Evaluation of Imputation Methods for Missing Data and Their Effect on the Reliability of Predictive Models

Xiao-Ou Ping, Feipei Lai, Yi-Ju Tseng
Graduate Institute of Biomedical Electronics and Bioinformatics, Department of Computer Science and Information Engineering
Dept. of Electrical Eng.
National Taiwan University
Taipei, Taiwan
pingxiaoou@gmail.com

Ja-Der Liang, Guan-Tarn Huang, Pei-Ming Yang
Department of Internal Medicine
National Taiwan University Hospital and National Taiwan University College of Medicine
Taipei, Taiwan
jdliang@ntuh.gov.tw

***Abstract* — In medical research, the problem of missing data occurs frequently. In this paper, eight imputation methods are evaluated based on accuracy and stability through a simulation experiment. The objective of this paper is to find appropriate methods for handling incomplete data sets during the development of predictive models which predict the recurrence status of liver cancer patients. Support vector machine (SVM) is employed for building predictive models. The data sources produced by different missing data handling methods (complete variable analysis and imputation method) are used for evaluating the impact on the development of the recurrence predictive model. Imputation methods show the potential benefit of features with missing values during the development of the recurrence predictive model.**

Keywords - incomplete data; missing value; predictive model; liver cancer

I. INTRODUCTION

According to a study reviewing 100 articles among seven cancer journals, up to 81 articles have evidence of missing data [1]. The problem of missing data occurs frequently. Therefore, how to handle incomplete data set is a crucial issue during data analysis. To handle incomplete data sets, several general handling methods are proposed [2]: (1) complete variable analysis: dropping the variables with missing data and analyzing only the variables without missing data, and (2) imputation method: estimating the missing values based on different methods. In the study, the complete variable data set, and the data sets imputed by different imputation methods are both employed for evaluating the impact of missing value handling methods for developing the predictive model. Performances of predictive models built based on these two types of data sets are compared for checking whether if the features with missing data have potential benefit for building the predictive model.

To estimate missing values in data sets, eight imputation methods are employed in this study and we design a simulation experiment for comparing the imputation performance on the stability and accuracy. Of eight imputation methods, six are single imputation (i.e., single imputed value for each missing value) and two are multiple imputations (i.e., multiple imputed values for each missing value). In this work, normalized root mean squared error (NRMSE) [3][4] is used for evaluating the accuracy of imputation methods; furthermore, the stability of imputation methods is also evaluated through repeated simulation experiments.

According to the global cancer statistics in 2011, liver cancer in men is the second most frequent cause of cancer death and in women, it is the sixth leading cause of cancer death. Hepatocellular carcinoma (HCC), as the most common primary liver cancer [5], has been the leading cause of cancer death in Taiwan.

For patients with early-stage HCC who are not suitable for surgical resection or liver transplantation, radiofrequency ablation (RFA) is the best alternative treatment [6]. In previous studies, researchers estimated that the cumulative 5 year recurrence rate is more than 70% for patients who received RFA [7]. The recurrence predictive models play a crucial role for physicians and patients in enabling the opportunities of supporting early prediction of a recurrence status.

In the work, support vector machines (SVM) [8][9] is employed as a classifier for developing the recurrence predictive model for newly diagnosed HCC patients receiving RFA treatment in one year. The predictive models built based on the data sources produced by different missing data handling methods (i.e., complete variable analysis and imputation method) are further evaluated and compared for presenting the impact of these methods. In the past few years, SVMs have been widely employed in medical specialties such as breast cancer [10] and liver diseases (fatty liver [11] and liver fibrosis [12]).

This study introduces a two-level approach to evaluating imputation methods and predictive models when the problem of missing data occurs. The performance of an imputation

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

method (i.e., accuracy and stability) may have variance according to its parameter settings and different data sets (e.g., different data sets from patients with different diseases). In the first level, an evaluation of imputation methods assists researchers in selecting appropriate imputation methods and their parameter settings for a specific data set through a designed simulation experiment. Furthermore, this simulation experiment can further provide information for evaluating reliability of predictive models which are developed based on imputed data sets (i.e., missing values are imputed by an imputation method). In the second level, the performance of an imputation method is further evaluated based on each clinical feature with missing values. When predictive models employ clinical features with estimated values (estimating by an imputation method), the performance of an imputation method for these features can be regarded as factors in evaluating reliability of these predictive models. When a predictive model relies heavily on a specific clinical feature and the performance of an imputation method for this feature is not good, the predictive model may be not a model with good reliability. This study focuses on a case study which is to find appropriate methods for handling incomplete data sets during the development of predictive models which predict the recurrence status of liver cancer patients.

