Imputation of Missing Data Using Ensemble Algorithms

Priyadharsini1st
Research scholar
Department of computer science
NGM College, Pollachi
E-mail: priyadharsini.ast@gmail.com

Dr. Antony Selvadoss Thanamani2nd
Associate Professor and Head
Department of computer science
NGM College, Pollachi
E-mail: selvdoss@gmail.com

Abstract: Missing data or incomplete data are very common in statistical situations. One way to deal with missing data is to conduct model imputation either one time or multiple times. One of the key problems in analyzing the imputed dataset is to give the valid statistical reference of the parameter estimated, that is, to give a right estimation of the standard error of the interested statistic. This paper proposes the new developed ensemble algorithms as imputation model. In order to realize multiple imputations, we suggest bootstrap sampling the prediction error several times. The properties of the proposed methods are studied by simulation and compared with existing methods. Finally, the methods are applied to analyze one real large dataset, taking the missing mechanism into consideration.

Keywords — data mining; Ensemble; random forest

I. INTRODUCTION

Missing data or incomplete data are very common in statistical situations [1]. It is a serious problem because they can lead bias and inefficiency in estimating the quantities of interest. Noticing its importance, many researchers have proposed methods to make statistical inferences when data are incomplete. A common way is to discard records with missing values and restrict the attention to those records for which all data are observed. However when the fraction of missing is large or the missing mechanism is not missing completely at random, this method is problematic. Another way to deal with missing data is model imputation, i.e., to fit a model on the observed cases for a given variable treating it as the outcome and then, for cases where this variable is missing, fill in the predicted value by the model to get an imputed value. It is also called single imputation [1].

In recent years, multiple imputation (MI) has emerged as a convenient and flexible paradigm for analyzing data with missing values [3,4]. Researchers use certain statistical methods to fill in missing values with plausible ones M (>1) times, where m is typically small (say, 3-10). Then each of the simulated complete datasets is analyzed by standard methods, and the results are later combined. One of the key problems in analyzing the imputed dataset is to give the valid statistical reference of the parameter estimated, that is, to give a right estimation of the standard error of the interested statistic.

The validity and efficiency of complete-data based methods cannot be guaranteed when data are incomplete because the uncertainties associated with the imputations need to be appropriately addressed since imputed values are not real observed ones [2].

This paper proposes the newly developed ensemble algorithms as imputation models. In order to realize multiple imputation, we suggest bootstrap sampling the prediction error a few times. The rest of this article is organized as follows. Section II gives a brief introduction on ensemble algorithms.

The detail of the proposed imputation method is introduced in Section III. The properties of the proposed method are studied with simulation and compared with existing methods in Section IV. After that the methods are applied to analyze one real large dataset, taking the missing mechanism into consideration in Section V. Finally Section VI concludes the paper.

II. ENSEMBLE ALGORITHM

In context of supervised learning, it is usually assumed that there is some unknown function f(x) that generates the data and our goal is to find an approximate off. Single learning algorithm usually pre-assumes a space of possible functions, namely hypotheses, then it tries to find out the function that matches f(x) best on the training data. However, this kind of algorithms may encounter statistical, computational and representable problems, as mentioned in [5].

To conquer those problems, ensemble algorithms combine a collection of weaker learners to produce a strong learner, or in other words, it combines multiple hypotheses to form a better hypothesis. The resulting hypothesis may not lies in the hypotheses of any of the weaker learners, thus ensemble learning algorithms have more flexibility than single learner. Empirically, ensemble algorithms usually yield better results than single learner if properly used. Bagging, or bootstrap aggregating, is a simple ensemble algorithm raised by Breiman [6].

It fits simple model to a collection of bootstrap samples, then gets the result by averaging the output (for regression) or voting (for classification). If the learning algorithm is unstable — that is, if small changes in the training data lead to large changes in the resulting hypothesis — then bagging will produce a diverse ensemble of hypotheses. Boosting is one of
the most powerful learning ideas introduced in the last twenty years [7].

The motivation for boosting was a procedure that combines the outputs of many weak classifiers to produce a powerful committee. The process is to fit the weaker learners iteratively and reweight the data after each iteration. To go into details, every time it fits the weaker learner to weighted training data and adds it to the current combined learner with some weight related to the weaker learner’s accuracy. Then it reweights the data with the prediction of the renewed learner: cases that are misclassified gain weight and cases that are classified correctly lose weight. Then it goes to next iteration until some conditions are satisfied. After certain times of iteration, we get the boosting model.

