Parimputation: From Imputation and Null-Imputation to Partially Imputation

Shichao Zhang, *Senior Member, IEEE*

***Abstract*— Missing data imputation is an important step in the process of machine learning and data mining when certain values are missed. Among extant imputation techniques, kNN imputation algorithm is the best one as it is a model free and efficient compared with other methods. However, the value of *k* must be chosen properly in using kNN imputation. In particular, when some nearest neighbors are far from a missing data, the kNN imputation algorithms are often of low efficiency. In this paper, a new imputation framework is designed. The imputation uses the left or right nearest neighbor for a missing data in a given dataset. Furthermore, a parimputation (partially imputation) strategy is proposed for dealing with the issue of missing data imputation. Specifically, some missing data are imputed when there are some complete data in a small neighborhood of the missing data and, other missing data without imputation are given up in applications, such as data mining and machine learning.**

***Index Terms*—Artificial intelligence; Data management; Data processing.**

I. INTRODUCTION

In real applications, missing value imputation is an actual and challenging problem confronted by machine learning and data mining. Therefore, there are great many efforts to missing value imputation. Traditional missing value imputation techniques can be roughly classified into regression imputation (RI) and nearest neighbor imputation (NNI) [33]. And missing values in a dataset are completed by replacing them with some plausible values. The plausible values are generally generated from the dataset using an imputation method.

RI can be classified into deterministic regression imputation (DRI) and stochastic regression imputation (SRI). Using a DRI method, missing values in a dataset are replaced with only the mean of all the known values in the dataset. Using an SRI method, each of missing values is replaced with the mean plus a random value. Experiment results have proven [22] that SRI methods are much better than DRI methods in many practical cases. However, it is usually more difficult to mathematically prove the efficiency for SRI methods.

NNI [33] is one of the hot deck techniques used to compensate for missing data. It has been successfully used in, for example, U.S. Census Bureau and Canadian Census Bureau. Using an NNI method, a missing value in a dataset is replaced with the value of the nearest neighbor in the dataset. kNNI (*k*-nearest-neighbors imputation) is an extension of NNI method (It is an NNI algorithm when $k = 1$). It takes into account *k* nearest neighbors when imputing. Yet, it is difficult to mathematically prove the efficiency for kNNI methods.

While having good randomicity, SRI methods are poor in efficiency when compared with kNNI techniques. However, the value of *k* must be selected properly when using kNNI methods. In particular, the nearest neighbor may be far from a missing data and the kNNI methods are thus of low efficiency. In this paper a new imputation framework is designed. Furthermore, it advocates giving up imputation if there is no close neighbors and only imputing those missing data that the nearest neighbor is not far from them. It is referred to a parimputation (partially imputation) strategy.

The rest of this paper is organized as follows. Section II briefly recalls related work on missing value imputation. In Section III we present an imputation framework. In Section IV, we design the parimputation strategy. We simply evaluate the proposed approach in Section V. This paper is concluded in Section VI.

Shichao Zhang is with the College of Computer Science and Information Technology, Guangxi Normal University, PR China; the State Key Lab for Novel Software Technology, Nanjing University, PR China; e-mail: zhangsc@mailbox.gxnu.edu.cn.

II. RELATED WORK

The missing data problem is faced in many application domains, such as, statistical analysis, machine learning, data mining, pattern recognition and information retrieval. Because imputation algorithms are designed independent of applications, we only review major related work in the application domains of statistical analysis and data mining in this section.

A. Research into Statistical Imputation for Missing Data

Statistical analysis with missing data has been noted in the literature for more than 70 years. Wilks [28] initiated a study on the maximum likelihood estimation for multivariate normal models with fragmentary data. Thereafter, extensive discussions on this topic continue. A useful reference for general parametric statistical inferences with missing data can be found in [16].

Little and Rubin [15] classified missing data mechanisms into three categories as follows.

1. **Missing Completely at Random** (MCAR): Cases with complete data are indistinguishable from cases with incomplete data. Heitjan [9] provided an example of

MCAR missing data and, Graham, Hofer and MacKinnon [12] illustrated the use of planned missing data patterns.

2. **Missing at Random** (MAR): Cases with incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing.
3. **Nonignorable**: The pattern of data missingness is non-random and it is not predictable from other variables in the database.