The construction of this paper is organized as follows. In Section II, an overview of our method is presented. Section III introduces the material about a specific data set used in this study. Section IV describes a designed simulation experiment, and introductions and evaluations of imputation methods. Section V contains a method of building predictive models and evaluations of these models. Sections VI and VII present results of imputation methods and predictive models. Section VIII discusses results and limitations of this study. Finally, conclusion and future work are given in Sections IX and X.

II. METHOD OVERVIEW

A two-level approach is introduced to evaluating imputation methods and predictive models when the problem of missing data occurs. Fig. 1 shows the overview of this work. A simulation experiment is designed for evaluating performance of imputation methods and reliability of a predictive model.

Before performing the simulation experiment, cases that have missing values (MVs) are removed from original data sets for producing complete cases data set without MVs. In each simulation round, partial original data (ten percent data) in the complete cases data set are masked as missing values and then these missing values are imputed by imputation methods. After the process of data imputation, the original values included in the complete cases data set and the imputed valued estimated by imputation methods can be compared for evaluating performances of these imputation methods based on NRMSEs.

Two major data sources are used for developing the predictive model: (1) complete variables data set: dropping the variables (features) with missing data and using only the variables without missing data, and (2) imputed data sets: imputing the missing values included in the original data set and using the imputed data set. The predictive models are evaluated based on criteria such as sensitivities and specificities.

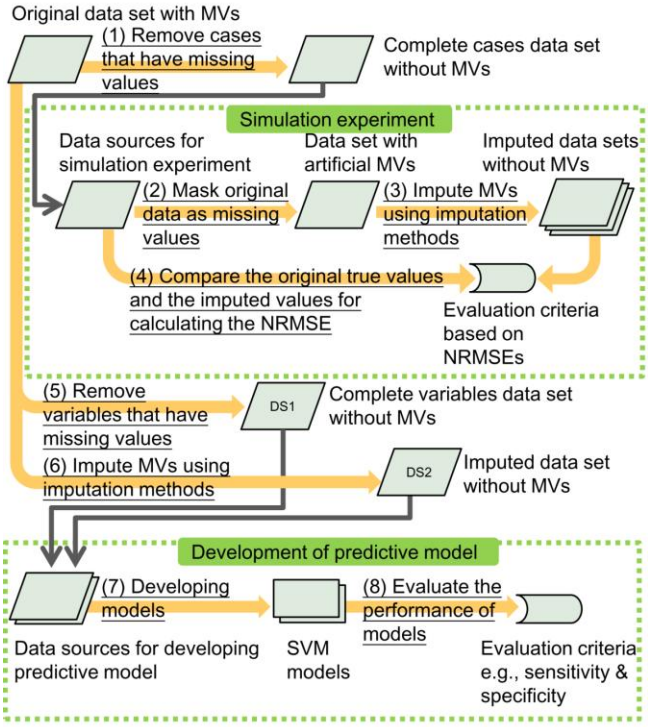

Figure 1. Simulation experiments and development of a predictive model based on two different data sets (DSs).

III. MATERIAL

83 HCC patients received ultrasound guided RFA were included in this study. RFA is their first treatment for HCC in NTUH between 2007 and 2009. Of the 83 patients, 18 patients had recurrent HCCs in one year after the RFA treatment and 65 patients were not recurrent in one year. A total of 20 clinical features included in this study are as follows: age, gender, tumor number, the size of the maximal tumor, liver cirrhosis, Barcelona Clinic Liver Cancer (BCLC) staging classifications [13], and 14 serum laboratory tests, including prothrombin time (INR), albumin, aspartate aminotransferase (AST), alanine transaminase (ALT), total bilirubin, creatinine, platelet count, alpha-fetoprotein (AFP), HBsAg, anti-HCV, alkaline phosphatase (ALP), direct bilirubin, total protein, and Gamma-glutamyl transpeptidase (GGT). In each feature, one value that is before and closest to the RFA treatment is used. The last four features include missing values. The missing rates of these features are as follows. ALP has 8.43% missing values, direct bilirubin has 15.66%, total protein has 22.89%, and GGT has 27.71%. The complete data set is produced by dropping these four features with missing values. The imputed data sets are produced by estimating missing values of these four features based on eight imputation methods and different parameter settings.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

IV. DATA IMPUTATION

A. Simulation experiment

There are two types of imputation methods, including single imputation (i.e., single imputed value for each missing value) and multiple imputation (i.e., multiple imputed values for each missing value). This study employed both types of imputation methods for comparing their influence on this specific data set in our study.

