A Survey: A Multiple Comparisons Algorithm based Ranking and Clustering of COCOMO and Putnam's Software Cost Estimation Models

Chetana D. Patil
PG student
Department of Computer Engineering, North Maharashtra University
SES's R. C. Patel Institute of Technology, Shirpur, MS, India

Tareek M. Pattewar
Assist. Proffesor
Department of Computer Engineering, North Maharashtra University
SES's R. C. Patel Institute of Technology, Shirpur, MS, India

ABSTRACT
Software project can be completely predicting the most realistic effort using Software Cost Estimation. There are variety of methods and models trying to improve the estimation procedure of Software project development and application. From the variety of methods emerged the need for comparisons to determine the best model. Here, we propose a statistical framework based on a multiple comparisons algorithm in order to rank several cost estimation models, identifying those which have significant differences in accuracy, and clustering them in non overlapping groups. The proposed framework is applied in a large scale setup of comparing prediction models over datasets.

General Terms
Software cost estimation, management, metrics/measurement, statistical method.

Keywords
Project estimation, Effort estimation, Cost models.

1. INTRODUCTION
Software cost estimation is the process of predicting the effort required to develop a software system. A software project can be completed predicting the most realistic effort using software cost estimation. There are variety of methods and models trying to improve the estimation procedure of Software project development and application [1]. estimation as one of the most important activities that is closely related with the success or failure of the whole The quickly increased need of large-scaled and complex software systems leads managers to settle software cost estimation as one of the most important activities that is closely related with the success or failure of the whole development process. Not exact estimates can be harmful for both developers and Customers since they can cause the delay of the product deliverables or the cancelation of a project.

The most common research topic of software cost estimation is the introduction and evaluation of estimation methods. On the other side, the variety of prediction methods is also related with inconsistent findings concerning the superiority of one technique.

Likewise, the correctness of the error measures used for the comparison of alternative models [2].While Mean Magnitude of Relative Error (MMRE) has been analyzed as a problematic accuracy measure to select the "best" model [3],it continues to be considered as the main indicator for the performance of software cost estimation methods.

The arithmetical process that is used when comparing multiple prediction techniques. The simple comparison between two competitive models, the null hypothesis is tested through a classical statistical test. When more than two models are used, the comparison of two models are come to be more complicated, and the problems related with it are identified in statistics as the "multiple comparisons problem".

Because of the large number of proposed cost estimation methods, it is essential for project managers to methodically base their choice of the most accurate model on fixed statistical procedures [4]. The problem of concurrent comparisons between multiple prediction models which can detect the major differences between a number of cost estimation methods and at the same time be able to rank and cluster that methods, defining the best ones.

The proposed methodology is based on the analysis of a design of experiment (DOE) or Experimental Design, a basic statistical tool in many applied research areas such as such as engineering, financial, and medical sciences. In the field of SCE it has not yet been used in a systematic manner.

Generally, DOE refers to the process of planning, designing, and analyzing an experiment in order to derive valid and objective conclusions effectively and efficiently by taking into account, in a balanced and systematic manner, the sources of variation. DOE analysis used to compare different cost prediction models by taking into account the blocking effect, i. e., the fact that they are applied repeatedly on the same training-test datasets.

Specifically, the algorithm we propose ranks and clusters the cost prediction models based on the errors measured for a particular dataset. Therefore, each dataset has its own set of "best" models. This is more realistic in SCE practice since each software development organization has its own dataset and wants to find the models that best fit its data rather than trying to find a globally best model which is unfeasible.

Furthermore, the clustering as an output is different from the output of pair wise comparisons tests. A pair wise test, for example, can possibly indicate that models A and B are equivalent, models B and C are also equivalent, but models A and C are different. The grouping of model B is therefore questionable. For larger numbers of models the overlapping homogeneous groups resulting from pair wise tests are ambiguous and problematic in interpretation. On the other hand, a ranking and clustering algorithm provides clear groupings of models, designating the group of best models for a particular dataset.

