]$) (Table 2c), and *iv*) the two intervals have the same lower and upper bounds ($A=B$) (Table 2d).

Based on the relative positions of PIs, we can give a score to each PI using one of the equations described in Table 2. The general idea of the proposed measure is to give an advance to the correct predictions through weights but also to take into consideration the overlapping and the width of PIs. The final score can be summarized through the mean value of these scores.

Table 2. Adjusted hit-rate scores for the comparison of PIs

	(a)		(b)		(c)		(d)	
	A	B	A	B	A	B	A	B
$x\in A\wedge x\notin B$	1	0	$1\frac{a_2-a_1}{b_1-a_1}$	0	$1\frac{a_2-a_1}{b_1-a_1}$ or $1\frac{b_1-b_2}{b_1-a_1}$	0	impossible	
$x\notin A\wedge x\in B$	0	1	0	$1\frac{b_2-b_1}{b_2-a_2}$	impossible		impossible	
$x\in A\wedge x\in B$	impossible		$1\frac{b_1-a_2}{b_1-a_1}$	$1\frac{b_1-a_2}{b_2-a_2}$	$1\frac{b_2-a_2}{b_1-a_1}$	1	1	1
$x\notin A\wedge x\notin B$	0	0	0	0	0	0	0	0

3. Experimentation

The publicly available dataset used in our experimentation comprises 77 completed software project data from a Canadian Software house [12] with both continuous and categorical variables, whereas the dependent variable is the actual effort.

From the measures presented in Table 3, we can observe that the OLS gives the highest value (98.70%) and CART the smallest value (50.65%) of hit-rate indicators, respectively. Thus, based solely on hit-rate, a practitioner would conclude that OLS is the best prediction technique for PI estimates.

Table 3. Comparison of PIs

	Hitrate	MedianARPI
OLS	**98.70%**	-0.18
EbA	63.64%	**-0.08**
NB	51.95%	0.10
CART	50.65%	-0.16

Table 4. Comparison via adjusted hit-rate measures

	OLS	EbA	NB	CART
OLS	-	0.45	0.53	0.54
EbA	**0.62**	-	**0.44**	**0.46**
NB	0.44	0.34	-	0.49
CART	0.41	0.34	**0.51**	-

In order to better illustrate the predictive power of the competitive models, in Figure 1 we present the actual lnefforts of the dataset along with the 95% PI by (a) OLS and (b) EbA and the predicted class of (c) NB and (d) CART models, respectively. For better interpretation, we present all the lower and all upper bounds connected with a line forming in every graph a prediction zone, whereas due to the high variability of the actual cost values, we decide to transform the *y*-axis to the natural logarithmic scale, so as to better illustrate the differences between the comparative models. Finally, the projects are sorted in ascending order according to the actual dependent variable, so the *x*-axis represents the ranks of each project in this order. As we can see from figures, the parametric OLS model provides in general extremely wide PIs. This is not true for the case of non-parametric EbA model, since it presents the narrowest PI-zone, which is significantly better than the corresponding PIs of the other methods, with a quite reasonable and high hit-rate. On the contrary, NB and CART, which estimate the cost within predefined intervals, present generally quite rough zones, whereas OLS and EbA evaluate more smooth PIs. Therefore, we should take a more careful examination of the predictive performances of the competitive techniques.

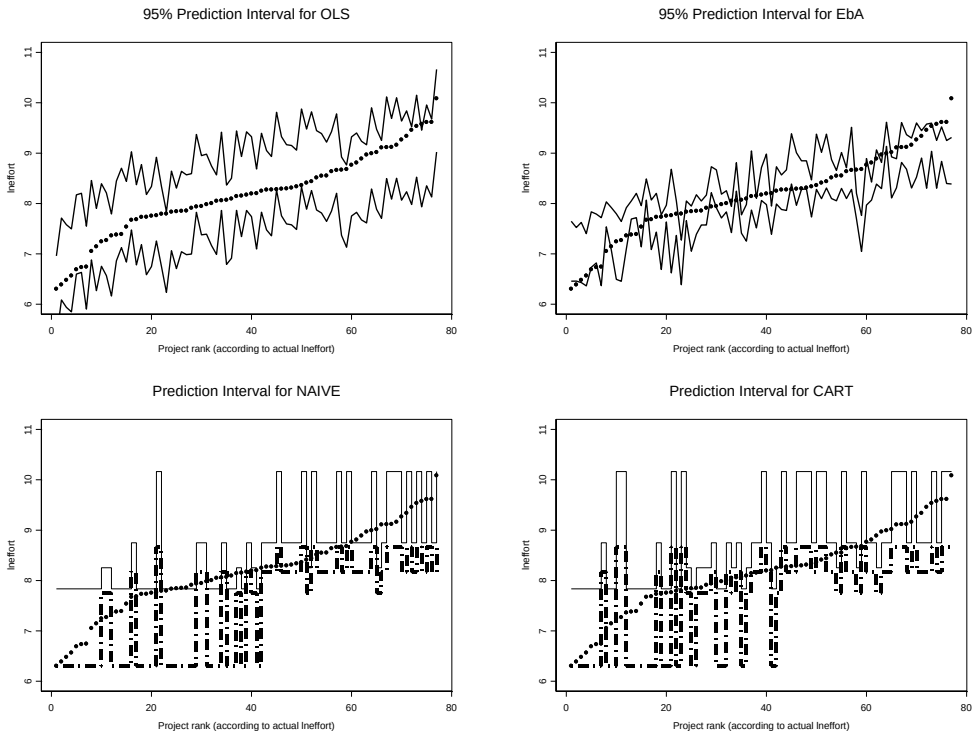

Fig. 1. (a) 95% prediction interval for OLS, (b) 95% prediction interval for EbA, (c) predicted class for NB and (d) predicted class for CART

