Expert Systems

Article

DOI: 10.1111/exsy.12155

Data preprocessing issues for incomplete medical datasets

Min-Wei Huang,[1] Wei-Chao Lin,[2] Chih-Wen Chen,[3,4]*
Shih-Wen Ke,[5] Chih-Fong Tsai[6] and William Eberle[7]

(1) Department of Psychiatry, Chiayi Branch, Taichung Veterans General Hospital, Chiayi, Taiwan
(2) Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan
(3) Department of Pharmacy, Kaohsiung Municipal Chinese Medical Hospital, Kaohsiung, Taiwan
(4) Graduate Institute of Natural Products, Kaohsiung Medical University, Kaohsiung, Taiwan
E-mail: chenchihwen@yahoo.com.tw
(5) Department of Information and Computer Engineering, Chung Yuan Christian University, Taoyuan City, Taiwan
(6) Department of Information Management, National Central University, Taoyuan City, Taiwan
(7) Department of Computer Science, Tennessee Technological University, Cookeville, TN, USA

Abstract: *While there is an ample amount of medical information available for data mining, many of the datasets are unfortunately incomplete – missing relevant values needed by many machine learning algorithms. Several approaches have been proposed for the imputation of missing values, using various reasoning steps to provide estimations from the observed data. One of the important steps in data mining is data preprocessing, where unrepresentative data is filtered out of the data to be mined. However, none of the related studies about missing value imputation consider performing a data preprocessing step before* imputation. Therefore, the aim of this study is to examine the effect of two preprocessing steps, feature and instance selection, on missing value imputation. Specifically, eight different medical-related datasets are used, containing categorical, numerical and mixed types of data. Our experimental results show that imputation after instance selection can produce better classification performance than imputation alone. In addition, we will demonstrate that imputation after feature selection does not have a positive impact on the imputation result.

Keywords: missing value, imputation, feature selection, instance selection, incomplete medical datasets

1. Introduction

While there is an ample amount of medical information available for data mining, many of the datasets are unfortunately incomplete. This can generally be attributed to, among other things, manual data entry procedures, incorrect measurements or equipment errors. As a result, many real-world datasets usually contain missing (attribute) values or missing data (Lakshminarayan *et al.*, 1999). Unfortunately, this data quality problem can adversely affect data mining performance.

Because most existing data mining and machine learning algorithms cannot deal with incomplete data, the simplest and most straightforward solution is the case deletion approach, which does not consider examples with missing values. However, this method is generally appropriate only when the chosen dataset contains a very small amount of missing data.

Consequently, the aim of *missing value imputation* in data mining is to provide estimations for missing values by reasoning from the observed data (i.e. complete data) (Batista & Monard, 2003; Garcia-Laencina *et al.*, 2010). Some novel imputation methods have been proposed in recent studies (e.g. Zhang, 2008; Zhu *et al.*, 2011), and there have been comparative studies between the different imputation methods (e.g. Batista & Monard, 2003; Acuna & Rodriguez, 2004; Farhangfar *et al.*, 2008). All of these studies have demonstrated the effectiveness of the imputation methods using small to large missing rates over different types of datasets.

Related studies have demonstrated that missing value imputation is useful, and it is a better choice than case deletion when the chosen datasets contain a certain proportion of missing values. However, none of the related studies consider data preprocessing tasks, that is, feature selection (Guyon & Elisseeff, 2003) and/or instance selection (Garcia *et al.*, 2012), *before* imputation. The goals of feature and instance selection are to preprocess the collected datasets to filter out unrepresentative features (i.e. attributes) and outliers.

Estimating missing values is based on the observed data. However, some of the data may be non-representative, such as data that contains noise. Therefore, the aim of this paper is to examine whether individually performing feature and instance selection over the observed data can affect the imputation results, leading to different mining analysis results. Specifically, because the real-world datasets can contain categorical (i.e. discrete), numerical (i.e. continuous) or both types of data, and the missing rates vary between different datasets, our research objective is to provide some guidelines for data mining practitioners to determine when the feature and/or instance selection task should be performed before missing value imputation, quantified over which different types of datasets with different 'missing' rates.