The recommended statistical approach is also constructed on an algorithmic procedure which is capable to produce non overlapping clusters of prediction models, consistent with respect to their predictive performance. Aimed at, apply a specific test from the generic class of multiple comparisons procedures, namely, the Scott-Knott test[5] which ranks the models and partitions them into clusters.

1.1 Multiple Comparisons Algorithm

The purpose of most multiple-comparisons procedures is to control the "overall significance level" for some set of inferences performed as a follow-up to ANOVA. This "overall significance level" or error rate is the probability, conditional on all the null hypotheses being tested being true, of rejecting at least one of them, or equivalently, of having at least one confidence interval not include the true value.

1.2 COCOMO Model

The Constructive Cost Model (COCOMO) is an algorithmic software cost estimation model developed by Barry W. Boehm. The model uses a basic regression formula with parameters that are derived from historical project data and current as well as future project characteristics. In the COCOMO model, the code size is given in thousand LOC and effort is in person-month.

1.3 Putnam's Model

Putnam derives his model based on manpower distribution and his finding in analyzing many completed projects. Putnam's model is also widely used in practice and SLIM is a software tool based on this model for cost estimation and manpower scheduling.

2. RELATED WORK

Magne Jorgensen et al. provide a basis for the development of software estimation research through orderly review of software development cost estimation studies [1].

Sweta Kumari et al. proposed comparison and analysis of different software cost estimation methods. The basic input for the software cost estimation is coding size and set of cost drivers, the output is Effort in terms of Person-Months (PM's) [6].

Martin Shepperd et al. offered the comparing software prediction techniques using simulation for exact software prediction systems rises as software come to be much larger and more complex. The fundamental characteristics: size, number of features, type of distribution, etc., of the data set impacts the choice of the prediction system to be used. The results advise that there are major differences depending upon the characteristics of the data set [7].

N. Mittas et al. comparing cost prediction models by resampling techniques. A certain limitation of several past studies is comparison without using proper statistical hypothesis testing. This can lead to incorrect results and unjustified simplifications about the predictive accuracy of estimation techniques [8].

Iman Attarzadeh et al. proposed a novel algorithmic cost estimation model based on soft computing technique. Newer soft computing methods to effort estimation based on non-algorithmic techniques like fuzzy logic (FL) may offer an alternative for solving the problem of long term procedure. The goal of this work suggests new fuzzy logic realistic model to achieve more accuracy in software effort estimation. The main objective of this work was to look into the role of fuzzy logic technique in improving the effort estimation accuracy by characterizing inputs parameters using two side Gaussian function which gave superior transition from one interval to another [9].

Vahid Khatibi et al. proposed software cost estimation methods. Project planning is the most important process in software projects development. Because of poor planning project faults and inaccurate outcomes are occurred for the project team. If cost and effort are determined negative in software projects, suitable occasions can be missed; whereas positive predictions can be caused to some resource losing. Nowadays software project managers should be aware of the increasing of project failures. So introducing and focusing on the estimation methods essential for achieving to the accurate and reliable estimations [10].

Salvatore Alessandro Sarcia et al. proposed a technique of detecting and handling scope error consists of analyzing the performance of estimation models from an uncertainty point of view [11].

Efi Papatheocharous et al. provide a Feature Subset Selection for Software Cost Modeling and Estimation that investigates the appropriateness of attributes, obtained from empirical project databases and aims to reduce the cost drivers used while preserving performance [12].

T. Foss et al. proposed the Mean Magnitude of Relative Error (MMRE), is may be the most commonly used estimation condition for evaluating the performance of conflicting software prediction models. Some determination of MMRE is to support us to select the best model from more than two models. Here, performed a simulation study representing that MMRE does not every time select the best model. Their conclusion cast some uncertainty on the conclusions of any study of competing software prediction models that used MMRE as a basis of model comparison [3].

