2012 International Conference on Information and Knowledge Management (ICIKM 2012)
IPCSIT vol.45 (2012) © (2012) IACSIT Press, Singapore

Systematic Literature Review of Missing Data Imputation Techniques for Effort Prediction

Beatrice Andrew [+] and Ali Selamat
Dept. of Software Engineering Faculty of Computer Science and Information System
Universiti Teknologi Malaysia Skudai, Johor 81310 Malaysia

Abstract. Background: The occurrences of missing data in software project data set bring significant impact on effort prediction. The significant impacts of this problem are wasted information and biased analysis. The common studies conducted to investigate in depth this area is to find sophisticated missing imputation techniques but there are only small numbers of research questioning the quality of study. Aims: This paper aims to systematically analyze researches done on missing data imputation techniques to estimate software effort and review the current approaches exist in the field. Method: We performed a systematic literature of empirical studies of data imputation techniques published in from 2000- 2012.

Keywords: software effort prediction, imputation technique, systematic literature review

1. Introduction

Accurate and unbiased software effort prediction is an important contributor to effective software project management [1]. Whole-project effort prediction is clearly important in terms of enabling software developers/managers to make a reasonable bid or form a plan of activities consequently, an extensive body of research has addressed this facet of project management .Like any real world data sets, incomplete or missing data is an unavoidable. Missing values result in less efficient estimates because of sample bias and reduced sample size. Most data mining algorithms cannot work with incomplete datasets. Hence, missing value imputation is mandatory. In statistics, imputation is the substitution of some value for a missing data point or a missing component of a data point. Once all missing values have been imputed, the dataset can then be analyzed using standard techniques for complete data. For analyzing the available data, its completeness and quality play a major role, because the inferences made from a complete data are more accurate than those made from an incomplete data [1]. For example researchers rarely find the survey dataset with complete entries [2]. The respondents may not give complete information because of negligence, privacy reasons or ambiguity of the survey questions. The missing parts of variables may be important things for analyzing the data. So in this situation data imputation plays a major role. The purpose of this study is to review and summarize the existing works done on missing data imputation techniques to improve effort prediction accuracy. Section 2 discusses related works and methodology is discussed in Section 3. Section 4 shows the results of the review. Section 5 discusses the findings and the summary is shown in Appendix I. The last section concludes this review.

2. Related works

Twala and Cartwright [3] analyze the ensembles of imputation methods in order to improve effort prediction accuracy and classifier learning efficiency. The issue addressed in the study is the impact of missing value to the prediction accuracy. In 2010, the same study under the same topic published addressing the concern of good quality data is required to enhance prediction accuracy [4].

[+] Corresponding author. Tel.: +06 016-3115747.
E-mail address: bea.andrew89@gmail.com.

Twala et al. [5] investigate the randomization of decision tree building algorithms to improve prediction accuracy. The main objectives are investigating the impact of missing values to prediction accuracy and how the ensemble missing data techniques could be implemented to improve effort prediction accuracy. However, the results do not support when the data set is small with many attributes and gives different performance as the proportion of the missing data is increased. Molloken and Jorgensen [6] mentioned that most of the predictions done in a software project are based on expert judgments because there is no evidence that a formal prediction models will lead to better prediction accuracy. Survey on the effort prediction done with and without unbiased data and the study reported several issues including project overrun and expert estimation being the most frequent used estimation technique.

Menzies and Shepperd [7] address the repeatable issues specifically in software engineering prediction. The unfortunate conclusion drawn from the study is the conclusion instability which means the conclusion stands true to one project but does not hold in other projects. [7] propose nine strategies to reduce conclusion instability. Unlike previous works, this paper explores the systematic review on the imputation techniques have been used predict effort in software engineering and analyze their strengths and weaknesses in terms of accuracy, robustness, and generalization.

3. Review process

The review processes follow the SLR guidelines for software engineering by Kitchenham et al.[8]. The guidelines consist of three phases: review planning, review execution, review reporting phase. The review protocol is shown in Figure below to assist the review process and the following sub-sections elaborate review planning phase.

3.1. Research questions

The research questions are formulated using criteria considered in this SLR: population, intervention, comparisons, outcomes, and context. Table 1 shows the criteria and scope of research question structure. Based on the criteria, the research questions are as below:

RQ1.How many reviews had done on the imputation techniques to improve effort prediction accuracy and what kind of studies involved?

RQ2.Who is leading researchers on data imputation technique for software effort prediction?

RQ3. What are the limitations of current research?