The rest of this paper is organized as follows. Section 2 provides an overview of related research including the 'missingness' mechanisms, as well as the use of the k-nearest neighbour imputation method as the baseline imputation approach. In addition, feature and instance selection are overviewed. Section 3 presents the experimental setup and results. Finally, Section 4 provides some concluding remarks and targeted future work.

2. Literature review

2.1. *The missingness mechanisms*

The randomness of missing data can be divided into three categories: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little & Rubin, 1987).

MCAR deals with the highest level of randomness. In this case, if $P(X \mid x \textit{ missing}) = P(X \mid x \textit{ observed})$, where X is a random attribute, then the distribution of X is not affected by missing values. This refers to data where the *missingness* mechanism does not depend on the attribute of interest or any other attribute that is observed in the data. In other words, this occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data. Any missing data imputation can be applied to this level of randomness without risk of introducing bias on the data.

In contrast, for MAR, if $P(X \mid x \textit{ missing}, Z) = P(X \mid x \textit{ observed}, Z)$, where X is a random attribute and Z is a set of predictor attributes, then the distribution of X is not affected by missing values for $X \in Z$. In other words, this occurs when the probability of an instance having a missing value for an attribute could depend on the value of that attribute or, better yet, when the distribution of an instance having a missing value for an attribute depends on the observed data but does not depend on the missing data.

NMAR occurs when the probability of an instance having a missing value for an attribute could depend on the value of that attribute. This is the most difficult condition to model because, in practice, it is difficult to judge the missing data mechanism, as the values for the missing data are unknown.

2.2. *k*-nearest neighbour imputation

In the k-nearest neighbour imputation (kNNI) (Dixon, 1979), missing values are imputed using values calculated from the k-nearest neighbours. In particular, the nearest neighbours can be identified by minimizing the distance function, such as the Euclidean distance. Once the k-nearest neighbours have been found, a replacement value must be estimated to substitute for the missing attribute value.

The advantages of kNNI are that it can predict both qualitative attributes (the most frequent value among the k-nearest neighbours) and quantitative attributes (the mean among the k-nearest neighbours). In addition, unlike model-based imputation methods, it does not require creating a predictive model for each attribute with missing data.

An important parameter for the kNNI method is the value of k, which is typically set to 1 but is sensitive to outliers. However, Jonsson and Wohlin (2004) show that the performance is fairly unaffected by the value of k. On the other hand, Batista and Monard (2003) report that $k = 10$ for large datasets.

The algorithm proceeds as follows:

(1) Divide the dataset D into two parts. Let D_m be the set containing the instances in which at least one of the attributes is missing. The remaining instances with complete attribute information form a set called D_c.
(2) For each vector x in D_m:

    (1) Divide the instance vector into observed and missing parts so that $x = [x_o; x_m]$.
    (2) Calculate the distance between x_o and all the instance vectors from set D_c. Use only those attributes in the instance vectors from the complete set D_c, which are observed in vector x.
    (3) Use the k closest instance vectors (k-nearest neighbours) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes, replace the missing value using the mean value of the attribute in the k-nearest neighbourhoods.

For more detailed information on the other imputation methods, please refer to De Leeuw (2001) and Garcia-Laencina *et al.* (2010).

2.3. *Feature selection*

In general, not all the collected features of a chosen dataset are informative nor can they provide high discriminative power (Powell, 2007). Therefore, irrelevant and/or redundant features should be removed from the chosen dataset by feature selection, which can improve the performance of classification and clustering when data mining.

Feature selection can be defined as the process of choosing a minimum subset of n features from the original dataset of m features ($n < m$), so that the feature space (i.e. the dimensionality) is optimally reduced.