In practice it is usually difficult to meet the nonignorable assumption. MAR is an assumption that is more often (MCAR is a special case of MAR), but not always tenable. The more relevant and related predictors one can include in statistical models, the more likely it is that the MAR assumption will be met.

B. *Research into Missing Data Imputation in Data Mining*

Recently, Magnani [17] has reviewed the main missing data techniques, including conventional methods, global imputation, local imputation, parameter estimation and direct management of missing data. He tried to highlight the advantages and disadvantages for all kinds of missing data mechanisms. For example, he revealed that statistical methods have been mainly developed to manage survey data and proved to be very effective in many situations. However, the main problem of these techniques is its strong model assumptions.

Batista and Monard [3] have analyzed the performance of 10-NNI as an imputation method, comparing its performance with other three missing data imputation methods: mean or mode imputation, C4.5 and CN2. This work proposed the advantages of the method: it can predict both qualitative attributes and quantitative attributes, and it does not create explicit modes (like a decision tree or a rules) because it is a lazy model. Their experiments showed that the method provides very good results than the other three methods, even for a large amount of missing data. A main drawback is that the algorithm must search through all the data set limiting in large databases only based on MCAR. Different imputations for industrial databases have also been studied in [12].

Yuan [30] reviewed three methods of multiple imputation for missing data, including regression method, propensity score method and MCMC (Markov Chain Monte Carlo) method. Also, he used standard statistical methods to evaluate the efficiency of multiple imputation.

Allison [2] has evaluated two algorithms for producing multiple imputations or missing data using simulated data based on the software of SOLAS. Software using a propensity score classifier with the approximate Bayesian boostrap was found to produce badly biased estimates of regression coefficients when data on predictor variables are MAR or MACR. Allison has also showed that listwise deletion produces unbiased regression estimates whenever the missing data mechanism depends only on the predictor variable, not on the response variable.

Other missing data imputation methods include a new family of reconstruction problems for multiple images from minimal data [11], a method for handling inapplicable and unknown missing data [8], different substitution methods for replacement of missing data values [20], robust Bayesian estimator [26], and nonparametric kernel classification rules derived from incomplete (missing) data [18].

III. An Imputation Framework for Dealing with Missing Values

Let X be a d-dimensional vector of factors and let Y be a response variable influenced by X. In practice, one often obtains a random sample (sample size = n) of incomplete data associated with a population (X, Y, δ),

$$(X_i, Y_i, \delta_i),\ i = 1, 2, \ldots, n$$

Where all the X_i's are observed and $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. Suppose that (X_i, Y_i) satisfies the following model:

$$Y_i = m(X_i) + \varepsilon_i,\ i = 1, 2, \ldots, n$$

Where $m(.)$ is an unknown function, and the unobserved ε_i (with population ε) are i.i.d. random errors with mean 0 and unknown finite variance σ^2, and are independent of the i.i.d. random variables X_i's.

To impute the missing values, $m(.)$ must be estimated. The $m(.)$ are often measured the statistical parameters of the response variable Y such as $\mu = EY$, $\theta = F(y)$ and θ_q, i.e. the mean, the distribution function and the q-th quantile of Y, where y is a fixed point in $\Re$, and $0<q<1$. EY stands for the average level of Y, the distribution function $F(y)$ is the probability of Y being smaller than or equal to the given y, and θ_q is the level of Y that satisfies $P(Y \le \theta_q) = q$. The median of Y (the case of $q = 1/2$) is the most important case of quantiles. The inference for them is a very important issue in practice.

In the situation where $m(.)$ is a linear function, i.e. Y and X fit a linear model, Wang and Rao [29] have compared the adjusted empirical likelihood methods and the normal approximation methods in terms of coverage accuracies and average lengths of the confidence intervals. They have indicated that the adjusted empirical likelihood methods perform competitively, the use of auxiliary information provides improved inferences and the deterministic imputation method performs well in making inference for the mean of Y. Qin et al. [21] have showed that one must use random imputation methods in making inference for distribution functions and quantiles of Y.

Yet in many complex practical situations, $m(.)$ (an unknown function) is not a linear function. When we do not know the form of $m(.)$, i.e. the nonparametric situation, Wang and Rao [29] have considered empirical likelihood inference on the mean of response Y when Y is missing at random (MAR) . They have only used the deterministic imputation method to infer the mean of Y, and left the inference for distribution functions and quantiles of Y unsolved.