Table 1. Structure of research question

Criteria	Scope
Population	Review on SLR of imputation techniques for effort prediction
Intervention	Issues addressed in related studies
Comparisons	Strength/s and weakness/es of existing approach
Outcomes	Suggests improvement in future research
Context	Missing data imputation techniques on software engineering prediction

3.2. Search strategy

This paper uses search strategy that comprises of the following steps:

1) *Preliminary search in major indexing databases*: by first identifying major keywords such as "review on imputation techniques", "review on imputation techniques for effort prediction" and "review on software engineering prediction".
2) *Research in major indexing databases*: use refined keywords which includes Boolean terms such as AND and OR in the initial keywords. The major indexing databases are Google Scholar, ScienceDirect, ISI Web of Knowledge, ACM Digital Library, Springerlink, and CiteSeerX.
3) *Record search results*.
4) *Classify papers according to types of publication*: The obtained search papers grouped and sorted according to journal, conferences, article in press, and book chapters.

3.3. Selection criteria

This phase involves ranking the source of papers from highest to lowest priority. The reviewed papers are mainly in English .The subject covered in this review is software engineering and computer sciences. All papers should explicitly contain text that attempts to define, propose, suggest, or describe a review on imputation techniques for effort prediction accuracy.

3.4. Qualitative analysis

The tabulation of data synthesis involves collating and summarizing the results of the included primary studies. The data tabulated to show the results obtained accordingly to the research questions .Appendix I shows the summary of the analysis for systematic literature review on imputation techniques in software engineering prediction. It will also highlight the comparisons and issues addressed by each study.

4. Results

After search being done through major indexing databases, there are no original reviews performed on how many systematic reviews or systematic literature reviews on imputation techniques for effort prediction accuracy. However, there are related studies found on reviewing the techniques. In the early 2000 until 2005, two reviews regarding the ensemble imputation methods done. Then there three studies on ensemble imputation techniques to improve prediction accuracy in 2005 and 2007. No increment in number of studies found from 2010 to latest May 2012, there are three studies conducted on missing data imputation techniques for effort prediction and no publication is made in 2011.

5. Discussions

5.1. How many reviews had done on the imputation techniques to improve effort prediction accuracy and what kind of studies involved?

This study includes seven researches that are doing review on imputation techniques for effort prediction accuracy. The earliest study is performed in 2005 by [2] No systematic review is done on imputation methods from 2000 to May 2012. From the selected studies, several types of imputation techniques that had been applied to estimate software development effort. They are listed as follows:

- ignoring methods
- mean imputation
- hot deck methods
- multiple imputation
- single imputation
- likelihood methods

Among the above listed missing data imputation techniques, ML based, HDI and RI are the most frequently used ones; they together were adopted by 80% of the selected studies, as illustrated in Fig. 1. The

study has shown that the publication peak appears around year 2006-2008 if based on the number of related research. Note that some studies contain more than one imputation techniques and types.

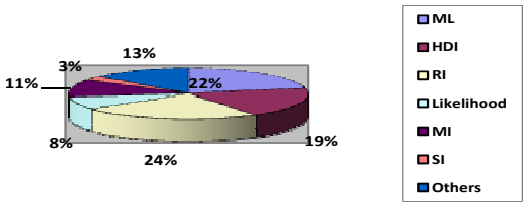

Fig. 1: Distribution of the studies over type of Imputation Techniques.

Wasito and Mirkin combine HDI, Likelihood Imputation with machine learning based imputation [9]. Song and Shepperd combine k-NN imputation with CMI which results to MINI algorithm, Sentas and Angelis used Kernel based imputation which is a machine learning method with RI [8]. Zhang and Wang adopted Expectation Maximization (EM) which is a likelihood method with Bayesian model which is a machine learning based model [10].

There are some fairly new techniques being evaluated especially after 2008 afterward such as Genetic Algorithm based imputation, Kernel based, Multi-Layer Perceptron and Neural Network with the concern of empirical evaluation of estimation accuracy such as parameter, feature subset, and outlier presence[8][11][12].

5.2. Who's leading researchers on data imputation technique for software effort prediction?

Most of the research topic related to empirical evaluation of imputation techniques and software effort prediction are done by renowned authors such as Shepperd, Twala, Cartwright, Menzies, Jorgensen, Song, and Scheffer with respect to the number of primary study. Martin J. Shepperd is the leading editor of PROMISE conference specialized in empirical software engineering which had done more than 20 studies and his studies were cited in many occasions. Most of these researchers are co-author with each other in more than 14 studies.

5.3. What are the limitations of current research?

As has been shown in this review, there are limited numbers of study performed on systematic literature review specifically in the discipline of software engineering prediction that focus on imputation techniques. Existing studies are mostly comparisons works, empirical evaluation on existing techniques and combining approaches to increase the imputation method's performance. This review has revealed that, on one hand, most of the available studies are carried out by the same researchers and none of the work assesses the quality of study regarding this field relatively large numbers of primary studies relate to practice rather than questions concerning the practices and techniques. This is not encouraging since this area is available for improvements.