For applying the single imputation methods, the "pcaMethods" is employed. The "pcaMethods" is a Bioconductor package and proposed by Stacklies et al. [14]. In the package, they implement a collection of principal component analysis (PCA) based methods and a non-PCA based method for estimating the missing values of the incomplete data. This package contains six single imputation methods, including singular value decomposition based imputation method (SVDImpute), local least squares imputation method (LLSImpute), probabilistic PCA (PPCA), Bayesian PCA (BPCA), non-linear PCA (NLPCA), and non-linear estimation by iterative partial least squares PCA (Nipals PCA). For applying the multiple imputation methods, the multivariate imputation by chained equations (MICE) [15] and multiple imputation (MI) [16] packages in R are employed. In this work, these methods included in these packages are employed for estimating missing data of the clinical features (e.g., serum laboratory tests).

For each feature with missing values, the cases with missing values are removed and simulation experiment is performed based on the complete data set (Fig. 2). The data set is separated into ten parts randomly. To reduce the bias due to just one simulation, the simulation experiment is repeated ten times and each case has one chance to be masked as missing values. The concept is similar to ten-fold cross validation. In each round of simulation experiment, each 10% data of this feature are masked as artificial missing values and eight imputation and different parameter settings are used for estimating the missing values. Then the true original values and the estimated values are compared for evaluating the performance of imputation method.

For each imputation method and its corresponding parameter setting, the experiment is repeated 10 times and each case has one chance to be masked as missing values. The distribution of these 10 NRMSEs is analyzed rather than calculating one mean value of these NRMSEs. The average of first quartile, third quartile, and the median of these 10 NRMSEs is regarded as the imputation method selection criterion for comparing the imputation performance in terms of stability and accuracy. A low NRMSE score means few imputation errors and high accuracy. Low imputation method selection criterion means high stability and accuracy.

Eight imputation methods and their corresponding parameter settings (total 37 combinations) are employed for estimating missing values of each feature. For each feature, what combinations achieve top 10 leading imputation performances are analyzed (i.e., top 10 scores of imputation method selection criterion). For example, the SVDImpute have five combinations (i.e., five parameter settings). If its four combinations achieve top 10 leading imputation performances, then its rate of combinations achieving top 10 leading imputation performances is 80%. After simulations of all features are done. The overall rate of combinations achieving top 10 leading imputation performances for all feature can be calculated by averaging the rate of combinations achieving top 10 leading imputation performances of each feature.

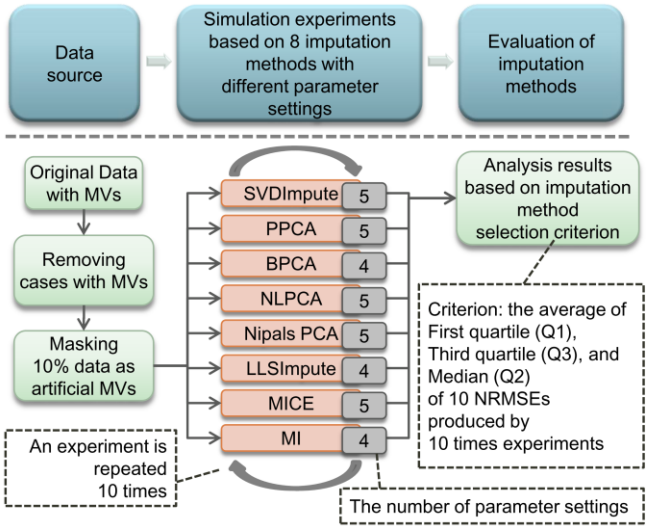

Figure 2. The simulation experiment based on data set without missing values using eight imputation methods.

B. Imputation methods

In SVDImpute, singular value decomposition (SVD) is used for obtaining a set of mutually orthogonal expression patterns (e.g., eigengenes in their study) [17]. These patterns can be used to approximate the expression of all features in data sets based on the linear combination of these patterns. PCA is popular approach for data analysis and data processing (e.g., dimension reduction). PCA is not based on a probability model and PPCA [18] includes an expectation–maximization (EM) approach for PCA with a probabilistic model [14]. BPCA is based on three processes, including principal component (PC) regression, Bayesian estimation, and an EM-like repetitive algorithm [19]. NLPCA is regarded as a non-linear generalization of standard linear PCA [20]. Nipals PCA [21] is a method at the root of PLS regression [14]. The parameter settings of SVDImpute, PPCA, BPCA, NLPCA, and Nipals PCA all include the number of principal components.