Concerning the overall performance of the models via ARPI measure, the researchers [5] propose the utilization of the MedianARPI (Table 3). In our dataset, three of the models present negative values, so the actual effort is generally, closer to the estimated lower bounds for these methods. Additionally, EbA seems to produce the least biased PIs, since the MedianARPI of the abovementioned method is the closest to zero value. The second best method is NB, whereas OLS presents the most biased PIs. However, a single measure is just a statistic and as such contains significant variability. Thus, when we compare models based solely on a single value we take the risk to consider as a significant difference which in fact may be not so significant. In order to remedy this issue, we can test the significance of the differences using the non-parametric *Wilcoxon signed rank* procedure. The Wilcoxon test indicates a statistically significant difference for all pair-wise comparisons except two cases (EbA-CART and NB-CART).

On the other hand, having in mind that there are two crucial issues regarding the comparisons of PIs, which are whether the actual values are contained within the intervals and the similarity of PIs with respect to their overlapping part and their widths, we should take a more careful examination of the PIs through the proposed adjusted hit-rate indicator (Table 4). Indeed, the adjusted hit-rates of EbA show that the methodology dominates in terms of efficiency meaning that the PI can be narrower without losing hit rate. For example, the comparison of EbA-OLS models demonstrates that the score of EbA is 0.62 compared with the value of 0.45 for OLS. This means that although OLS presents higher value of the simple hit-rate indicator, the adjustment through the new measure incorporates further information concerning the power of EbA model. Taking into account the Wilcoxon tests, the results reveal that there are three methods with similar performances {EbA, CART, NB}, which in turn outperform OLS.

4. Conclusions

The paper suggested a comprehensive framework for comparisons between models built from different prediction techniques presenting ways of measuring and comparing interval prediction accuracy. Although a plethora of studies indicates the requirement that each estimation technique should be accompanied with a prediction interval, since they can play a critical role in the well-balanced management and planning of a software project, there has been noticed a little concern regarding the formal comparison of PIs. In this study, the comparison was conducted through the usage of the well-known hit-rate indicator and the Actual effort Relative to PI measuring the bias of the interval. Finally, as the crucial issue of a comparison of two alternative prediction interval techniques is the similarity of their intervals, we highlight how a recently new measure taking into account the width and overlapping points of the intervals can constitute a well-defined methodology in order to assess the superiority of one model against the other.

Summarizing the findings of the experimentation, the constructed PIs of four comparative methods revealed that although OLS presents an especially high value of hit-rate measure, it seems not to be the best prediction interval technique regarding the width and the overlapping points compared with the PIs of the other models. On the contrary, the EbA model achieves to produce the narrowest PIs and presents statistically significant differences when the adjusted hit-rate measure is evaluated for the comparison purposes. Finally, regarding the ARPI measure, EbA attained the least biased PIs, whereas in the case of the OLS, the measure indicated biased predictions. Concerning the comparison between the techniques that produce point estimates accompanied by prediction intervals and the techniques that estimate the cost within predefined intervals, although the latter form does not result in smooth PIs, it produces generally accurate intervals.

References

1. M. Jorgensen and M. Shepperd, "A Systematic Review of Software Development Cost Estimation Studies," IEEE Transactions on Software Engineering, vol. 33, no. 1, 33, Jan. 2007.
2. M. Shepperd, "Software Project Economics: A Roadmap," Proc. of the IEEE International Conference on Future of Software Engineering (FOSE '07), pp. 304-315, 2007.
3. C. Mair, and M. Shepperd, "The Consistency of Empirical Comparisons of Regression and Analogy-based Software Project Cost Prediction," Proc. of the International Symposium on Empirical Software Engineering (ISESE'05), pp. 509–518, Nov. 2005.
4. B. Kitchenham, and S.Linkman, "Estimates, Uncertainty and Risk," IEEE Software, vol. 14, no. 3, pp 69-74, 1997.
5. M. Jørgensen, K. Teigen, and K. Moløkken, "Better Sure than Safe? Overconfidence in Judgment Based Software Development Effort Prediction Intervals," Journal of Systems and Software, vol. 70, no. 1-2, pp. 79-93, 2004.
6. N. Mittas, and L. Angelis, "Bootstrap Prediction Intervals for a Semi-parametric Software Cost Estimation Model," Proc. of the IEEE 35th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 293-299, 2009.
7. I. Myrtveit, and E. Stensrud, "A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models," IEEE Transactions on Software Engineering, vol. 25, no. 4, pp. 510–525, 1999.
8. M. Shepperd, and C. Schofield, "Estimating Software Project Effort Using Analogies. IEEE Transactions on Software Engineering, vol. 23, no. 12, pp. 736–743, 1997.
9. D. Hand, H. Mannila, and P. Smyth. Principles of Data Mining. MIT Press, US, 2001.
10. I. Rish, "An empirical study of the naive Bayes classifier," Workshop on Empirical Methods in Artificial Intelligence (IJCAI), 2001.
11. D. Montgomery, E. Peck, and G. Vining. Introduction to Linear Regression Analysis. Wiley Series in Probability and Statistics, 4th edition, 2001.
12. J. Desharnais. Analyse statistique de la productivitie des projets informatique a partie de la technique des point des fonction. Master's thesis, University of Montreal, 1989.