In terms of research trends, [2][7][9][13-21] focus on the algorithms used to improve imputation's performance and this type of study usually involves comparisons work. Unsurprisingly, this research trends attracts more primary studies.

6. Conclusion and future works

This study finally identified eight primary studies on the review of imputation techniques for effort prediction that are pertaining to the three research questions (RQs) raised in this review. Although the studies are cited in many other primary and secondary studies, it is clear the topics related to SLR and the

assessment of quality of primary studies is limited. It is also important for the leading researchers to increase the publication in SLR of imputation techniques to increase possible valuable researches.

7. References

[1] M. O. Elish (2009). "Improved estimation of software project effort using multiple additive regression trees." Expert Systems with Applications **36**: 10774–10778.

[2] P. Garc(2009). "K nearest neighbours with mutual information for simultaneous classification and missing data imputation." Neurocomputing **72**(7-9): 1483-1493.

[3] M. Cartwright, B. Twala. and. Menzies. (2005). Ensemble Imputation Methods for Missing Software Engineering Data. METRICS

[4] M. Cartwright, B. Twala . (2010). "Ensemble missing data techniques for software effort prediction." Intelligent Data Analysis **14**: 299-331.

[5] B. Twala, M. Cartwright., and M. Shepperd (2006). Ensemble of Missing Data Techniques to Improve Software Prediction Accuracy. ICSE.

[6] M.Jørgensen and M.Shepperd (2007). "A Systematic Review of Software Development Cost Estimation Studies." IEEE Transactions on Software Engineering **33**.

[7] T. Menzies and M. Shepperd (2012). "Special issue on repeatable results in software engineering prediction." Empirical Software. Engineering. **17**(1-2): 1-17.

[8] B.Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman (2009). Systematic literature reviews in software engineering- A systematic literature review. Journal of Information and Software Technology **51**: 7-15.

[9] B.Mirkin and I. Wasito(2006). "Nearest neighbours in least-squares data imputation algorithms with different missing patterns." Computational Statistics & Data Analysis **50**: 926 – 949.

[10] K. Molloken, . and M. Jorgensen (2003). A Review of Surveys on Software Effort Estimation. Proceedings of the 2003 International Symposium on Empirical Software Engineering, IEEE Computer Society**:** 223.

[11] P. Sentas, and L. Angelis. (2006). "Categorical missing data imputation for software cost estimation by multinomial logistic regression." Journal of Systems and Software **79**(3): 404-414.

[12] J. Scheffer(2002). "Dealing with Missing Data." R.L.I.M.S **3**.

[13] I. Wasito, B. Mirkin. (2005). "Nearest neighbour approach in the least-squares data imputation algorithms." Information Sciences **165**: 1-25.

[14] Z. Shichao; J. Zhi; Z. Xiaofeng; Z. Jilian (2009). "Missing Data Analysis: A Kernel-Based Multi-Imputation Approach." Transactions on Computing Science **8**.

[15] G. Ridgeway and D. Madigan. (2003). "A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets." Data Mining and Knowledge Discovery **7**(3): 301-319.

[16] B.Mirkin and I. Wasito (2005). "Nearest neighbour approach in the least-squares data imputation algorithms." Information Sciences **165**: 1-25.

[17] M. Shepperd and Q. Song . (2007). "A new imputation method for small software project data sets." The Journal of Systems and Software **80**: 51-62.

[18] S. G. MacDonell and M. J. Shepperd. (2003). "Combining techniques to optimize effort predictions in software project management." The Journal of Systems and Software **66**: 91-98.

[19] I. Myrtveit and E. Stensrud. (2012). "Validity and reliability of evaluation procedures in comparative studies of effort prediction models." Empirical Software Engineering.

[20] P. Jönsson and C. Wohlin. (2004). An Evaluation of k-Nearest Neighbour Imputation Using Likert Data. METRICS.

[21] Y.S. Seo, K. A Yoon, D. H. Bae (2008). An Empirical Analysis of Software Effort Estimation with Outlier Elimination. PROMISE '08. Liepzig, Germany.

[22] E. Mendes and S. Counsell (2006). Web Effort Estimation. Web Engineering, Springer.

[23] J.Moses. and M. Farrow (2005). "Assessing Variation in Development Effort Consistency Using a Data Source with Missing Data." Software Quality Journal **13**(1): 71-89.

[24] Q. Song , C. Xiangru , C.L. Jun (2008). "Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation." The Journal of Systems and Software **81**.

[25] G. A. Liebchen and Martin Shepperd. (2008). Data Sets and Data Quality in Software Engineering. PROMISE '08. Leipzig, Germany.

[26] K. Mollokken and M. Jorgensen. (2003). A Review of Surveys on Software Effort Estimation. Proceedings of the 2003 International Symposium on Empirical Software Engineering, IEEE Computer Society**:** 223.