A. Random Forest

Random forest is one of the classification methods. It is a method of aggregating the prediction ability of multiple classifiers, known as ensemble classification. Ensemble classification is a machine learning method that aims to achieve better prediction accuracy compared to single models by utilizing the advantages of multiple models [15].

It tends to generate base models that are competent yet complementary resulting in a comprehensive classifier. Many traditional machine learning algorithms generate a single model such as decision tree and neural network.

However, the ensemble learning method generates multiple models. Europides et al. [16] illustrated an experiment to show how an ensemble method can improve a classifier’s performance.

The experiment showed the calculated error rate of 0.06 in the ensemble classifier, which was lower compared to the binary classifier 0.35 in the experiment. Random forest develops many tree base classifiers, in which each tree depends on the values of a random input vector sampled independently from the total data set with replacement and with the same distribution of all trees within the forest [16].

The randomness of the random forest originated from the random selected samples and variables. The random forest approach has shown good accuracy in its overall performance.

Mohammed Khalilia et al. [17] indicated that Random Forest outperformed the support vector machine, bagging and boosting in terms of the area under the receiver operating characteristic (ROC) curve (AUC) to predict disease risks from highly imbalanced data.

However, there is a limitation in the data obtained which may result in the use of the same data for the same patient multiple times that may cause a slight bias in prediction result.

In general, random forest algorithm produces high prediction accuracy. Hence, in this research, this algorithm is applied to another area of health, which is mental health, to determine its suitability and performance.

III. METHODS

WEKA library was used in this implementation. For missing data problem, Reference [8] considers a nonparametric approach which employs tree based models for missing data imputation. Reference [9] proposes to use boosting method for missing data imputation. Here in this article we propose to use ensemble algorithms such as bagging and boosting as the imputation model when the missing variable is continuous. The detailed procedure is as follows.

S1: Using the data without missing values to build the model and get the prediction error.

S2: Applying the model to predict the missing values.

S3: Bootstrap sampling the prediction error
In step 1 and adding them to the predicted values in

Step 2

S4: Repeating Step 3 M time to obtain MI.

S5: Analysing each of the M completed datasets using standard method.

S6: Adjust the results in S5,

taking the uncertainties due to imputation into consideration. There are two main differences between our proposed methods with [8, 9]. First, we only consider one variable with missing values while they consider several ones. Second, we conduct MI by bootstrapping the prediction error while they are single imputation. The imputation method introduced here by itself is simple. The important issue is to give correct statistical inference of the interested parameter using the imputed data. We will study this problem by simulation in next section.

IV. SIMULATION STUDIES

Samples of data were simulated as follows. For each observation, covariates X=(X1,….X5) were generated with each Xi distributed independently as uniform in the interval [-5, 5]. Then Y was generated from the model Y=10sin(πX1X2)+ 20(X3-0.5)^2 +10X3+e, where e~N(0,4). The values of y are missing with conditional probability 1/(1+exp(-1+0.5(X1+X3))).

The population parameters to be estimated are expectation (u) and variance (s) of Y. When there is no missing value, we use sample mean and sample variance as the estimates and give their associated standard errors, 95% confidence intervals. When there are missing values and they are “filled in” by the imputation method, uncertainties due to imputations need to be appropriately addressed. We use the following procedure to study this question.

To study the effect of
(1) Sample size,
(2) Fraction of missing and
(3) Times of multiple imputation,

i.e., the value of M, we design a 3 by 3 by 6 factorial experiment with n equals 1000, 2000, 5000, fraction of missing equals 0.2, 0.3 and 0.5 by setting different cutoff values for the conditional

Probabilities; M equals 2, 5, 10, 20, 50, 10000. For comparison purpose, we also consider regression model and KNN method as single imputation methods without bootstrapping the prediction errors. Table 1 shows the results of n=1000, M=5 and percent of missing is 30%. Results for other settings are similar and available from the authors.
Imputation are substituting errors. In future, we may use other binary classification method to model missing mechanism. Fourth, in real data analysis, we can use other binary classification method to model missing mechanism. Third, in real data analysis, we can consider more data generating scenarios. We also suggest bootstrap sampling the prediction error in order to conduct multiple imputation. The properties of the proposed methods are compared with existing ones by simulation studies. Then they are applied to analyze a real large dataset to obtain useful results. Lots of work can be done in future. For example, first, the prediction errors we use for bootstrap sampling to generate multiple imputation are substituted errors. In future, we may consider split the non-missing data into training set for building imputation model and test set for generating prediction error to bootstrapping. Second, in simulation studies, we can consider more data generating scenarios. Third, in real data analysis, we can use other binary classification method to model missing mechanism. Fourth, in both simulation and real analyses, missing values can be obtained by generating uniform random variable and comparing its value with the conditional missing probability.