LLSimpute [22] estimates missing values based on a linear combination of k selected similar variables. The k variables are selected by the Euclidean distance or by Pearson correlation coefficients. The optimal combination is found by local least squares (LLS) regression [14]. The parameter setting of LLSImpute is the number of variables selected for regression.

MICE is used for generating multiple imputations [15] and it contains different imputation functions, including the

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

predictive mean matching (pmm), Bayesian linear regression (norm), linear regression ignoring model error (norm.nob), unconditional mean imputation (mean), and random sample from the observed values (sample). The parameter setting of MICE is based on the number of iterations and different imputation functions used for multiple imputations.

Multiple imputation (MI) is used for generating multiple imputations based on iterative regression imputation [16] and it contains different models for different variable types. For example, the "binary" regression model is used for binary data, and the "categorical" regression model is used for unordered categorical data. The parameter setting of MI is based on the number of iterations and the functions used for adding noise in multiple imputation procedure (e.g., reshuffling and fading).

In multiple imputation methods, MICE and MI, the number of imputed data sets is set as 2 in this study and two imputed data sets are generated. Both of these data sets are used for developing predictive models and a model with better performance is selected.

C. Evaluation of imputation methods

NRMSE is frequently applied for evaluating the performance of imputation methods. The root mean squared error (RMSE) is used for calculating the error between the estimated values of the missing entries and original true values in the complete data set. The RMSE is further normalized by the following constant: the range of the original true values over the missing entries [3][4].

$$NRMSE = \frac{\sqrt{mean\ [(y_{estimated} - y_{true})^2]}}{\max(y_{true}) - \min(y_{true})} \quad (1)$$

The "*max()*" function denotes the maximum value of a listing numbers. The "*min()*" function denotes the minimum value of a listing numbers. The $y_{estimated}$ means the estimated values over the missing entries and the estimated values are imputed by different imputation methods. The y_{true} means the original true values of the missing entries and the true values are from the original complete data set.

V. Recurrence Predictive Model

In the work, the SVM is employed as a classifier for the prediction of the recurrence status of the patients with HCC after RFA treatment in one year.

A. SVM for classification

The SVM was proposed by Boser, Cortes, and Vapnik and it is widely used for solving classification problem [8][9]. The mapping function is used for mapping input feature vectors into higher dimensional feature space for SVM to find the linear separating hyperplane for separating the instances into two classes. In the work, radial basis function (RBF), is selected as the kernel function and the cost parameter *C* and parameter of kernel function gamma are the parameters that can be adjusted during the development of the SVM classification model. In this work, we perform SVM based on LIBSVM [23]. The grid search is employed for searching the appropriate parameters, *C* and gamma, of SVM models. In the work, the grid search based on 5-fold cross-validation (inner loop) is adopted for finding appropriate parameters [23][24]. The sensitivity is regarded as the criterion for finding parameters in grid search to achieve better sensitivity in our data set.

B. Feature Selection

In the work, the hybrid feature selection method was employed. We combine simulated annealing (SA) [25] with random forests [26]. First, SA was employed for selecting a subset of features from all features. Random forests (RF) was employed for assigning the weight of importance for the feature in the selected subset. The selected features were added stepwise as the input data to train our SVM model. SA was developed from the idea of annealing of metallurgy. The raw material can be heated for growing a crystal. The temperature is reduced until the crystal structure is frozen and the better results can be achieved through slower process of cooling [27]. In this study, the "better result" denotes the better subset of features. RF creates many classification trees. The importance of a feature is determined by permuting this feature and all others were reserved and then calculating the increased amount of prediction error [26]. In the work, SA is performed using R package, named "Subselect" [28] and RF was performed using R package, named "FSelector" [29].

Double five-fold cross-validation loop method is adopted. An inner loop five-fold cross-validation is performed on the training data set of an outer loop for finding appropriate parameter settings of SVM model. Then, the selected parameters are used for training the whole training data set of an outer loop and getting a trained SVM model and a training classification result. The selected features were added stepwise to train different SVM models. This training classification result is further used as criteria for selecting appropriate feature combinations. Finally, the average classification results of outer loop cross-validation are presented.