A. Scott and M. Knott proposed a cluster analysis method for grouping means in the analysis of inconsistency.

3. METHODOLOGY

Now, define the processes which are used to compare a set of candidate prediction methods and then define the algorithm based on the Scott-Knott test. The Scott-Knott test discourses the limitations of the well-known techniques.

3.1 Problems related with Comparison of Multiple Prediction Techniques

The problem belongs to a general class in statistics known as "multiple hypothesis testing". The multiple hypothesis testing can be determined as the technique of testing more than one hypothesis at the same time.

Concisely labeling the problem, the conclusions obtained from a statistical hypothesis test are every time focus to improbability. For this purpose, identify an satisfactory maximum probability of rejecting the null hypothesis.

The problem of the increasing error can be covered by adjusting the total error, but still the execution of a large number of pair wise comparisons is not so straightforwardly interpretable from that time the resulting groupings are overlapping. For this purpose, several methods have been offered to perform targeted multiple comparisons [13]. The problem of ranking and clustering cost estimation model according to their accuracy can be controlled by the algorithm, which is based on the Scott-Knott test.

3.2 Comparative Prediction Models

The comparative prediction methods can be grouped into three main categories that are regression-based models, analogy-based techniques, and machine learning methods [8, 13].

All these methods are well-established methods, there is a huge literature on them, and they have been applied in software cost estimation.

3.2.1Regression-based Models

Use data to identify relationships among variables and use these relationships to make predictions. For example, to predict the selling price of a house without any particular knowledge of the house, we use the average selling price of all of the houses in the data set.

The relationship between dependent and independent variables is expressed as a known explicit function with unknown parameter. Then the parameters are estimated in a such way so as to minimize a predefined criteria.

3.2.2 Analogy-based Techniques

Estimates costs by comparing proposed programs with similar, previously completed programs for which historical data is available.

A type of nonparametric regression procedure, where the unknown value of dependent variables are estimated by the known values, of the same variable, corresponding to neighbors of the estimated case.

3.2.3 Machine Learning Methods

Construction of a relationship between dependent and independent variables based on empirical data and past experience i.e. relationships are constructed on the basis of data observed and historical data in similar types of projects or product.

A branch of artificial intelligence techniques that are concern with construction of relationship between dependent and independent variables based on empirical data and past experience.

3.3 Accuracy Measures

The accuracy measures [13] that are most frequently used for validating cost models are based on theyA(actual) and the*yE*(estimated from a model) cost values. There is a need for utilization of three different error functions measuring three aspects of the prediction performance of comparative models.

More precisely, Absolute Error (AE) is used in order to evaluate the accuracy of models, whereas error ratio *z* has been adopted as a measure of bias accounting for underestimations ($z < 1$) or overestimation ($z > 1$) with an optimum value of 1.The most widely known MRE indicator was also used, it provides a measure of the spread of the error ratio *z*.

The local measures of error that are computed through the actual (yA) and the estimated (yE) values of each single project *i* constitute the basis for the evaluation of the overall prediction performance of the comparative models by computing a statistic (i.e., mean) for a set of *n* test cases.

3.4 Scott-Knott Test

The Scott-Knott test is a multiple comparison procedure based on principles of cluster analysis. The clustering states to the models compared and multiple cases, though the standard for clustering together treatments is the statistical implication of differences between their mean values. The Scott-Knott test depends on a specific suitable feature of the method, which is capable to isolate the models into non overlapping clusters. In this case, the values of the response variable that is impressed by the models are converted to expressions of the prediction errors gained from the models being compared. The algorithm able to rank and cluster prediction models according to their accuracy.

The Scott-Knott technique follows and uses the one-way analysis of variance (one-way ANOVA). The one-way ANOVA checks the null hypothesis that the treatment means are all the same or, evenly, that there is no arithmetical difference between the accuracy measures achieved by the compared models.

However, the another possibility is that the models can be partitioned into two mutually exclusive nonempty subsets [5,13].