To avoid estimating $m(.)$, NNI method replaces a missing value in a dataset with the value of the nearest neighbor in the dataset. Further, kNNI method is proposed, which replaces a missing value in a dataset with the mean of k nearest neighbors when imputing. NNI or kNNI algorithms have experimentally been proved more efficient than other existing imputation methods [33]. They have widely been used in applications. However, as mentioned before, (1) it must seek a proper k when using kNN imputation methods; and (2) when some nearest neighbors are far from a missing datum, the kNN imputation algorithms are often of low efficiency. The first issue is tackled in this section and the second one will be dealt with in next section.

A. Imputation Model

This subsection builds a new imputation model that uses the left or right nearest neighbor for a missing data in a given dataset.

For a 2-dimensional imputation problem, let T_1 = $(X_i, Y_i, 1)$, T_2 = $(X_j, Y_j, 1)$, T = $(X_l, Y_l, 0)$ in a dataset, where T_1 and T_2 are the left and right nearest neighbors of an incomplete data T with respect to the factor X, respectively. That is, for any complete data T_3 = $(X_k, Y_k, 1)$ in the dataset, we have either

$$X_k \le X_i \text{ or } X_k \ge X_j$$

With T_1 and T_2, we can replace Y_l with the mean of Y_i and Y_j, or

$$Y_l = \frac{1}{2}(Y_i + Y_j)$$

In the same reason, for an $(n+1)$-dimensional imputation problem, we select such $2n$ complete data, T_1^-, T_1^+, …, T_n^-, T_n^+ from a given dataset, where T_i^-, T_i^+ are the left and right nearest neighbors of an incomplete data T with respect to the factor X_i , respectively. Formally, let T = $(X_{l1}, X_{l2}, ..., X_{ln}, Y_l, 0)$ in the dataset, NN is a set of all nearest neighbors of T in the dataset, and T's left and right nearest neighbors with respect to the factor X_i are as follows:

$$T_i^- = (X_{i1}^-, X_{i2}^-, ..., X_{in}^-, Y_{i-}, 1),\ i = 1, 2, \ldots, n$$

$$T_i^+ = (X_{i1}^+, X_{i2}^+, ..., X_{in}^+, Y_{i+}, 1),\ i = 1, 2, \ldots, n$$

where T_i^- or T_i^+ may not exist in NN. They satisfy that, for a nearest neighbor $(X_{j1}, X_{j2}, ..., X_{jn}, Y_{j+}, 1)$ in NN, either $X_{ji} \le X_{ii}^-$ if there is a T_i^- in NN, or $X_{ji} \ge X_{ii}^+$ if there is a T_i^+ in NN.

With these nearest neighbors, we can replace Y_l with the mean of all the Y_{i-} and Y_{i+} . Or

$$Y_l = \frac{1}{2n}\sum_{i=1}^{n}(Y_{i-} + Y_{i+}) \quad (2)$$

B. Model Enhancement

In Section III.A we have proposed a simple and easy-implemented imputation model. From the selection of the left and right nearest neighbors of a missing datum with respect to the factor X_i , there are three cases as follows.

1. There may be no left or right nearest neighbor for a missing data in a given dataset, with respect to the factor X_i .
2. A complete data may be selected multiple times in the set of left /right nearest neighbors of a missing data in a given dataset, with respect to the factor X_i .
3. Some left or right nearest neighbors of a missing data in a given dataset, with respect to the factor X_i may be far from the missing data.

For the first case, we can simply give up all the missed left or right nearest neighbors when estimating the missing data. The second case shows that fact: the more times a complete data is selected, the closer to the missing data the complete data is.

For the third case, we can use weighting technique to weaken their impact to the missing data when estimating the missing data. The weight of a left or right nearest neighbor of a missing data can be determined as follows.