C. Evaluation of the recurrence predictive model

The sensitivity, specificity, accuracy, balanced accuracy (BAC), positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic (ROC) curve (AUC) are used as evaluated metrics for evaluating the predictive models. The definitions are as follows:

$$Sensitivity = TP / (TP + FN) \quad (2)$$

$$Specificity = TN / (TN + FP) \quad (3)$$

$$Accuracy = (TP + TN) / (TP + FP + FN + TN) \quad (4)$$

$$BAC = ((Sensitivity + Specificity) / 2) \quad (5)$$

$$PPV = TP / (TP + FP) \quad (6)$$

$$NPV = TN / (TN + FN) \quad (7)$$

TP (True Positive) is the patient predicted with recurrent HCC and the patient actually has recurrent HCC. TN (True Negative) is the patient predicted without recurrent HCC and the patient is actually without recurrent HCC. FP (False Positive) is the patient predicted with recurrent HCC but the patient is actually without recurrent HCC. FN (False

Negative) is the patient predicted without recurrent HCC but the patient actually has recurrent HCC. In this study, the ROC is created based on decision values of the SVM [30].

VI. PERFORMANCE OF SIMULATION EXPERIMENTS

The rate of imputation methods which has top 10 leading performances with different parameter settings for four features with missing values are shown in Fig. 3. SVDImpute and Nipals PCA perform well (in top 10 leading performances) in 80% of cases (different parameter settings and different features). The PPCA performs well in 55% of cases and the BPCA performs well in 44% of cases. Other imputation methods perform well under 30% of cases, especially the LLSImpute only performs well in 6% of cases.

For each imputation method, a specific parameter setting which has the best rate of top 10 leading performances for four features with missing values are selected and further analyzed, including ALP, direct bilirubin (DB), total protein (TP), and GGT. In TABLE I, the SVDImpute with the parameter value of 15, performs well (in top 10 leading performances) in all four features with missing values. The Nipals PCA with the parameter value of 15, also performs well in all four features. The PPCA, BPCA, and MICE can find parameters to perform well in three features (ALP, total protein, and GGT). The NLPCA can find parameters to perform well in two features (Total protein and GGT). The LLSImpute and MI methods can only find parameters performing well in a single feature (The LLSImpute can perform well in GGT, and MI can perform well in direct bilirubin).

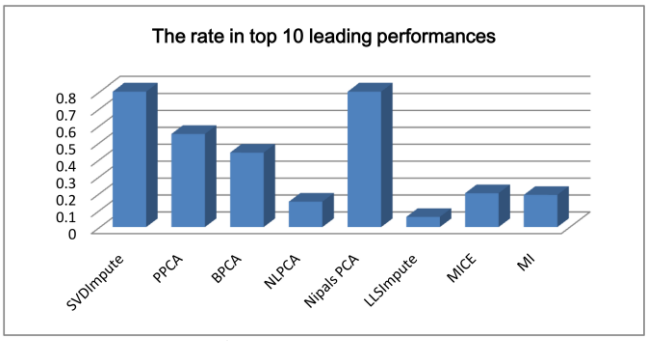

Figure 3. The rate of imputation methods which has top 10 leading performances with different parameter settings for four features with missing values.

TABLE I. THE PARAMETER SETTINGS WITH THE BEST RATE OF LEADING 10 PERFORMANCES FOR FOUR FEATURES WITH MISSING VALUES.

Method	Parameter	ALP	DB	TP	GGT
SVDImpute	15	V	V	V	V
PPCA	20	V		V	V
BPCA	1	V		V	V
NLPCA	1			V	V
Nipals PCA	15	V	V	V	V
LLSImpute	5				V
MICE	mean, 20	V		V	V
MI	fading, 100		V		

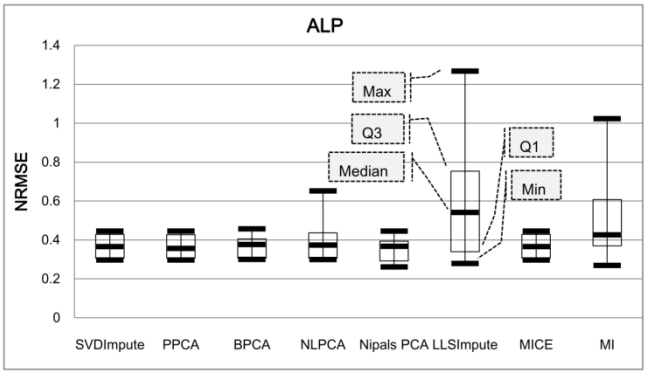

Figure 4. The experiment results based on eight imputation methods for ALP.