In this study, the F-score is used (Chen & Lin, 2006). This is a simple feature selection technique, which measures the discrimination of two sets of real numbers. Given training vectors x_k, $k = 1, 2, \ldots, m$, if the number of positive and negative instances are n_+ and n_-, respectively, then the *F*-score of the *i*th feature is defined as

$$F(i) = \frac{\left(\overline{x_i}^{(+)} - \overline{x_i}\right)^2 + \left(\overline{x_i}^{(-)} - \overline{x_i}\right)^2}{\frac{1}{n_+-1} \sum_{k=1}^{n_+} \left(\overline{x_i}_{k,i}^{(+)} - \overline{x_i}^{(+)}\right)^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_+} \left(\overline{x_i}_{k,i}^{(-)} - \overline{x_i}^{(-)}\right)^2} \tag{1}$$

where $\overline{x}_i$, $\overline{x}_{i^{(+)}}$ and $\overline{x}_{i^{(-)}}$ are the average of the *i*th feature of the whole, positive and negative datasets, respectively; $\overline{x}_{k,i^{(+)}}$ is the *i*th feature of the *k*th positive instance; and $\overline{x}_{k,i^{(-)}}$ is the *i*th feature of the *k*th negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the *F*-score, the more likely this feature is more discriminative.

Figure 1 shows an example of a dataset[1] with and without feature selection by multidimensional scaling (MDS) (Cox & Cox, 2001). As we can see, the data distributions in the 200 and 95 dimensional feature spaces are different. Therefore, estimations of missing values by the data with 200 dimensions and 95 dimensions could be different.

2.4. *Instance selection*

Similar to feature selection, the collected data are not all equally informative, and some data points can be considered noisy points or outliers. Outliers are unusual observations (or bad data points) that are far removed from the mass of data. In other words, an outlier is a value further away from the sample mean than what is deemed reasonable.

Therefore, the aim of instance selection, or record reduction, is to reduce the size of a dataset while still maintaining the integrity of the original (Wilson & Martinez, 2000). In some cases, generalization accuracy can increase when noisy instances are removed and when decision boundaries are smoothed to more closely match the true underlying function.

Instance selection can be defined as follows. Let X_i be an instance where $X_i = (X_{i1}, X_{i2}, \ldots, X_{im}, X_{ic})$, meaning that X_i is represented by m-dimensional features and X_i belongs to class c given by X_{ic}. Then, assume that there is a target set TA that consists of M instances, which is used for instance selection. Consequently, the subset of selected samples S are produced, where $S \subseteq TA$. Given a testing set TS, we can classify a new pattern T from TS over the instances of S and TA. If the instance selection algorithm has been chosen appropriately, the classifier performance trained by S should be better than TA.

In this work, IB3 (Aha *et al.*, 1991) is used as the instance selection algorithm because it can perform instance selection efficiently (where the computational complexity is $O(n^2 \log_2 n)$) and can provide reasonably good performance (Garcia *et al.*, 2012). This method utilizes an acceptable instance concept to carry out the selection. That is, instance x from the training set is added to a new set S if the nearest acceptable instance in S (if there is no acceptable instance a random one is used) is in a different class than x. Acceptability is defined by a confidence interval

$$\frac{p + \frac{z^2}{2n} \pm \sqrt{\frac{p(p-1)}{n} + \frac{z^2}{2n^2}}}{1 + \frac{z^2}{n}} \tag{2}$$

where z is the confidence factor (in IB3 0.9 is used to accept, 0.7 to reject), p is the classification accuracy of a given instance (while added to S) and n is equal to a number of classification trials for the given instance (while added to S).

Figure 2 shows the MDS results of the dataset used in Figure 1 with and without instance selection. We can observe that the data distribution in the reduced dataset after instance selection is different from the one in the original dataset. Therefore, the results of missing value imputation for 7809 and 3720 data samples could be different.