For a left or right nearest neighbor T_i = $(X_{i1}, X_{i2}, ..., X_{in}, Y_i, 1)$ of a missing data T = $(X_{l1}, X_{l2}, ..., X_{ln}, Y_l, 0)$, we obtain

$$d_i = \sqrt{(X_{i1} - X_{l1})^2 + ... + (X_{in} - X_{ln})^2}$$

Hence, we can get the weight w_i of T_i as follows.

$$w_i = 1 - \frac{d_i}{d_1 + d_2 + ... + d_m} \quad (3)$$

Where, "m" is the number of the selected left or right nearest neighbors of the missing data. With these weights, we can estimate Y_l as follows.

$$Y_l = \sum_{i=1}^{n} (w_{i-} Y_{i-} + w_{i+} Y_{i+}) \qquad (4)$$

Further, we can waive all the left or right nearest neighbors that are far from the missing data according to d_i or w_i . In other words, we can select those left or right nearest neighbors that are very close to the missing data. After filtering some nearest neighbors, it is easy to estimate Y_l by improving Eqns. (3) and (4).

C. Imputation Framework

From the above, our new approach, called ENI (encapsidated-neighbor imputation), is similar to kNNI method. There are two main differences between ENI and kNNI as follows:

1. The ENI approach takes into account the left and right nearest neighbors of a missing data, whereas the kNNI method selects *k* nearest neighbors.
2. In ENI approach, the number of the selected nearest neighbors is a variable determined by data when imputing missing data, whereas the kNNI method uses a fixed *k*.

With the ENI approach, the process of missing data imputation is as follows.

Let *X* be a *n*-dimensional vector of factors, *Y* a response variable influenced by *X*, a dataset of incomplete data associated with a population (X, Y, δ) be as follows

$$(X_i, Y_i, \delta_i),\ i = 1, 2, \ldots, N$$

1. For each incomplete data *T* = $(X_{l1}, X_{l2}, ..., X_{ln}, Y_l, 0)$, search all the left or right nearest neighbor of Y: $T_1, T_2, \ldots, T_m$;
2. Use the Eqn (3) to calculate the weight w_i of T_i , *i* = 1, 2, …, *m*;
3. Estimate Y_l with Eqn (4);
4. Repeat Steps 1-3 until no incomplete data in the dataset.

This process is simple and easy to be understood and implemented.

IV. PARIMPUTATION: PARTIALLY IMPUTATION

From Section III, the ENI method takes into account all the left or right nearest neighbors of missing data when imputing them. However, like kNNI algorithms, it still suffers from the fact: sometimes all the left or right nearest neighbors can be far from a missing data in a dataset. When this case happens and the missing data is imputed with ENI or kNNI method, the results from the dataset can be inaccurate. To deal with this issue, this paper advocates a parimputation strategy: some missing data are imputed when there are some complete data in a small neighborhood of the missing data and, other missing data without imputation are given up in applications, such as data mining and machine learning.

For understanding the strategy, in this section, we first review the known value strategy and the null strategy that have been widely used in machine learning and data mining applications for dealing with missing data [22], and then propose the parimputation strategy, regarded as a new strategy.

A. Known Value Strategy for Missing Data

In cost-sensitive learning, the first tree building and test strategy for "missing is useful" is called the Known Value Strategy [14] [31]. It utilizes only the known attribute values in the tree building for each test example. For each test example, a new (and probably different) decision tree is built from the training examples with only those attributes whose values are known in the test example. That is, the new decision tree only uses attributes with known values in the test example, and thus, when the tree classifies the test example, it will never encounter any missing values.

The Known Value Strategy was proposed in [14] but its ability of handling unknown values was not studied. Clearly, the strategy utilizes all known attributes and avoids any missing data directly.

In [14], an internal node strategy was also proposed. It keeps examples with missing values in internal nodes, and does not build branches for them during tree building. When classifying a test example, if the tree encounters an attribute whose value is unknown, then the class probability of training examples falling at the internal node is used to classify it. As unknown values are dealt with using internal nodes, this strategy is called as the Internal Node Strategy.

As there might be several different situations where values are missing, leaving the classification to the internal nodes may be a natural choice. This strategy is also quite efficient as only one tree is built for all test examples.

B. Null Strategy

As values are missing for a certain reason – unnecessary and too expensive to test – it might be a good idea to assign a special value, often called "null" in databases [6], to missing data. The null value is then treated just as a regular known value in the tree building and test processes. This strategy has also been proposed in machine learning [1].