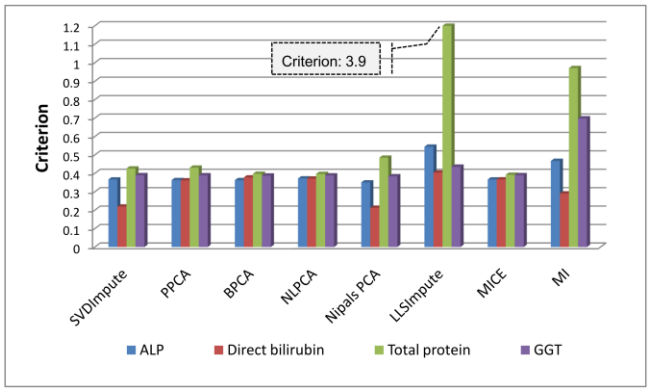

Figure 5. The imputation method selection criterion (averaging Q1, median and Q3) of eight imputation methods for four features.

The further information relevant to eight imputation methods with the parameter settings which have the best rate of top 10 leading performances are shown in Fig. 4 (using ALP as an example). The figure shows the maximum, first quartile (Q1), third quartile (Q3), median, and minimum of 10 NRMSEs in 10 repeated simulation experiments. In Fig. 4, the maximum, and the median in NLPCA, LLSImpute, and MI are larger than those of other imputation methods. Fig. 5 shows the performance of eight imputation methods for four features with missing values. For example, for total protein, the LLSImpute and MI did not perform better than the other six imputation methods.

VII. PERFORMANCE OF A PREDICTIVE MODEL

The performance of predictive models produced using complete data set and imputed data sets are shown in TABLE II. The average of performance in five-fold cross-validation is presented. The complete data set contains 16 clinical features with no missing values. The imputed data set contains 16 features as complete data set and other four features with missing values. These missing values are imputed using eight imputation methods. Therefore, there are eight different data sets named by the imputation methods.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

Fig. 6 shows the used frequencies of four features with missing values which selected for building SVM models in five-fold cross-validation.

TABLE II. PERFORMANCE OF PREDICTIVE MODELS PRODUCED BY DIFFERENT DATA SETS

Data Set	Sen	Spe	BAC	Acc	PPV	NPV	AUC
Complete	66.67	85.64	76.16	81.91	68.57	90.12	69.06
SVDImpute	71.67	82.66	77.16	80.59	63.50	91.50	76.00
PPCA	73.33	79.93	76.63	78.31	52.48	91.50	75.14
BPCA	60.00	85.64	72.82	80.66	68.57	88.88	67.90
NLPCA	71.67	79.73	75.70	78.16	53.81	91.67	73.57
Nipals PCA	88.33	71.78	80.06	75.66	49.60	95.96	80.43
LLSImpute	78.33	78.06	78.20	78.16	52.48	92.87	72.70
MICE	66.67	86.23	76.45	81.98	65.33	90.84	66.52
MI	73.33	84.67	79.01	81.99	63.57	91.96	80.28

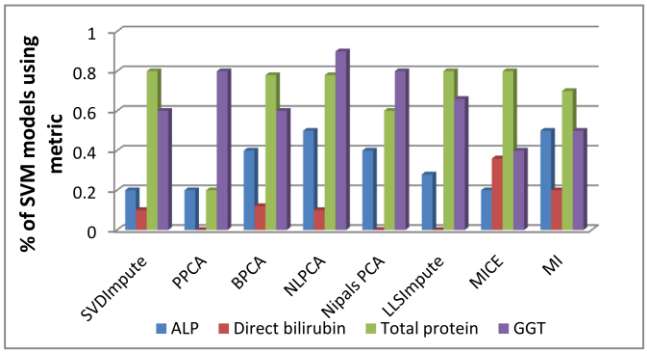

Figure 6. The used frequency of four features for predictive models.

VIII. DISCUSSION

In a study relevant to missing values proposed by Janssen et al., they apply logistic regression to modeling the risk of deep venous thrombosis (DVT). They conduct simple methods for dealing with missing data which will lead to misleading results [2]. Therefore, before building predictive model, the simulation experiments are performed firstly for evaluating the imputation methods in accuracy and stability of estimating missing values in our data sets. Through this simulation, the imputation methods with their parameter settings which are suitable for our data sets can be selected.

