International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

Systematic Mapping Study of Missing Values Techniques using Naive Bayes.

Tejal Patil

ME Student, Department of Computer Science Engineering,

G H Raisoni Institute of Engineering and Management Jalgaon, Maharashtra, India

Abstract- *Missing Values present a common problem facing research in software engineering which is mainly based on statistical or data mining analysis of Software engineering data. The method of handling with Missing values is to ignore data with the missing observations. This leads to losing valuable information and then obtaining biased results. There are many techniques have been developed to deal with Missing Value, especially those based on imputation methods.*

In Proposed Work the Naive Bayes Classifier is use For Classification in this study. Naive Bayesian Classifier is Popular Classifier; not only for its good performance, but also for its simple form it is not sensitive to missing data and the efficiency of calculation is very high.

The study shows an becoming more intense in machine learning techniques especially the K-nearest neighbour algorithm (KNN) to handle with Missing values in SE datasets and found that most of the Missing Value techniques are used to serve software development effort estimation techniques.

Keywords— *Systematic mapping study, Missing values, Machine learning, Naive Bayes*

1. INTRODUCTION

Nowadays, larger historical software project datasets are often used with a higher number of missing values for a significant number of variables, therefore making their use rather challenging for research aims [3][4]. Thereby, a proper handling of Missing value is necessary when analyses are performed in a domain, such as empirical Software Engineering, where accuracy and precision are key factors. Indeed, up until the early 2000's, most of the empirical research in Software Engineering field have been carried out with small samples. Various approaches have been examine to handle Missing Value in SE datasets, such as (1) deletion techniques which delete the data with the missing values from the datasets and use only when the dataset complete ones [5][6]; (2) toleration techniques which perform analysis directly on the incomplete data sets. there are no need to complete dataset [7]; and (3) Imputation techniques which first fill the incomplete features or data and then do analysis on complete data sets [4][8][9]. It is important to consistently identify, classify and analyze the state of art and provides an overview of the trend in the field of Missing value techniques in SE data sets. This paper presents a systematic mapping of Missing Value techniques in SE datasets. A mapping study is the defined method building a classification scheme and structuring a field of interest, in order to acquire an overview of existing approaches, to resume the coverage of the research

area in different facets of the classification scheme and to use the identified lacunas as basis for future research directions". To the right of our knowledge, this study is the first systematic mapping study in the area of missing Value techniques in SE datasets. The dividation of this paper are:(1) A classification scheme categorizing the data in the field of Missing Value techniques in Software Engineering datasets;(2) A systematic mapping study of Missing value techniques in SE, structuring related research work over the past decade by analyzing 30 selected papers; and (3) An analysis of the demographic trends in the area of Missing Value techniques in SE datasets;(4) A repository of the papers collected and analyzed through this systematic study. The results summarize the existing MV techniques in SE datasets, the types of Missing Value treated, the type of data analyzed, and the objective investigated behind the use of MV techniques. Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high. Bayesian Iteration Imputation uses Naive Bayesian Classifier to impute the missing data. It is consisted of two phases: a) Decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, weighted index, etc.; b) Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data and the times for iterative can be reduced.

2. Related work

2.1. Research Methodology in Missing Value

A systematic mapping study is used for provide an overview of the research area, to investigate the existence of research of interest in research methodology there are some operation are to be performed. And these are mapping questions, search strategy, study selection, Quality assessment, and data extraction in this study there are some mapping questions to investigate the work of mapping study the aim of the study is to, analyze, identify and synthesize the work published during the past decade in the field of Missing Value techniques in Software Engineering datasets. The mapping questions are formulated in such a way that they reflect the main purposes of this systematic mapping [11].

The studies were identified by consulting six online libraries: (1) IEEE xplore, (2) ACM digital libraries, (3) Science Direct, (4) Springer Link and (5) DBLP, (6) Google Scholar.

The search in the six electronics databases resulted in 280 candidate papers. To identify selected papers, this screening was performed by applying two selection phases and these phases are Phase 1: Inclusion and exclusion criteria which are used to identify the relevant papers. Phase 2: Quality assessment criteria which are applied to the relevant papers of phase 1 so as to select the papers with acceptable quality, which were eventually used in the data extraction. It defined the following

inclusion and exclusion criteria, which have been refined through pilot selection.

Inclusion criteria: (1) Use of missing value technique to deal with Missing value in Software Engineering datasets; (2) Comparison between two or more Missing Value techniques using SE datasets (3) Using of this techniques for predicting and evaluating any software project attribute (4) Where several papers reported the same study, only the most recent was included.

Exclusion criteria: (1) Studies dealing with MV but are not in the SE domain; (2) Papers treating Missing Values in contexts other than experimentation in SE (3) Studies not available in full text (4) Studies not in English language (5) Papers published before 2000. By applying the inclusion and exclusion criteria in selection phase 1, we identified 32 relevant papers. Then, by scanning the references of these 32 relevant papers, 6 extra Relevant papers were added in the initial search. Therefore, we identified in total 38 relevant papers. In phase 2, the quality assessment criteria were applied to the 38 relevant papers and 35 papers were finally selected which were then used in data extraction. Quality assessment is usually carried out in systematic literature reviews, but less often in systematic mapping studies. Data extraction the 30 selected studies ware used to collect data that would provide the answer that shows the data extraction from which was created as an excel sheet and filled in by one another for each of the paper selected.

2.2Results and discussion: in this section the mapping results and presents the answer to the publication channels. That included publication channels, publication trend, research types and research approaches, missing value techniques in that missing value techniques two imputation techniques are involved ML (machine learning) and non ML(Machine learning) and ML(machine learning) imputation techniques distribution.

