CLUMPING AND RANKING SOFTWARE COST ESTIMATION MODELS THROUGH MULTIPLE COMPARISONS ALGORITHM

1G. PARVATHI, 2K.L.N.C. PRAKASH

Student of Computer Science & Engineering, Student of Computer Science & Engineering
parvathi200328@gmail.com, klnc.praakash@gmail.com

Abstract—Software Cost Estimation can be described as the process of predicting the most realistic effort required to complete a software project. Due to the strong relationship of accurate effort estimations with many crucial project management activities, the research community has been focused on the development and application of a vast variety of methods and models trying to improve the estimation procedure. From the diversity of methods emerged the need for comparisons to determine the best model. However, the inconsistent results brought to light significant doubts and uncertainty about the appropriateness of the comparison process in experimental studies. Overall, there exist several potential sources of bias that have to be considered in order to reinforce the confidence of experiments. In this paper, 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 11 prediction models over 6 datasets. The results illustrate the benefits and the significant information obtained through the systematic comparison of alternative methods.

Keywords— Cost estimation, Management, Metrics/Measurement, Statistical methods

I. INTRODUCTION

The importance and the significant role of the Software Cost Estimation (SCE) to the well-balanced management of a forthcoming project are clearly portrayed through the introduction and utilization of a large number of techniques during the past decades. The rapidly increased need of large-scaled and complex software systems lead managers to settle SCE as one of the most vital activities that is closely related with the success or failure of the whole development process. Inaccurate estimates can be proved catastrophic to both developers and customers since they can cause the delay of the product deliverables or even worse to the cancellation of a contract.

Due to the abovementioned requirements, the interest has been focused on the open research problem of the selection of the “best” estimation technique. According to an extended systematic review of studies, the most common research topic of SCE is the introduction and evaluation of estimation methods. On the other hand, the variety of prediction methods is also associated with contradictory and inconsistent findings regarding the superiority of a technique over others. The most determining factor for these controversial results seems to be an inherent characteristic of prediction systems, i.e. their strong dependency on the kind of available data (types and number of project attributes and sample size) used in model fitting. The complexity of building an accurate model is swiftly increased, if we consider the alternative variations of a generic estimation method (e.g. regression analysis). In several studies the researchers base their inferences on a small number of datasets so generalization of findings may be quite misleading.

A certain limitation of several past studies is the comparison without using appropriate statistical hypothesis testing. This can lead to erroneous results and groundless generalizations regarding the predictive accuracy of estimation techniques. Although the comparison of methods without statistical tests may lead to unsound results, many recent papers still base their findings solely on single indicators.

Another source of bias can also be the statistical procedure that is used when comparing multiple prediction techniques. In the case of a simple comparison between two competitive models, the null hypothesis is examined via a classical statistical test (i.e. paired t-test or Wilcoxon signed rank test). With more than two comparative models, the meaning of "significant difference" becomes more complicated, and the problems associated with it, are known in statistics as the "multiple comparisons problems". Due to the large number of proposed cost estimation methods, it is necessary for project managers to systematically base their choice of the most accurate model on well-established statistical procedures in order to diminish the uncertainty threatening the estimation process. However, to the best of our knowledge, the problem of simultaneous comparisons among multiple prediction models has not been studied yet, in the sense that there is no statistical procedure which can identify the significant differences between a number of cost estimation methods and at the same time to be able to rank and cluster them, designating the best ones.
All of the issues discussed above lead us to conclude that there is an imperative need to investigate what is the state-of-the-art in statistics before trying to derive conclusions and unstable results concerning the superiority of a prediction model over others for a particular dataset. The answer to this problem cannot constitute a unique solution since the notion of “best” is quite subjective. In fact, a practitioner can always rank the prediction models according to a predefined accuracy measure but the critical issue is to identify how many of them are evidently the best, in the sense that their difference from all the others is statistically significant. Hence, the research question of finding the “best” prediction technique can be restated as a problem of identifying a subset or a group of best techniques. The aim of the paper is therefore to propose a statistical framework for comparative SCE experiments concerning multiple prediction models. It is worth mentioning that the setup of the current study was also inspired by an analogous attempt dealing with the problem of comparing classification models in Software Defect Prediction, a research area that is also closely related with the improvement of software quality. 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 engineering, financial and medical sciences. In the field of SCE it has not been used in a systematic manner yet. 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. In the present study, DOE analysis is 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.