The performance of the predictive model with the complete data set is regarded as a reference. Of eight models imputing using different imputation methods, six of them can achieve higher sensitivity than that of complete data set. Especially, the model with Nipals PCA increases about 20% in sensitivity (comparing to the model of complete data set). Although their PPVs are lower than that of complete data set, the model with higher sensitivity can identify more patients with recurrent status. According to the results, the imputation methods reveal the potential of features with missing values in improving sensitivity.

The model with MICE has similar performance with complete data set. In our data set, MICE may not be a suitable method to impute missing values for developing predictive models. The model with BPCA cannot achieve better performance than that of the complete data set. In our data set, the BPCA may not be a suitable method to impute missing values for developing predictive model.

In Fig. 5, the LLSImpute has high imputation selection criterion (not accurate and not stable) in total protein (3.9). MI has high imputation selection criterion in total protein (0.97) and GGT (0.70). However, in Fig. 6, the predictive model with LLSImpute relies heavily on total protein (which appears about four times in five-fold cross-validation). The predictive model with MI relies on total protein and GGT (which appear about 3.5 times and 2.5 times in five-fold cross-validation). Because of above reasons, the reliability of the models with the LLSImpute and MI may not be as good as the models with other imputation methods in our data set. Several limitations of this work are listed in the following content.

Most previous studies concerning patients' recurrence statuses after RFA were focused on risk factors analysis, but not development of predictive models. For example, among these four studies concerning risk factors [7][31][32][33], sample sizes are 118, 124, 190, and 273, respectively. In these study, patients received RFA within specific ranges from four years to five years (e.g., within four years between 2003 and 2007). In our study, 83 patients are collected and they received RFA within a specific range (i.e., within two years between 2007 and 2009). Our sample size is smaller than theirs.

The relationship between patients with missing values and patients without missing values are not further analyzed and discussed in this study. However, we hope these predictive models can also predict patients' statuses when they have missing values. In this study, two ways for handling missing values, complete variable analysis and imputation method, do not remove the patients with missing values. For complete variable analysis, only the features (i.e., variables) with missing values are removed and the number of patients is still 83. For imputation method, features with missing values are reserved and the number of patients is also 83. These missing values are estimated before classification is performed. Therefore, predictive models developed based on these ways can also be used when patients have missing values.

Different feature selection methods may select different feature sets from the same data set which is handled by a specific imputation method. The selection of features may affect predictive models and classification results. In this study, predictive models which built by selected features have better performance than predictive models which built by all features (i.e., not performing this feature selection method). In this study, relationship between imputation methods and feature selection methods is not further discussed.

IX. CONCLUSION

Before imputation methods are employed for estimating missing values during data analysis (e.g., classification), the performance (i.e., accuracy and stability) of imputation

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

methods for a specific data set can be evaluated through the first level evaluation and suitable imputation methods for this data set can be selected. In the second level evaluation, we can not only evaluate the correction of predictive models, but also the reliability of these models. A two-level evaluation method proposed in this study may be applied to other data sets and other predictive targets for finding appropriate imputation methods and providing information of evaluating the reliability of predictive models.

X. FUTURE WORK

In the data set of this study, a clinical feature only has single value. Actually, a clinical feature may have various values which are measured at different time points. Data analysis based on data sets with multiple measurements would be one of our future works. Because SVM can work on higher dimensional feature space, we select this algorithm for finding solutions in different feature spaces. Maybe the comparisons between different algorithms based on this data set can be regarded as one of our future works.

ACKNOWLEDGMENT

The authors acknowledge the members of the research team lead by F Lai of National Taiwan University for their contributions toward this study. F Lai received grants from the National Science Council, Taiwan (NSC 101-2221-E-002-203-MY3).

REFERENCES

[1] A. Burton and D. G. Altman, "Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines," British Journal of Cancer, vol. 91, Jul 5 2004, pp. 4-8.

[2] K. J. M. Janssen, et al., "Missing covariate data in medical research: To impute is better than to ignore," Journal of Clinical Epidemiology, vol. 63, Jul 2010, pp. 721-727.

[3] S. Delepoulle, F. Rousselle, C. Renaud, and P. Preux, "A comparison of two machine learning approaches for Photometric Solids Compression," in Intelligent Computer Graphics 2010, ed: Springer, 2010, pp. 145-164.

[4] R. Jagannathan and S. Petrovic, "Dealing with missing values in a clinical case-based reasoning system," in Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on, 2009, pp. 120-124.

[5] A. Jemal, et al., "Global cancer statistics," CA Cancer J Clin, vol. 61, Mar-Apr 2011, pp. 69-90.