One potential problem with the Null Strategy is that it does not deliberately utilize the known values, as missing values are treated just as a known value. Another potential drawback is that there might be more than one situation where values are missing. Replacing all missing values by one value (null) may not be adequate. In addition, subtrees can be built under the "null" branch, suggesting oddly that the unknown is more discriminating than known values. The advantage of this strategy is its simplicity and high efficiency compared to the Known Value Strategy, as only one decision tree is built for all test examples.

Also, C4.5 [23][24] does not impute missing values explicitly, and it is shown to be quite effective [3]. And C4.5's missing-value strategy is applied directly in cost-sensitive trees. During training, an attribute is chosen by the maximum cost reduction discounted by the probability of missing values of that attribute. During testing, a test example with missing value is split into branches according to the portions of training examples falling into those branches, and goes down to leaves simultaneously. The class of the test example is the weighted classification of all leaves.

C. Parimputation Strategy

As described previously, the parimputation is a strategy for dealing with the issue of missing data imputation. The parimputation strategy is proposed for addressing those missing data in a given dataset that all the left or right nearest neighbors are far from them.

From the observed part of an incomplete datum in a dataset, if there are some complete data in a small neighborhood of the incomplete data, we refer it to a predictable missing data; otherwise, we refer it to an unpredictable missing data. With the observed part of an unpredictable missing data in a dataset, seeking the unpredictable missing data is similar to that of detecting outliers (or isolation points) in machine learning and data mining. This means that there are many well-established outlier detection techniques (such as [10] [25]) that can be applied to determining whether a missing data is unpredictable.

With the parimputation strategy, we can deal with the missing data in two ways as follows:

1. Impute all the predictable missing data in a dataset; remove all the unpredictable missing data from the dataset, and then discover patterns from the dataset that contains complete data and imputed data.
2. Impute only the predictable missing data in a dataset; and then discover patterns from the dataset with the known value strategy, or the null strategy.

From the above, the parimputation strategy is simple and easy to be understood and implemented.

V. EXPERIMENTS

In order to show the effectiveness of the proposed methods, extensive experiments were done on a real dataset with the algorithm implemented in C++ and executed using a DELL Workstation PWS650 with 2G main memory, and 2.6G CPU.

A. Algorithm Design

As mentioned previously, the ENI is simple and easy to be understood and implemented. However, the description is only used to state the problem. We should select the left and right nearest neighbors of a missing data from a set of nearest neighbors of the missing data. There three cases as follows.

(1) There is no nearest neighbor in the set, i.e., the missing data is unpredictable and it is not imputed in our experiments.
(2) The number of left and right nearest neighbors of a missing data is often lesser than *k*. This means that there are only few data observed in the set of nearest neighbors.
(3) The number of left and right nearest neighbors of a missing data is greater than *k*. This means that there are plenty data observed in the set of nearest neighbors.

These indicate that the number of selected left and right nearest neighbors is variable when imputing missing data. In particular, we can only select the left and right nearest neighbors from the *k* nearest neighbors that are selected for a kNNI algorithm.

Because the goal of this paper is to introduce a new imputation strategy, we simply evaluate the ENI in next subsection with compared with the kNN method for imputing continuous missing target attributes in terms of imputation accuracy.

B. Experimental Results

The first set of experiments was conducted on a real dataset of a class in a high school. The dataset contains 711 instances in total and 12 attributes for each instance (non missing attribute values). The average score was selected as the target attributes (response variable, Y) and, the Math (X_1), Chinese (X_2) and English (X_3) as the factors, where $Y = X_1 + X_2 + X_3$. We used the missing mechanisms MCAR and MAR on Y at different missing rates of 5%, 10% and 20%. Then the ENI and kNNI algorithms were utilized to fill out the missing values of Y. Our experiments have demonstrated that the ENI is much better than kNNI method at the efficiency for this linear function.

The second set of experiments was conducted on a real dataset, *Abalone*, downloaded from UCI machine learning repository. We selected 1528 instances where 7 attributes were picked as the factors and another one as the response variable. We use the missing mechanisms MCAR and MAR on Y at different missing rates of 5%, 10% and 20%. Then the ENI and kNNI algorithms were utilized to fill out the missing values of Y. The experimental results are listed in Tables 1-3.