The proposed statistical methodology is also based on an algorithmic procedure which is able to produce non-overlapping clusters of prediction models, homogeneous with respect to their predictive performance. For this purpose, we utilize a specific test from the generic class of multiple comparisons procedures, namely the Scott-Knott test, which ranks the models and partitions them into clusters.

The proposed statistical framework is applied on a relatively large-scale set of 11 methods over 6 public-domain datasets from the PROMISE repository and the International Software Benchmarking Standards Group (ISBSG). Finally, in order to address the disagreement of the performance measures, we apply the whole analysis on three functions of error that measure different important aspects of prediction techniques: accuracy, bias and spread of estimates. The rest of the paper is organized as follows: In Section 2 we summarize related work and we specify the contribution of the current study. In Section 3, we present the limitations of current approaches for multiple comparisons of models and we analytically describe the proposed procedure based on the Scott-Knott test. In Section 4, we demonstrate the experimental setup of this study in a systematic manner, whereas in Section 5 we accumulate the results of the analysis. In Section 6, we perform some sensitivity analysis for two small datasets. Finally, certain limitations of the study are given in Section 7, whereas in Section 8, we conclude by discussing the novel findings of the proposed framework.

II. RELATED WORK AND CONTRIBUTION

During the last decades there is an evolving research concerning the identification of the best SCE method. The researchers strive to introduce prediction techniques including expert judgment, algorithmic, statistical and machine learning methods. The usual practice of these studies is to compare the proposed estimation method with established models on a small number of datasets. Earlier studies based their inference regarding the superiority of a prediction method against a comparative one on accuracy measures computed through a validation procedure. Despite the novelty and the promising results of each estimation technique, the researchers’ interest has been rapidly focused on the problem of inconsistent findings regarding the determination of the best estimation method, while at the same time they started investigating the reasons behind unstable results.

Mittas and Angelis showed that the usual practice of promoting a model against a competitive one just by reporting an indicator, can lead to erroneous results, since these indicators are single statistics of error distributions, usually highly skewed and non-normal. In this regard, they proposed resampling procedures for hypothesis testing, such as permutation tests and bootstrap techniques for the construction of robust confidence intervals.

The Scott-Knott test presented here was used in another context in, for combining classifiers applied to large databases. Specifically, the Scott-Knott test and other statistical tests were used for the selection of the best subgroup among different classification algorithms and the subsequent fusion of the models’ decisions in this subgroup via simple methods, like weighted voting. In that study extensive experiments with very large datasets showed that the Scott-Knott test provided
the highest accuracy in difficult classification problems. Hence, the choice of the test for the present paper was motivated by former results obtained by one of the authors.

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 fit best to its data 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, like the Nemenyi test. 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 goal of this paper is to further extend the research concerning the comparison and ranking of multiple alternative SCE models. We propose a framework for conducting comparative experiments and we present an evaluation of this analysis over different datasets and prediction models. An algorithm based on the multiple comparison Scott-Knott test ranks and clusters prediction models by indentifying statistically significant differences between them.

III. LIMITATIONS OF ESTABLISHED PROCEDURES AND DESCRIPTION OF THE PROPOSED METHODOLOGY

In this section, we present the main aspects of the proposed methodology. First, we discuss problems of the procedures used to compare a set of candidate prediction methods. Second, we describe in detail the algorithm based on the Scott-Knott test, which addresses the limitations of the established techniques.

3.1 Problems with Comparison of Multiple Prediction Techniques

The important role of well-established statistical comparisons in SCE is highlighted in many recent studies, especially during the last decade where the findings are derived through formal statistical hypothesis testing. Indeed, the researchers use parametric as well as non-parametric procedures, whereas there has been also an increasing interest for more robust statistical tests such as permutation tests and bootstrapping techniques for the construction of confidence intervals.

The problem we address in this paper belongs to a generic class in statistics, known as "multiple hypothesis testing" and can be defined as the procedure of testing more than one hypothesis simultaneously.

Describing briefly the problem, the conclusions derived from a statistical hypothesis test are always subject to uncertainty. For this reason, we specify an acceptable maximum probability of rejecting the null hypothesis when it is true and this is referred as a "Type I error". In the case of multiple comparison problems, when several hypotheses are carried out, the probability that at least a Type I error occurs, increases dramatically with the number of hypotheses.