The journals and conferences in which the selected papers of this systematic mapping study ware published. Conferences are the main target venue of studies on MV techniques in SE dataset in the field of publication trend it presents the number of articles published per year. According to [17],[18] carried out the first study on the effect of missing data techniques on software development effort prediction. In Missing value techniques toleration techniques discussed in four paper and mainly used in comparison with the imputation techniques[3][7][8][9].

3. Proposed Work

In the systematic mapping study of missing value techniques there are three techniques are used for finding the missing values these are deletion techniques, imputation techniques and toleration techniques. In the proposed work naive Bayes classifier are used for classification. Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high. Bayesian Iteration Imputation uses Naive Bayesian Classifier to impute the missing data. It is consisted of two phases: a) first decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, and weighted

index b) By Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data.

Naive Bayes learners and classifiers can be very fast compared to more another methods. The notion of classification is very general and has many applications. For instance, in computer vision, a classifier may be used to divide text into classes such as text data, numeric value. Naive Bayes model is easy to build and particularly useful for very large data sets. It is easy and fast to predict class of test data set. It also performs well in multi class prediction it first decide the order of the attribute to be treated according to same measurement such as information Gain, missing Rate, and weighted index.

4. Conclusion

Aim of the systematic mapping study was examine the current research on the use of MV techniques in SE datasets by selecting 3o papers from a total of 280 candidate articles. These 350 selected studies are classified according to the six classification criteria and these classification criteria are the research approaches, research type, MV types, and objectives using MV techniques. Publication channels and trends are identified. The principal of the findings this study following. Conferences are the main target of the missing values. The time periods for included articles extend from 2000 to 2012 and the trends of MV publications in SE are characterized by discontinuity since no paper was published in 2002 and 2011. Most of the papers presented of MV techniques, solutions propose were presented as well as only two papers gave opinions about Missing Values techniques. Most of the selected studies belong to the history based evaluations research approach and it is the most used dataset is the international software benchmarking standard group (ISBSG) Dataset. The results reveal that imputation techniques were the most investigated, followed by deletion methods and then toleration techniques. The results indicate that 54% of selected papers treated only MCAR (Missing completely at random), which means they assumed that data are missing at randomly. However most papers taking the missing mechanisms into account while investigating MV techniques and show how these mechanisms can significantly affect their performance. The main motivation behind investigating in SE dataset was to predict software development effort.

The study help to practioners to identify techniques with which to extend the Missing Value treatment in their projects, and it may also help researchers to identify both the datasets to be used in the valuation of their studies and channels in which to publish their research result. In This study naive Bayes classifier help make the classification easy to find of missing values in software engineering data. Because the efficiency of the naive Bayes classifier is very high.

References

[1]Ali Idri Ibtissam Abnane, Alain Abram.Systematic Mapping Study Of missing values Techniques In Software Engineering Data

[2] ISBSG, Data R8. International Software Benchmarking Standards
Group, www.isbsg.org, October 18, 2005.

[3] S. Panagiotis, A. Lefteris. Categorical missing data imputation for
software cost estimation by multinomial logistic regression. The journal
of system and software 79, 2006. 404-4014
http://promise.site.uottawa.ca/SERepository.

[4] M. Cartwright, M.J. Shepperd, Q. Song. Dealing with Missing Software
Project Data. Proc. 9th IEEE International Software Metrics Symposium
(Metrics'03), Sydney, Australia, 2003, pp.154-165

[5] R.F. Olanrewaju, W. Ito. Development of an imputation technique - INI
for software metric database with incomplete data. 4th Student
Conference on Research and Development, 2006, pp 76 – 80.

[6] A. Bala. Impact analysis of a multiple imputation technique for handling
missing value in the ISBSG repository of software projects, Ph.D thesis,
supervised by Alain Abran, Ecole de technologie superieure – ETS,
University of Quebec, Montreal Canada, 2013.

[7] K. Tamura, T. Kakimoto, K. Toda, M. Tsunoda, A. Monden, K.
Matsumoto. Empirical Evaluation of Missing Data Techniques for
Effort Estimation, 2009. DOI= doi:
10.1.1.145.780<http://citeseerx.ist.psu.edu/viewdoc/summary?>.

[8] Q. Song, M. Shepperd, X. Chen, J. Liu, 2008. Can k-NN imputation
improve the performance of C4.5 with small software project data sets?
A comparative evaluation. Journal of Systems and Software, Volume 81,
Issue 12, December 2008, pp. 2361–2370.

[9] L. Jingzhou , A. Al-Emran, G. Ruhe,. 2007. Impact Analysis of Missing
Values on the Prediction Accuracy of Analogy-based Software Effort
Estimation Method AQUA, First International Symposium on Empirical
Software Engineering and Measurement, 2007. Pp 126 – 135

[10] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson. Systematic mapping
studies in software engineering, 12th International Conference on
Evaluation and Assessment in Software Engineering (EASE), 2008, pp.
71–80.

[111] K. El-Emam, A. Birk . Validating the ISO/IEC 15504 Measure of
Software Requirements Analysis Process Capability. IEEE Transactions
on Software Engineering, Volume:26 , Issue: 6 , 2000, pp 541 – 566.

[12] B. Twala , M. Cartwright. Ensemble missing data techniques for

software effort prediction. Intelligent Data Analysis, 2010, pp 299-331.

[13] K. Strike, K. El Emam, N. Madhavji. Software Cost Estimation with

Incomplete Data. IEEE Transactions on Software Engineering. Volume:

27, Issue: 10, 2001,pp 890 - 908

[14] W. Zhang, Y. Yang, Q. Wang, Q. Handling missing data in software

effort prediction with naive Bayes and EM algorithm. Promise '11 Pro.

7th International Conference on Predictive Models in Software

Engineering, 2011. DOI=10.1145/2020390.202