<table>
<thead>
<tr>
<th>TRUE</th>
<th>Reg</th>
<th>KNN</th>
<th>boost</th>
<th>Bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>171.678</td>
<td>137.498</td>
<td>144.400</td>
<td>170.941</td>
<td>170.8</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>221</td>
</tr>
<tr>
<td>23.414</td>
<td>44.8574</td>
<td>24.722</td>
<td>23.5</td>
<td>24.20</td>
</tr>
<tr>
<td>162.477</td>
<td>124.455</td>
<td>134.82</td>
<td>161.444</td>
<td>161.4</td>
</tr>
<tr>
<td>181.009</td>
<td>150.49</td>
<td>154.021</td>
<td>180.3</td>
<td>180.7</td>
</tr>
<tr>
<td>23115.2</td>
<td>21441.6</td>
<td>21600.0</td>
<td>22490.2</td>
<td>2239</td>
</tr>
<tr>
<td>704922</td>
<td>738568</td>
<td>7716</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2113470</td>
<td>1034084</td>
<td>4</td>
<td>93.2</td>
</tr>
<tr>
<td>V(s)</td>
<td>V(u)</td>
<td>S</td>
<td>UL</td>
<td>U</td>
</tr>
</tbody>
</table>

The true values in Table 1 were obtained by 2000 times of simulation. On each trial, a sample of size n=1000 was generated. Sample mean and sample variance for variable y were calculated and denoted as \( y(i) \), \( i=1,...,2000 \). The mean, variance, 2.5% quantile and 97.5% quantile of these 2000 values of \( y(i) \) are treated as true values. For imputation with regression method, again, we performed 2000 times of simulation. On each trial, after generating 1000 data points, 30% of y were deleted with the highest conditional probability of missing. After imputation, we calculate the sample mean and sample variance for variable y with completed data and denoted as \( y_{\text{Reg}}(i) \) and \( y_{\text{Reg}}(i) \), \( i=1,...,2000 \). The mean, variance, 2.5% quantile and 97.5% quantile of these 2000 values of \( y_{\text{Reg}}(i) \) and \( y_{\text{Reg}}(i) \) are the numbers shown in the column Reg of Table 1.

**Figure 1: Density estimates of sample variance for true values and different imputed methods when fraction of missing equals 0.3.**

From Table 1, for using sampling mean to estimate the population expectation when missing values are implemented by different methods, boosting and bagging are almost unbiased estimation, while regression and KNN methods highly underestimate the true value with 95% confidence interval not covering the true value. For using sampling variance to estimate the population variance when missing values are implemented by different methods, regression and KNN methods again highly underestimate the true values. Boosting and bagging also underestimate the true value, but in a lower degree. Figure 1 shows the density estimates of sample variance for the true value obtained from the four imputation methods.

**V. REAL DATA ANALYSES**

S1: applying logistic regression to model the missing mechanism. That is, the dependent variable is missing or not and the independent variables are contrast’s type, technical area, intellectual property, payment method, application target, contract’s value. All variables are significant and the overall prediction accuracy is 69.3%. This shows that missingness depends on covariates.

S2.1: bootstrapping sampling 1000 times and calculating bootstrap sample mean of execution amount along with its variance and confidence interval. These values are treated as true ones.

S2.2: applying the logistic regression in step 1 to obtain conditional missing probability for each case and deleting 51.3% of cases with highest missing probabilities.

S2.3: filling in the missing values with each imputation method (boosting and bagging), bootstrap sampling the completed data 1000 times and calculating bootstrap sample mean along with its variance and confidence interval.

**VI. CONCLUSION**

In this article, propose ensemble algorithm such as bagging and boosting as imputation methods for missing data. We also suggest bootstrap sampling the prediction error in order to conduct multiple imputation. The properties of the proposed methods are compared with existing ones by simulation studies. Then they are applied to analyze a real large dataset to obtain useful results. Lots of work can be done in future. For example, first, the prediction errors we use for bootstrapping to generate multiple imputation are substituted errors. In future, we may consider split the non-missing data into training set for building imputation model and test set for generating prediction error to bootstrapping. Second, in simulation studies, we can consider more data generating scenarios. Third, in real data analysis, we can use other binary classification method to model missing mechanism. Fourth, in both simulation and real analyses, missing values can be obtained by generating uniform random variable and comparing its value with the conditional missing probability.

**REFERENCES**


2. Laura Lazzeroni, Nathaniel Schenker, and Jeremy Taylor, Robustness of Multiple-Imputation Techniques to Model Misspecification, 260-265.


