Available online at www.sciencedirect.com
ScienceDirect
The Journal of Systems and Software 80 (2007) 51–62
The Journal of Systems and Software
www.elsevier.com/locate/jss

A new imputation method for small software project data sets

Qinbao Song [a,*], Martin Shepperd [b]

[a] Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b] Brunel University, Uxbridge, UB8 3PH, UK

Received 17 March 2005; received in revised form 30 April 2006; accepted 3 May 2006
Available online 16 June 2006

Abstract

Effort prediction is a very important issue for software project management. Historical project data sets are frequently used to support such prediction. But missing data are often contained in these data sets and this makes prediction more difficult. One common practice is to ignore the cases with missing data, but this makes the originally small software project database even smaller and can further decrease the accuracy of prediction. The alternative is missing data imputation. There are many imputation methods. Software data sets are frequently characterised by their small size but unfortunately sophisticated imputation methods prefer larger data sets. For this reason we explore using simple methods to impute missing data in small project effort data sets. We propose a class mean imputation (CMI) method based on the *k*-NN hot deck imputation method (MINI) to impute both continuous and nominal missing data in small data sets. We use an incremental approach to increase the variance of population. To evaluate MINI (and *k*-NN and CMI methods as benchmarks) we use data sets with 50 cases and 100 cases sampled from a larger industrial data set with 10%, 15%, 20% and 30% missing data percentages respectively. We also simulate Missing Completely at Random (MCAR) and Missing at Random (MAR) missingness mechanisms. The results suggest that the MINI method outperforms both CMI and the *k*-NN methods. We conclude that this new imputation technique can be used to impute missing values in small data sets.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Software effort prediction; Missing data; Data imputation; Class mean imputation; *k*-NN imputation

1. Introduction

Given that 75% of software projects reported overruns (Moløkken and Jørgensen, 2003), there is considerable demand from industry for accurate software project effort prediction. Unfortunately, a barrier to accurate effort prediction is incomplete and small (in terms of the number of cases) software engineering data sets. Therefore, in order to improve effort prediction, we must first carefully deal with missing data. Although a wide range of missing data techniques have been proposed, none of them specifically focuses upon missing values in small data sets with both nominal and continuous values. For this reason we wish to develop an imputation method specifically to address this problem, that is so typical in software engineering data sets.

There are three types of missing data techniques. Identifying the proper missing data method from these techniques for incomplete small software data sets is the precondition of tackling missing software engineering data. Therefore, in this section, we firstly introduce the taxonomy of missing data techniques, present a big picture of missing data techniques to readers; then briefly summarize the related work in the software engineering field; and lastly raise the research issue of this paper.

1.1. Missing data techniques taxonomy

The missing data problem has been studied by researchers in many fields for more than 30 years. There are three approaches to this problem. First, there are missing data ignoring techniques, e.g. (Haitovsky, 1968; Roth, 1994).

* Corresponding author. Tel.: +86 29 82668645; fax: +86 29 82668971.
E-mail addresses: qbsong@mail.xjtu.edu.cn (Q. Song), martin.shepperd@brunel.ac.uk (M. Shepperd).

0164-1212/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2006.05.003

Second, there are missing data toleration techniques (Aggarwal and Parthasarathy, 2001; Schuurmans and Greiner, 1997). Third, there are missing data imputation techniques which are the emphasis of this paper, e.g. (Friedman, 1998; Little, 1988; Schafer and Olsen, 1998; Shirani et al., 2000; Troyanskaya et al., 2001).