3. Experiments

3.1. *Experimental setup*

3.1.1. *The datasets and baseline classification accuracy*

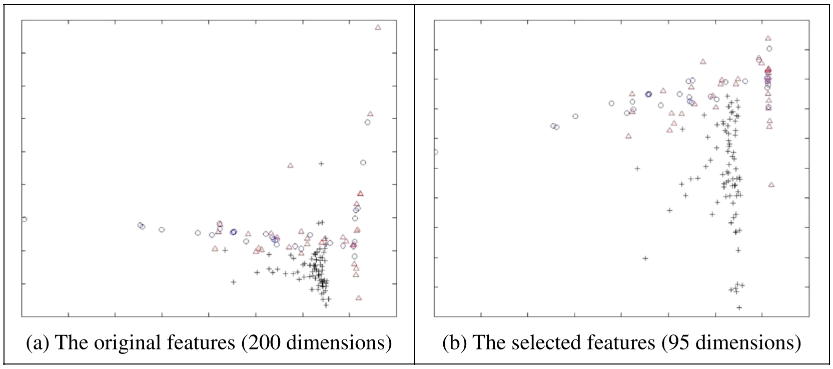
Eight medical-related datasets, which contain different types of data, are chosen from the UCI Machine Learning Repository.[2] They contain both categorical, numerical and mixed attribute types of data. Moreover, each type of dataset contains differing numbers of attributes, samples and classes, allowing us to determine the effect of varying different types of datasets with different missing rates on the final classification accuracy.

The *k*-nearest neighbour (k-NN) classifier ($k = 1$) is used for the classifier because 1-NN can be conveniently used as the baseline classifier, and this method is likely to provide a reasonable classification performance in most applications (Jain *et al.*, 2000). Moreover, after performing case deletion on each dataset, 10-fold cross-validation is used to divide each dataset into 90% training and 10% testing sets to train and test the 1-NN classifier. An example of these datasets and their classification accuracy is shown in Table 1.

Missing values for each dataset are randomly introduced into all attributes (using missing data rates of 5, 10, 15, 17 and 20 to 50% at 5% intervals) by the MCAR mechanism. In order to reduce the likelihood of obtaining biased results by randomly introducing missing values, each missing rate calculation is performed 20 times over each dataset, and the final classification accuracy for that dataset is based on averaging the 20 different classification results.

[1] The dataset is composed of 7809 data samples, and each data sample contains 200 features.

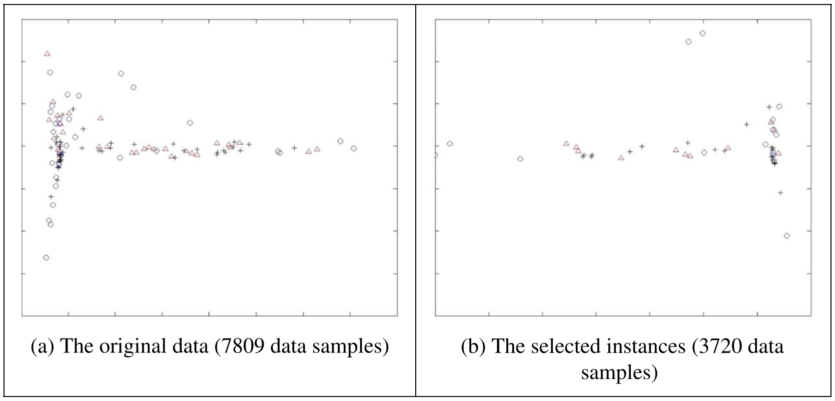
[2] http://archive.ics.uci.edu/ml/



(a) The original features (200 dimensions)

(b) The selected features (95 dimensions)

Figure 1: *An example of a dataset before and after feature selection.*



(a) The original data (7809 data samples)

(b) The selected instances (3720 data samples)

Figure 2: *An example of a dataset before and after instance selection.*

Table 1: *Dataset information*

Dataset	No. of instances	No. of attributes	No. of classes	Classification accuracy
Categorical datasets				
Lymphograph	148	18	4	75.68%
Numerical datasets				
Breast cancer	286	9	2	95.75%
Escherichia coli genes	336	8	8	80.36%
Pima Indian diabetes	768	8	2	70.18%
Yeast	1484	8	10	52.29%
Mixed data types of datasets				
Liver disorders	345	7	2	62.9%
Statlog	270	13	2	75.19%
Statlog_German	1000	20	2	69.7%

3.1.2. The process of combining feature selection and imputation In order to examine the effect of performing feature selection on missing value imputation, the experimental process is as follows. Given a dataset D with a missing value, the data with and without missing values can be grouped into complete ($D_{complete}$) and incomplete subsets ($D_{incomplete}$), where $D \in D_{complete} + D_{incomplete}$.