From Tables 1, 2 and 3, the ENI is much better than kNN method. In particular, when the missing rate is 20%, the ENI is better than kNN method in each imputation times. This demonstrates that using only the left and right nearest neighbors can improve the imputation performance kNNI methods.

This research is focused on the case that only one attribute is with missing values. If several attributes are with missing values. The use of the ENI is as follows.

(1) Select such an attribute as the response variable that the number of its missing values is minimal among the attributes with missing values.

(2) Use the ENI to impute the missing values based on all complete attributes [1] (without missing values).

(3) Repeat Steps (1) and (2) until all predictable missing values are imputed.

[1] From the second imputation, the imputed attributes are taken as complete attributes.

Table 1. When the missing rate is 5%, there are 76 instances with missing data in *Abalone*.

Imputation times	1	2	3	4	5	6	7	8	9	10
ENI	**41**	**42**	**42**	35	**44**	**47**	**39**	**38**	**43**	**48**
kNNI	35	34	34	**41**	32	29	37	**38**	33	28

Table 2. When the missing rate is 10%, there are 153 instances with missing data in *Abalone*.

Imputation times	1	2	3	4	5	6	7	8	9	10
ENI	**80**	**78**	**86**	75	**80**	**78**	**82**	76	**88**	**90**
kNNI	73	75	67	**78**	73	75	71	**77**	65	63

Table 3. When the missing rate is 20%, there are 306 instances with missing data in *Abalone*.

Imputation times	1	2	3	4	5	6	7	8	9	10
ENI	**170**	**165**	**162**	**158**	**174**	**168**	**155**	**159**	**154**	**163**
kNNI	136	141	144	148	132	138	151	147	152	143

VI. Conclusions

In this paper we have proposed a new imputation, called ENI. It is different from the kNNI method because

1. The ENI approach takes into account the left and right nearest neighbors of a missing data, whereas the kNNI method selects *k* nearest neighbors.
2. In ENI approach, the number of the selected nearest neighbors is variable when imputing missing data, whereas the kNNI method uses a fixed *k*.

From the extrapolation, the ENI approach is more reasonable than the kNNI method. Further, a parimputation strategy has been advocated for dealing with the unpredictable missing data in a dataset. The experimental results have demonstrated that the ENI is much better than the kNNI method.

The future work is to apply the ENI approach and the parimputation strategy to real machine learning and data mining applications, so as to improve the methods.

VII. Acknowledge

Thanks for the experiments carried out by my student, Mr Manlong Zhu. Thanks for the comments on the early version of this paper from Dr Yongsong Qin, Mr Xiaofeng Zhu, and Dr. William K. Cheung.

This work was supported in part by the Australian Research Council (ARC) under grant DP0985456, the Nature Science Foundation (NSF) of China under grant 90718020, the China 973 Program under grant 2008CB317108, the Research Program of China Ministry of Personnel for Overseas-Return High-level Talents, the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (07JJD720044), and the Guangxi NSF (Key) grants.

References

[1] Ali, K.M. and Pazzani, M.J. (1993). Hydra: A noise-tolerant relational concept learning algorithm. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (IJCAI93), pp. 1064-1071. Morgan Kaufmann, 1993.

[2] Allison, P. (2001). *Missing Data*. Sage Publication, Inc, 2001. [place of publication]

[3] Batista G. and Monard, M.C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, Vol. 17, pp. 519-533, 2003.

[4] Caruana, R, (2001). A Non-parametric EM-style algorithm for Imputing Missing Value. *Artificial Intelligence and Statistics*, January 2001.