The *missing data ignoring techniques* simply delete the cases that contain missing data. Because of their simplicity, they are widely used (Roth, 1994) and tend to be the default for most statistics packages, but this may not lead to the most efficient utilization of the data and incurs a bias in the data unless the values are missing completely at random. Consequently they should be used only in situations where the amount of missing values is very small. This approach has two forms:

- *Listwise deletion* (LD) is also referred to as case deletion, casewise deletion or complete case analysis. This method omits the cases containing missing values. It is easy, fast, does not 'invent data', commonly accepted and is the default of most statistical packages. The drawback is that its application may lead to a large loss of observations, which may result in too small data sets if the fraction of missing values is high and particularly if the original data set is itself small, as is often the situation for software project estimation (Myrtveit et al., 2001).
- *Pairwise deletion* (PD) is also referred to as the available case method. This method considers each feature separately. For each feature, all recorded values in each observation are considered (Strike et al., 2001) and missing data are ignored. This means that different calculations will utilise different cases and will have different sample sizes, an undesirable effect. The advantage is that the sample size for each individual analysis is generally higher than with complete case analysis. It is necessary when the overall sample size is small or the number of cases with missing data is large.

The *missing data toleration techniques* use a probabilistic approach to handle missing data. They do not predicte missing data but assign a probability to each of the possible values. Thus they are internal missing data treatment strategies, which perform analysis directly using the data set with missing values.

Breiman et al. (1984) proposed the CART algorithm which may be used to address the missing data problem in the context of a decision tree classifier. If some cases contain missing values, CART uses the best surrogate split to assign these cases to branches of a spilt on a feature where these cases' values were missing. C4.5 (Quinlan, 1993) is an alternative method to CART. C4.5 uses a probabilistic approach to handle missing data. Missing values can be present in any variables except the class variable. This method calculates the expected information gain by assuming that the missing value is distributed according to the observed values in the subset of the data at that node of the tree. From amongst the simpler methods, it seems to be one of the better techniques to deal with missing values (Grzymala-Busse and Hu, 2000).

If the objective is not to predict the missing values, missing data toleration is a nice choice. This is because any prediction of missing values will incur bias thereby making prediction results doubtful. However, most data analysis methods only work with a complete data set, so first we must fill in missing values or delete the cases with missing values, and then use the resulting data set to perform subsequent analysis. In this case, toleration techniques cannot be used. Moreover, in cases where the data set contains large amounts of missing data, or the mechanism causing to the missing data is non-random, imputation techniques are likely to perform better than ignoring techniques (Haitovsky, 1968).

The *missing data imputation techniques* estimate missing values for the missing cases and insert estimates obtained from other reported values to produce an estimated complete case. The common forms are as follows:

- *Mean imputation* (MI) is also referred to as unconditional mean imputation. This method imputes each missing value with the mean of reported values. It is fast, simple, easily implemented and no observations are excluded. The disadvantage is that it leads to underestimation of the population variance. It is also a rather naïve approach.
- *Regression imputation* (RI) is also referred to as conditional mean imputation. This method replaces each missing value with a predicted value based on a regression model. The regression model is built using the complete observations. It tends to perform better than MI, but still underestimates variance.
- *Hot-deck imputation* (HDI) methods fill in missing data by taking values from other observations in the same data set. The choice of which value to take depends on the observation containing the missing value. Randomly choosing observed values from donor cases is the simplest hot-deck method. The similar response pattern imputation (SRPI) (Joreskog and Sorbom, 1993), which identifies the most similar case without missing observations and copies the values of this case to fill in the holes in the cases with missing data, and the k nearest neighbours (k-NN) imputation (Fix and Hodges, 1952; Cartwright et al., 2003; Song et al., 2005; Jönsson and Wohlin, 2004), which searches for the k most similar cases to the missing value and replaces the missing value by the mean or modal value of the corresponding feature values of the k nearest neighbours all belong to this class. This approach preserves the sample distribution by substituting different observed values for each missing observation, but the data set must be large enough to find appropriate donor cases.
- *Multiple imputation* means that the missing data are imputed $m > 1$ times, with a different randomly chosen error term added in each imputation. In this method, each missing value is replaced by a set of m plausible

Download English Version:

https://daneshyari.com/en/article/494337

Download Persian Version:

https://daneshyari.com/article/494337

Daneshyari.com