First of all, feature selection based on the F-score is performed over the complete subset $D_{complete}$. In particular, 50% of the important features of $D_{complete}$ are kept. That is, the top 50% of features having higher F-scores are selected. This leads to a reduced subset, denoted as $D_{complete\_feature\_reduced}$.[3] Next, the unimportant features of the incomplete subsets $D_{incomplete}$ identified in the preceding

[3] If too few features are filtered out, $D_{complete\_feature\_reduced}$ will be similar to $D_{complete}$, which makes it unlikely that performance differences will be shown. In fact, if too many features are filtered out, the discriminative power of $D_{complete\_feature\_reduced}$ will be much less than that of $D_{complete}$. Therefore, we think that selecting 50% of the original features should be a reasonable decision.

texts are removed, which results in a new reduced subset, denoted by $D_{incomplete\_feature\_reduced}$. Then, $D_{complete\_feature\_reduced}$ and $D_{incomplete\_feature\_reduced}$ are combined (denoted as $D_{feature\_reduced}$) for imputation by kNNI. Note that the number of data samples in $D_{feature\_reduced}$ is the same as D. Finally, the reduced dataset $D_{feature\_reduced}$ becomes complete after performing the imputation step, denoted as $D'_{feature\_reduced}$. Besides, another feature selection algorithm, that is, the genetic algorithm (GA) (Raymer *et al.*, 2000), is also used for comparison.

After performing imputation, the incomplete dataset D, denoted by D', is used as the baseline. In addition, 10-fold cross-validation over D' and $D'_{feature\_reduced}$ is applied to train and test the 1-NN classifier. Consequently, the classification accuracy of 1-NN over D' and $D'_{feature\_reduced}$ is compared to investigate the effects of feature selection.

3.1.3. The process of combining instance selection and imputation In order to combine instance selection and imputation, first instance selection based on IB3 is performed over the complete subset $D_{complete}$. As a result, a reduced subset, denoted as $D_{complete\_instance\_reduced}$, is produced. Note that unlike feature selection by *F*-score, using IB3 is not necessary to determine how many data samples should be removed. Next, $D_{complete\_instance\_reduced}$ and $D_{incomplete}$ are combined (denoted as $D_{instance\_reduced}$) for imputation by kNNI. Note that the number of features in $D_{instance\_reduced}$ is the same as D and the number of data samples in $D_{instance\_reduced}$ is smaller than the one in D. Finally, after performing imputation, the reduced dataset $D_{instance\_reduced}$ becomes complete, denoted as $D'_{instance\_reduced}$. Besides IB3, another instance selection algorithm, that is, DROP3 (Wilson & Martinez, 2000), is also used for comparison.

For performance comparison, 10-fold cross-validation over D' and $D'_{instance\_reduced}$ is applied to train and test the 1-NN classifier individually. Consequently, the classification accuracy of 1-NN over D' and $D'_{instance\_reduced}$ are compared to examine the effects of instance selection.