The problem can be easily described through the following example. Suppose we wish to compare five hypothetical models setting the significance level at \( \alpha = 0.05 \). For each test of an overall set of ten pair-wise tests, the probability of making a Type I error equals \( \alpha = 0.05 \) and therefore the probability of not making a Type I error equals 1 — 0.05 = 0.95. In the case of ten comparisons the probability of no Type I error is 0.95\(^10\) ~ 0.60. So, with a level of \( \alpha = 0.05 \) for each of the ten tests, the probability of erroneously rejecting a null hypothesis is 0.40.

The problem of this error inflation can be treated by adjusting the overall (or family) error, but still the execution of a large number of pair-wise comparisons is not so straightforwardly interpretable since the resulting groupings are overlapping. For this reason, various intelligent techniques have been proposed in the statistical literature in order to perform targeted multiple comparisons, i.e. comparisons that answer specific research questions.

3.2 Experimental Design

The whole study is organized according to the formal principles of the experimental design or design of experiment (DOE). DOE refers to a systematic planning in order to maintain control over all factors that may affect the result of an experiment. We have to point out that the term "experimental design" is often used informally in SCE by the researchers to describe generally the conditions, the assumptions, the data and the procedures of a study.

DOE constitutes an entire branch area in statistics involving fundamental concepts that have to be specified and controlled in advance. The basic element of a DOE is the experimental unit which is
the “object” on which the researcher wishes to measure a response variable. The purpose is to study the effect of one or more factors (categorical variables) on the response variable. The different categories of a factor are known as levels or treatments.

In the case of our experimental setup, the predictive performance of each competitive model is evaluated through a k-fold cross-validation approach in which the original dataset is randomly partitioned into k subsamples of equal size. During a repeated procedure, each one of the subsamples is considered as the validation sample (test set) and the remaining k−1 subsamples as the training sets used for fitting the models. Then, the local measures of error (Table1) are computed for each project of the test set. Each test set gives a global measure of error. The error measures from all the experimental runs are transformed so as to be normally distributed. These values are used as an input for the Scott-Knott algorithm. In our study, we used k = 10 as the number of folds.

Following the terminology of a DOE, the k = 10 different folds of the abovementioned procedure can be considered as the experimental units of our context, the comparative prediction models represent the treatments and the response variable is the normalized expression of a measure in Table 2. The purpose of the experiment is to investigate the effect of different treatments (models) on a response variable (error measure), i.e. to test the differences of the predictive performance of different models.

<table>
<thead>
<tr>
<th>Prediction Model</th>
<th>Block</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set 1</td>
<td>MemE1A</td>
<td>MemE1B</td>
<td>MemE1C</td>
<td>MemE1D</td>
<td></td>
</tr>
<tr>
<td>Test set 2</td>
<td>MemE2A</td>
<td>MemE2B</td>
<td>MemE2C</td>
<td>MemE2D</td>
<td></td>
</tr>
<tr>
<td>Test set 3</td>
<td>MemE3A</td>
<td>MemE3B</td>
<td>MemE3C</td>
<td>MemE3D</td>
<td></td>
</tr>
<tr>
<td>Test set 4</td>
<td>MemE4A</td>
<td>MemE4B</td>
<td>MemE4C</td>
<td>MemE4D</td>
<td></td>
</tr>
<tr>
<td>Test set 5</td>
<td>MemE5A</td>
<td>MemE5B</td>
<td>MemE5C</td>
<td>MemE5D</td>
<td></td>
</tr>
<tr>
<td>Test set 6</td>
<td>MemE6A</td>
<td>MemE6B</td>
<td>MemE6C</td>
<td>MemE6D</td>
<td></td>
</tr>
<tr>
<td>Test set 7</td>
<td>MemE7A</td>
<td>MemE7B</td>
<td>MemE7C</td>
<td>MemE7D</td>
<td></td>
</tr>
<tr>
<td>Test set 8</td>
<td>MemE8A</td>
<td>MemE8B</td>
<td>MemE8C</td>
<td>MemE8D</td>
<td></td>
</tr>
<tr>
<td>Test set 9</td>
<td>MemE9A</td>
<td>MemE9B</td>
<td>MemE9C</td>
<td>MemE9D</td>
<td></td>
</tr>
<tr>
<td>Test set 10</td>
<td>MemE10A</td>
<td>MemE10B</td>
<td>MemE10C</td>
<td>MemE10D</td>
<td></td>
</tr>
</tbody>
</table>

3.3 The Scott-Knott Test
The Scott-Knott test is a multiple comparison procedure based on principles of cluster analysis. The clustering refers to the treatments (methods or in our case models) being compared and not to the individual cases, while the criterion for clustering together treatments is the statistical significance of differences between their mean values. Our preference for the Scott-Knott test relies to a specific desirable characteristic of the method, i.e. that it is able to separate the models into non-overlapping groups. In our case the values of the response variable that is affected by the models are translated to expressions of the prediction errors derived from the models being compared. The algorithm we describe next is therefore able to rank and cluster prediction models according to their accuracy.