4. CONCLUSION

In this paper, finding the most important reason for the software project failures has been the subject of many researches in last decade. The reason for software project failures is incorrect estimation in early stages of the project. So introducing and focusing on the estimation methods seems necessary for achieving to the accurate and reliable estimations.

Performance of each estimation method depends on several parameters such as complexity of the project, duration of the project, expertise of the staff, development method and so on.

Software cost estimation about the concurrent comparison of

alternative prediction models, ranking of that prediction models and clustering on the basis of similar performance.

The entire process is decided on well-established statistical methodologies, which is taking into consideration the multiple comparison problems. Keeping in mind the critical role of the adoption of reliable practices in the development process for both project managers and customers, here proposed a formal framework and structured guidelines in order to reinforce the knowledge acquisition and diminish the inherent uncertainty in SCE.

The proposed statistical hypothesis testing through the Scott-Knott test verifies that the predictive accuracy of a set of methods does not confirm a statistically significant difference among them.

For future work, we would like to compare the more than one software cost estimation models. Ranking of the prediction models and clustering is provided on the basis of similar performance to the compared prediction models. Increase accuracy of output using more models such as Putnam's model.

5. ACKOWNLEGDEMENT

We are thankful to all the personalities who helped us throughout this survey.

6. REFERENCES

[1] M. Jorgensen and M. Shepperd, A Systematic Review of Software Development Cost Estimation Studies, IEEE Trans. Software Eng.,vol. 33, no. 1, pp. 33-53, Jan. 2007.

[2] B. Kitchenham, S. MacDonell, L. Pickard, and M. Shepperd.What accuracy statistics really measure. IEEE Proc. SoftwareEng, vol. 148, pp. 81–85, June 2001.

[3] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, A Simulation Study of the Model Evaluation Criterion MMRE,IEEE Trans. Software Eng., vol. 29, no. 11, pp. 985-995, Nov. 2003.

[4] I. Myrtveit, E. Stensrud, and M. Shepperd, Reliability and Validity in Comparative Studies of Software Prediction Models, IEEE Trans. Software Eng., vol. 31, no. 5, pp. 380-391, May 2005.

[5] A. Scott and M. Knott, A Cluster Analysis Method for Grouping Means in the Analysis of Variance, Biometrics, vol. 30, no. 3, pp. 507-512, Sept. 1974.

[6] SwetaKumari, ShashankPushkar, Comparison ans Analysis of Different Cost Estimatoion Methods, IJACSA, vol. 4,no.1,pp. 153-156, 2013.

[7] M. Shepperd and G. Kadoda, Comparing Software Prediction Techniques Using Simulation,IEEE Trans. Software Eng., vol. 27, no. 11, pp. 1014-1022, Nov. 2001.

[8] N. Mittas and L. Angelis, Comparing Cost Prediction Models by Resampling Techniques, J. Systems and Software, vol. 81, no. 5, pp. 616-632, May 2008.

[9] ImanAttarzadeh and Siew Hock Ow,A Novel Algorithmic Cost Estimation Model Based on Soft Computing Technique, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia,2010.

[10] Vahid Khatibi and Dayang N. A. Jawawi**,**Software Cost Estimation Methods: A Review,Faculty of Computer Science and Information System University Technology Malasia(UTM),JohorMalasia,vol. 2,no. 1,pp.1-9,2010-11.

[11] Salvatore Alessandro Sarcia, Victor Robert Basili,GiovanniCantone, Scope Error Detection and Handling concerning Software Estimation Models, pp. 1-8.

[12] Efi Papatheocharous, Harris Papadopoulos and Andreas S. Andreou, Feature Subset Selection for Software Cost Modeling and Estimation, Department of Computer Science, pp. 1-5.

[13] N. Mittas and L. Angelis, Ranking and Clustering Software Cost Estimation Models through a Multiple Comparisons Algorithm, vol. 39, no. 4, pp. 539-542, April 2013.