3.2. Results from datasets with specific missing rates

Table 2 shows the classification results obtained by case deletion, imputation (by kNNI), imputation after feature selection (FS) and imputation after instance selection (IS) over different datasets. Note that the missing rate allowed for each dataset means that the classification accuracy by case deletion is similar to the one over the complete dataset (i.e. the baseline). In particular, the level of performance similarity between case deletion and the baseline is defined by less than a 5% classification difference. For instance, using the complete dataset of Lymphograph, the classification accuracy of k-NN is 75.68%. Therefore, the threshold for using case deletion for this dataset is that the classification accuracy should be higher than 71.9% (i.e. 75.68% × 0.95). In this case, the classification accuracy of k-NN over the Lymphograph dataset with missing rates of 10 and 15% will be 71.95% (higher than the threshold) and 55.14% (lower than the threshold), respectively. Therefore, the allowable missing rate is 10%. So, in this example, a Lymphograph dataset with 10% missing values is used for imputation, FS with imputation and IS with imputation.

From these results, we can determine the suitability of using case deletion for different types of incomplete datasets containing different rates of missing values. In addition, incomplete dataset imputation does not necessarily provide better results than the case deletion method when the datasets contain certain missing rates. That is, the average classification accuracy by case deletion is higher than the one by imputation.

Our research objective is to compare the performances obtained by imputation and imputation after FS and IS. The average results indicate that performing FS first over incomplete datasets does not positively affect the final imputation results. In other words, the classification results obtained by both approaches are very similar. On the other hand, the average classification accuracy obtained by performing IS first and missing value imputation second is much higher than the one obtained by performing missing value imputation alone.

3.3. Results on specific datasets with various missing rates

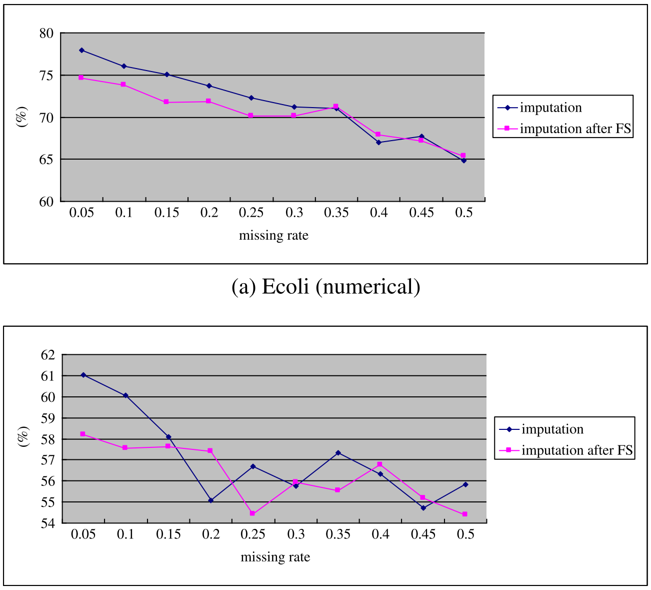
We further examine the datasets where there are significant performance differences between imputation and imputation after FS and IS. Note that we use *F*-score and IB3 to represent FS and IS, respectively, because they provide (slightly) better performances than GA and

Table 2: *Classification results over different datasets*

Datasets	Missing rate allowed	Case deletion	Imputation	Imputation after FS (*F*-score/GA)	Imputation after IS (IB3/DROP3)
Lymphograph	10%	68.05%	72.57%	70%/69.8%	76.83%/76.39%
Breast cancer	45%	96.95%	93.06%	91.77%/92.31%	92.75%/88.8%
Escherichia coli genes	30%	79.78%	71.25%	70.12%/68.52%	71.51%/ 67.44%
Pima genes	10%	68.02%	67.73%	67.63%/67.66%	78.07%/75.11%
Yeast	10%	49.3%	47.59%	44.53%/45.77%	65.31%/ 52.14%
Liver disorders	10%	60.3%	60.06%	57.57%/58.35%	66.16%/ 59.45%
Statlog	15%	72.16%	73.26%	73.26%/71.26%	76.58%/65.56%
Statlog_German	10%	64.89%	66.64%	65.8%/65.02%	69.52%/69.93%
Avg.		69.93%	69.02%	67.59%/67.34%	74.59%/69.35%

DROP3, respectively. Figures 3 and 4 show the results of imputation after FS versus imputation and imputation after IS versus imputation, using datasets with various missing rates (i.e. from 5 to 50%). Note that for some datasets with certain missing rates, the imputation algorithm cannot be performed because there are not enough 'complete' examples in the training data for missing value imputation.