Suppose that we want to compare the predictive accuracy of d alternative models on a specific dataset through the utilization of a functional expression of the error e . We also assume that following a standard validation procedure there are k pairs of training-test datasets, i.e. the original dataset is divided k times in training-test subsets. All models are applied repeatedly to each one of these pairs, i.e. each model is trained and validated using the same pair. The predictions of the test dataset yield an overall measure of accuracy which is the value of the response variable for the unique combination of a model and a training-test dataset. Therefore, for a specific dataset we have d × k measurements.
IV. EXPERIMENTAL SETUP

This section provides details concerning the setup of the framework and the experimental design of the study. The basic idea of the experimental setup was to take into account: (a) different cost prediction methods covering a major part of the variety of the proposed methodologies appeared so far in the SCE literature and which are governed by a diversity of principles, (b) different datasets and (c) different measures of error. Moreover, the experiment was designed to take into account the effect of training-test splitting of each dataset.

4.1 Comparative Prediction Models

The eleven selected methods can be grouped into three main categories that are regression-based models, analogy-based techniques and machine learning methods.

All these models are well-established methods, there is a vast literature on them and they have been already applied in SCE. In Table we present a list of the methods being compared along with a brief description, in order to give the basic principles of each method.

The choice of the alternative prediction techniques was also based on the conclusions of a systematic review on SCE studies. Jørgensen and Shepperd pointed out that the regression-based models dominate, since half of all studies deal with the problem of fitting, improvement or comparison of a regression model. Furthermore, they claim that the researchers' interest for the analogy-based techniques is steadily increased during the last decade. Finally, the distribution of estimation methods also reveals that the proportion of machine learning techniques (i.e. Classification and Regression Trees and Neural Networks) presents an increasing trend.

4.2 Datasets

The datasets for the experimentation are derived from two sources, namely the PROMISE repository and the International Software Benchmarking Standards Group (ISBSG, release 10).

The main reason for this selection was that these datasets have been extensively used to empirically validate or justify a large amount of research results, whereas they are also publically available. Each dataset contains different number of projects and a set of independent variables with mixed-type characteristics, whereas the dependent variable that has to be predicted is the actual effort. Another criterion for the selection of the datasets was the ability to apply all the competitive prediction methods on them. Therefore, we did not...
consider datasets with too many categorical variables which cause problems to certain methods like regression and Neural Networks.

The ISBSG repository contains 4106 software projects from more than 25 countries but most of the variables have a large amount of missing values. Having in mind the guidelines of ISBSG suggesting filtering of the data projects, we decided to discard the projects with missing values. Moreover, an important issue in SCE is the utilization of datasets with high quality in the process of evaluation and comparison of prediction models. Due to this fact and the instructions of ISBSG organization that point out not taking into account projects with low quality, we based our analysis on projects marked with "A" in Data Quality Rating and UFP Rating. Finally, the independent characteristics utilized in the construction of the alternative models are the same as in order to retain the compatibility with other studies.

<table>
<thead>
<tr>
<th>Name</th>
<th>#projects</th>
<th>#categorical</th>
<th>#scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kemerer (KEM)</td>
<td>15</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Telecom (TEL)</td>
<td>18</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Albrecht (ALB)</td>
<td>24</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>Miyazaki (MIY)</td>
<td>48</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>Desharnais (DES)</td>
<td>77</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ISBSG10 (ISBSG)</td>
<td>506</td>
<td>-</td>
<td>10</td>
</tr>
</tbody>
</table>

### 4.3 Accuracy Measures

In Section 2, we have already mentioned that the overwhelming majority of SCE studies base their inferences on MMRE. However, MMRE has been criticized as an inappropriate measure since it tends to favor models that underestimate the actual cost of projects.

During the last decade there has been a thorough discussion concerning the determination of the most appropriate error function without essential convergence. Having in mind that all the accuracy indicators exhibit certain advantages while on the same time they suffer from either a flaw or a limitation, the basic key for resolving the problem is to realize what is really measured by each error function.