[5] Chen, J., and Shao, J., (2001). Jackknife variance estimation for nearest-neighbor imputation. *J. Amer. Statist.* Assoc. 2001, Vol. 96: 260-269.
[6] Date, C.J. and Darwen, H. (1989). The default values approach to missing information. In: *Relational Database Writings 1989-1991*, pp. 343-354, 1989.
[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, series B, Vol. 39, pp. 1–38.
[8] Gessert, G., (1991). Handling Missing Data by Using Stored Truth Values. *SIGMOD Record,* 2001, Vol. 20(3): 30-42.
[9] Heitjan, D.F. Annotation (1997). What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4): 548-550
[10] John, G. H. (1995). Robust Decision Trees: Removing Outliers from Databases. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 174–179. Menlo Park, CA: AAAI Press.
[11] Kahl, F., et al., (2001). Minimal Projective Reconstruction Including Missing Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, Vol. 23(4): 418-424.
[12] Lakshminarayan, K., et al., (1996). Imputation of Missing Data Using Machine Learning Techniques. *KDD-1996*: 140-145.
[13] Lakshminarayan K. et al. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, 11: 259-275.
[14] Ling, C.X., Yang, Q., Wang, J. & Zhang, S. (2004). Decision trees with minimal costs. *ACM International Conference Proceeding Series*, ICML 2004.
[15] Little, R.J.A. and Rubin, D.A. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.
[16] Little R. and Rubin D. (2002). *Statistical Analysis with Missing Data*. Wiley, 2002.
[17] Magnani, M. (2004). *Techniques for Dealing with Missing Data in Knowledge Discovery Tasks*, (available at http://magnanim.web.cs.unibo.it/index.html).
[18] Pawlak, M. (1993). Kernel classification rules from missing data*. IEEE Transactions on Information Theory*, 39(3): 979-988.
[19] Pearson R.K. (2006). The Problem of Disguised Missing Data. *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 1: 83 - 92.
[20] Pesonen, E., Eskelinen, M. and Juhola, M., (1998). Treatment of missing data values in a neural network based decision support system for acute abdominal pain. *Artificial Intelligence in Medicine*, 13(3): 139-146.
[21] Qin, Y.S., Zhang, S.C., Zhu, X.F., Zhang, J.L. and Zhang, C.Q. (2007). Semi-parametric Optimization for Missing Data Imputation. *Applied Intelligence*, 27(1): 79-88.
[22] Qin, Z.X. (2007). *Multiple costs and their combination in cost-sensitive learning*. PhD Thesis, University of Technology Sydney, 2007.
[23] Quinlan, J. (1989). Unknown Attribute values in Induction. In *Proc 6th Int workshop on machine learning*: Ithaca, pp 164-168.
[24] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
[25] Ramaswamy, S., Rastogi, R. & Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Data Sets. *Proceedings of the ACM SIGMOD Conference on Management of Data*, Dallas, TX, 427–438.
[26] Ramoni, M. and Sebastiani, P. (2001). Robust Learning with Missing Data. *Machine Learning*, 2001, Vol. 45(2): 147-170.
[27] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, Wiley: New York, 1987.
[28] Wilks, S. (1932). Moments and distributions of estimates of population parameters from fragments samples. *Ann. Math. Statist.* 3: 163-203.
[29] Wang, Q.H. and Rao, R.N.K. Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist*. 30: 896-924.
[30] Yuan Y.C. (2001). Multiple imputation for missing data: concepts and new development SAS/STAT 8.2. (Available at http://www.sas.com/statistics) SAS Institute Inc. Cary, NC.
[31] Zhang, SC, Qin, YS, Zhu, XF, Zhang, JL, and Zhang, CQ. (2006). Optimized Parameters for Missing Data Imputation. *PRICAI06*, 2006: 1010-1016.
[32] Zhang, S.C., Qin, Z.X., Sheng, S.L. and Ling, C.L. (2005). “missing is useful”: Missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17 No. 12: 1689-1693.
[33] Zhang, S.C., Qin, Y.S., Zhang, J.L., Zhu, X.F., Zhang, C.Q. (2008). Missing Value Imputation Based on Data Clustering. *Transactions on Computational Science Journal*, LNCS 4750, pp 128-138.

Shichao Zhang: Shichao Zhang is a professor and the dean of College of Computer Science and Information Technology at the Guangxi Normal University, Guilin, China. He holds a PhD degree in Computer Science from Deakin University, Australia. His research interests include data analysis and smart pattern discovery. He has published about 50 international journal papers, including 7 in IEEE/ACM Transactions, 2 in Information Systems, 6 in IEEE magazines; and over 40 international conference papers, including 3 AAAI, 2 ICML, 1 KDD, and 1 ICDM papers. He has won 6 China NSF/863/973 grants, 1 Overseas-Returning High-level Talent Research Program of China Hunan-Resource Ministry, 3 Australian large ARC grants. He is a senior member of the IEEE; a member of the ACM; and serving as an associate editor for IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, and IEEE Intelligent Informatics Bulletin.