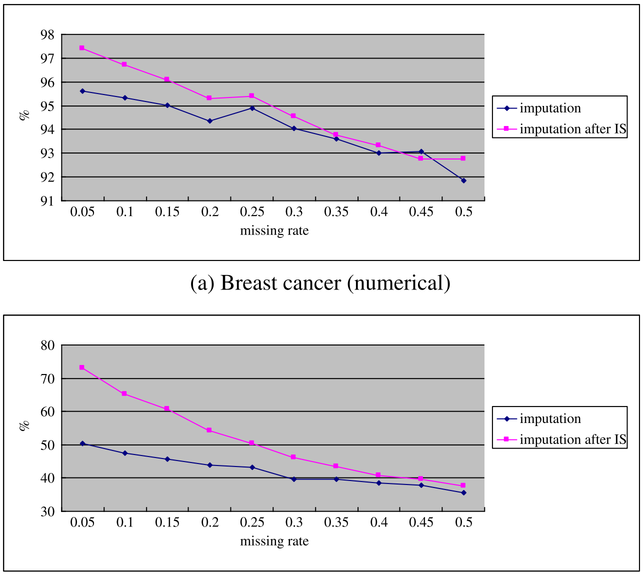
Performing FS before imputation is likely to a have negative impact on the imputation results, which leads to poorer classification performance than performing imputation alone. The results indicate that performing IS before imputation can have a positive impact that results in higher classification accuracy than performing imputation alone. In particular, the performance improvement is significant, that is, $p < 0.05$.



(a) Ecoli (numerical)

(b) Liver_disorders (mixed)

Figure 3: *Classification accuracy after feature selection and imputation.*



(a) Breast cancer (numerical)

(b) Yeast (numerical)

Figure 4: *Classification accuracy after instance selection and imputation.*

These results are consistent with results presented earlier that performing IS first and imputation second can improve the final classification accuracy. However, performing FS before imputation cannot provide reasonable improvement.

In other words, removing some outliers (or noisy data) from the (complete) training dataset allows the imputation algorithm to produce better estimations for missing values, which results in higher classification accuracy than that obtained without performing IS.

4. Conclusion

Many real-world medical datasets are usually incomplete, containing some missing attribute values. Missing value imputation is one of the common approaches taken to solve the incomplete dataset problem. Although there are many different types of imputation algorithms mentioned in the literature, there have been no related studies examining whether performing data preprocessing, that is, FS and IS, has any impact on the final imputation results.

In this study, our aim is to compare the classification performance obtained with both data preprocessing tasks before imputation and by imputation alone. Three types of medical datasets, including categorical, numerical and the mixed type of data, are used in our determination of the effect of FS and IS on missing value imputation and understanding when we should consider FS or IS before imputation.

Our experimental results show that imputation after IS can produce better classification performance than imputation alone, while imputation after FS does not have a positive impact on the imputation result.

In the future, we hope to further examine large scale, or 'big data', datasets, which contain a very large number of features and data samples, as well as high-dimensional datasets, where each data sample is represented by a large number of attributes. In addition, it is worth investigating whether we should directly ignore examples with missing values without performing imputation. The initial results indicate that better classification accuracy is obtained with the case deletion method than performing imputation over some datasets containing specific missing rates. Moreover, other statistical and supervised learning based imputation methods, such as naïve Bayes, neural networks, etc., can be employed for further comparison.

References

Acuna, E. and C. Rodriguez (2004.) The treatment of missing values and its effect in the classifier accuracy. Banks D, House L, McMorris FR, Arabie P, and Gaul W (eds). *Classification,* *clustering and data mining applications*, Springer-Verlag: Chicago, 639–648.

AHA, D.W., D. KIBLER and M.K. ALBERT (1991) Instance-based learning algorithms, *Machine Learning*, **6**, 37–66.