The lesson learned from the abovementioned discussion is that there is 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 under- \((z < 1)\) or over- \((z > 1)\) estimations with an optimum value of 1. The most widely known MRE indicator was also used since according to Kitchenham et, it provides a measure of the spread of the error ratio \( z \). The local measures of error that are computed through the actual \((Y_A)\) and the estimated \((Y_E)\) 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.

### V. EXPERIMENTAL RESULTS

In this section, we present the results of the experiments conducted on six datasets. Summarize the results of all experimental runs. Each dataset corresponds to a different DOE where each model was applied to 10 test sets and then the values of the error measures were transformed and used as an input for the Scott-Knott algorithm. So, the ranking was based on the means of the transformed values and not the means of the original values of the global measures.

Although the rankings of models clearly portray an overview of their predictive performance, it is essential to statistically test whether some models are superior to others, since these observed differences could reasonably occur "just by chance". Next, we present the results of the formal comparison of models through the inferential statistical procedures of RCBD and Scott-Knott test. Generally, the application of Scott-Knott tests for all datasets reveals one of the most appealing findings of this study: Despite the large divergences of error functions among alternative prediction models, there are no statistical evidences that some methods differ significantly. Hence, the notion of the "best" estimation technique should be revised, whereas at the same time it is probably more proper to refer to the "best group of estimation techniques".

Moreover, the size of the samples, in which the alternative prediction models are built, plays a critical role in SCE experiments. This pattern in the results also brings to light a conviction concerning the vain of using many estimation methods on small datasets, since they cannot be more informative despite their promising and sophisticated principles.

### CONCLUSIONS

In this paper, we deal with a critical research issue in software cost estimation concerning the
simultaneous comparison of alternative prediction models, their ranking and clustering in groups of similar performance. We examined the predictive power of 11 models over 6 public domain datasets. The whole procedure is settled on well-established statistical methodologies taking into consideration the multiple comparison problems. Having in mind the critical role for the adoption of reliable practices in the development process for both project managers and customers, we proposed a formal framework and structured guidelines in order to reinforce the knowledge acquisition and diminish the inherent uncertainty in SCE.

In this regard, we proposed certain directions concerning the utilization of alternative prediction models from different classes of estimation techniques, different datasets from public domain repositories, alternative error functions measuring different aspects of predictive performance, an experimental design in order to overcome the problem of splitting the datasets in the validation process and more importantly an efficient hypothesis testing procedure that signifies whether a set of prediction models gives statistically better results than another set of comparative models.

We have to emphasize that the experimentation section is used as means for illustrating how the whole framework can be evaluated on the comparison setup contributing to the systematic research of the performances of any kind of prediction techniques. Thus, it is not our purpose to determine the superiority of any prediction method and even more, it is not wise to generalize the derived findings for the population of software projects.

On the other hand, the derived results of our experimentation either bring to surface a few significant conclusions or confirm other essential results from past studies. Although the indicators from alternative models designate generally different predictive performances for the datasets, the proposed statistical hypothesis testing through Scott-Knott test verifies that the predictive accuracy of a set of methods does not confirm a statistically significant difference among them. This suggests that the notion of the "best" estimation method may have not been so well defined so far in the previous SCE research. Hence, there is a need to relocate the basis of the whole research.

In other words, when a practitioner wishes to perform a comparative study, she or he ought to seek for a set of best estimation models and not just a single one. Moreover, it seems that there is not a global solution, since alternative methods can exhibit a few advantages in terms of certain aspects of prediction performance but on the same time these candidates may suffer in terms of another aspect of accuracy. In order to overcome these inconsistencies, the managers should utilize their experience and lead the whole process through the necessities that are arisen in each case. Therefore, it is important to emphasize here that the proposed method is an aid to the process of decision making by ranking and clustering the candidate models. However it does not make decisions itself. The final decision is left to the expert and depends on several issues, even on personal criteria like experience, preference of statistical software, etc.

Another interesting finding concerns the utilization of complicated and more sophisticated models. It seems that very often a linear model is adequate enough to catch the trend between effort and other cost drivers of projects. Therefore, in certain cases it may be useless to strive to introduce new, highly complicated algorithms which in practice they just cannot provide any further improvement. Finally, it is our strong belief that new estimation techniques should be tested and compared using appropriate statistical procedures.

**REFERENCES**