[6] H. B. El-Serag, "Hepatocellular carcinoma," N Engl J Med, vol. 365, Sep 22 2011, pp. 1118-1127.

[7] W. Y. Kao, et al., "Risk factors for long-term prognosis in hepatocellular carcinoma after radiofrequency ablation therapy: the clinical implication of aspartate aminotransferase-platelet ratio index," European Journal of Gastroenterology & Hepatology, vol. 23, Jun 2011, pp. 528-536.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144-152.

[9] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, 1995, pp. 273-297.

[10] X. A. Zhao, et al., "A support vector machine (SVM) for predicting preferred treatment position in radiotherapy of patients with breast cancer," Medical Physics, vol. 37, Oct 2010, pp. 5341-5350.

[11] G. Li, et al., "Computer aided diagnosis of fatty liver ultrasonic images based on support vector machine," Conf Proc IEEE Eng Med Biol Soc, vol. 2008, 2008, pp. 4768-4771.

[12] Y. Sela, et al., "fMRI-based hierarchical SVM model for the classification and grading of liver fibrosis," Biomedical Engineering, IEEE Transactions on, vol. 58, 2011, pp. 2574-2581.

[13] J. M. Llovet, C. Bru, and J. Bruix, "Prognosis of hepatocellular carcinoma: the BCLC staging classification," Semin Liver Dis, vol. 19, 1999, pp. 329-338.

[14] J. Selbig, W. Stacklies, H. Redestig, M. Scholz, and D. Walther, "pcaMethods - a bioconductor package providing PCA methods for incomplete data," Bioinformatics, vol. 23, May 1 2007, pp. 1164-1167.

[15] S. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, 2011, pp. 1-67.

[16] Y.-S. Su, M. Yajima, A. E. Gelman, and J. Hill, "Multiple imputation with diagnostics (mi) in R: opening windows into the black box," Journal of Statistical Software, vol. 45, 2011, pp. 1-31.

[17] O. Troyanskaya, et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, Jun 2001, pp. 520-525.

[18] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," Neural Computation, vol. 11, Feb 15 1999, pp. 443-482.

[19] S. Oba, et al., "A Bayesian missing value estimation method for gene expression profile data," Bioinformatics, vol. 19, Nov 1 2003, pp. 2088-2096.

[20] M. Scholz, F. Kaplan, C. L. Guy, J. Kopka, and J. Selbig, "Non-linear PCA: a missing data approach," Bioinformatics, vol. 21, Oct 15 2005, pp. 3887-3895.

[21] H. Wold, "Estimation of principal components and related models by iterative least squares," Multivariate analysis, vol. 1, 1966, pp. 391-420.

[22] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," Bioinformatics, vol. 21, Jan 15 2005, pp. 187-198.

[23] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, 2011, pp. 27:21-27:27.

[24] C. W. Hsu, C. C. Chang, and C. J. Lin, "A Practical Guide to Support Vector Classification," 2010.

[25] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," Science, vol. 220, 1983, pp. 671-680.

[26] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, 2002, pp. 18–22.

[27] Z. Michalewicz, M. Schmidt, M. Michalewicz, and C. Chiriac, Adaptive Business Intelligence: Springer, 2007.

[28] J. Cadima, J. O. Cerdeira, P. D. Silva, and M. Minhoto, "The subselect R package, Version 0.11," 2011.

[29] P. Romanski, "Selecting attributes**,** Package 'FSelector'," 2012.

[30] E. R. Delong, D. M. Delong, and D. I. Clarkepearson, "Comparing the Areas under 2 or More Correlated Receiver Operating Characteristic Curves - a Nonparametric Approach," Biometrics, vol. 44, Sep 1988, pp. 837-845.

[31] V. W. T. Lam, et al., "Risk factors and prognostic factors of local recurrence after radiofrequency ablation of hepatocellular carcinoma," Journal of the American College of Surgeons, vol. 207, Jul 2008, pp. 20-29.

[32] K. Shiozawa, et al., "Risk factors for the local recurrence of hepatocellular carcinoma after single-session percutaneous radiofrequency ablation with a single electrode insertion," Molecular Medicine Reports, vol. 2, Jan-Feb 2009, pp. 89-95.

[33] B. W. Yang, et al., "Risk factors for recurrence of small hepatocellular carcinoma after long-term follow-up of percutaneous radiofrequency ablation," European Journal of Radiology, vol. 79, Aug 2011, pp. 196-200.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3