BATISTA, G. and M. MONARD (2003) An analysis of four missing data treatment methods for supervised learning, *Applied Artificial Intelligence*, **17**, 519–533.

CHEN, C.-C. and C.-J. LIN(2006.)Combining SVMs with various feature selection strategies. Guyon I, Gunn S, Nikravesh M, Zadeh LA (eds). *Feature extraction, foundations and applications*, Springer: Berlin.

COX, T.F. and M.A.A. COX(2001.)*Multidimensional scaling*, Chapman and Hall: Boca Raton, Florida.

DE LEEUW, E. (2001) Reducing missing data in surveys: an overview of methods, *Quality & Quantity*, **35**, 147–160.

DIXON, J.K. (1979) Pattern recognition with partly missing data, *IEEE Transactions on Systems, Man, and Cybernetics*, **10**, 617–621.

FARHANGFAR, A., L. KURGAN and J. DY (2008) Impact of imputation of missing values on classification error for discrete data, *Pattern Recognition*, **41**, 3692–3705.

GARCIA, S., J. DERRAC, J.R. CANO and F. HERRERA (2012) Prototype selection for nearest neighbor classification: taxonomy and empirical study, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 417–435.

GARCIA-LAENCINA, P.J., J.-L. SANCHO-GOMEZ and A.R. FIGUEIRAS-VIDAL (2010) Pattern classification with missing data: a review, *Neural Computing and Applications*, **19**, 263–282.

GUYON, I. and A. ELISSEEFF (2003) An introduction to variable and feature selection, *Journal of Machine Learning Research*, **3**, 1157–1182.

JAIN, A.K., R.P.W. DUIN and J. MAO (2000) Statistical pattern recognition: a review, *IEEE Transitions on Pattern Analysis and Machine Intelligence*, **22**, 4–37.

JONSSON, P. and C. WOHLIN (2004) An evaluation of k-nearest neighbor imputation using likert data, *IEEE International Symposium on Software Metrics*, 108–118.

LAKSHMINARAYAN, K., S.A. HARP and T. SAMAD (1999) Imputation of missing data in industrial databases, *Applied Intelligence*, **11**, 259–275.

LITTLE, R.J.A. and D.B. RUBIN (1987.) *Statistical analysis with missing data*, John Wiley and Sons: New Jersey.

POWELL, W.B. (2007.) *Approximate dynamic programming: solving the curses of dimensionality*, Wiley-Interscience: New Jersey.

RAYMER, M.L., W.F. PUNCH, E.D. GOODMAN, L.A. KUHN and A.K. JAIN (2000) Dimensionality reduction using genetic algorithms, *IEEE Transactions on Evolutionary Computation*, **4**, 164–171.

WILSON, D.R. and T.R. MARTINEZ (2000) Reduction techniques for instance-based learning algorithms, *Machine Learning*, **38**, 257–286.

ZHANG, S. (2008) Parimputation: from imputation and null-imputation to partially imputation, *IEEE Intelligent Informatics Bulletin*, **9**, 32–38.

ZHU, X., S. ZHANG, Z. JIN, Z. ZHANG and Z. XU (2011) Missing value estimation for mixed-attribute data sets, *IEEE Transactions on Knowledge and Data Engineering*, **23**, 110–121.

The authors

Min-Wei Huang

Dr Min-Wei Huang is currently a director of Taichung Veterans General Hospital, Chiayi branch.

Wei-Chao Lin

Dr Wei-Chao Lin is an associate professor of Asia University.

Chih-Wen Chen

Mr Chih-Wen Chen is currently a doctor of Kaohsiung Municipal Chinese Medical Hospital.

Shih-Wen Ke

Dr Shih-Wen Ke is an assistant professor of Chung Yuan Christian University.

Chih-Fong Tsai

Dr Chih-Fong Tsai is professor of National Central University.

William Eberle

Dr William Eberle is an associate professor of Tennessee Technological University.