A Review On Missing Value Estimation Using Imputation Algorithm

Roslan Armina1, Azlan Mohd Zain2, Nor Azizah Ali3 and Roselina Sallehuddin4

Soft Computing Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
E-mail: roslanarmina@gmail.com, azlanmz@utm.my, nzah@utm.my, roselina@utm.my

Abstract. The presence of the missing value in the data set has always been a major problem for precise prediction. The method for imputing missing value needs to minimize the effect of incomplete data sets for the prediction model. Many algorithms have been proposed for countermeasure of missing value problem. In this review, we provide a comprehensive analysis of existing imputation algorithm, focusing on the technique used and the implementation of global or local information of data sets for missing value estimation. In addition validation method for imputation result and way to measure the performance of imputation algorithm also described. The objective of this review is to highlight possible improvement on existing method and it is hoped that this review gives reader better understanding of imputation method trend.

1. Introduction

Data quality is considered important and concerned as big role in Machine learning and other correlated area Knowledge Discovery from databases. In Machine Learning, most of algorithm run the data mining straight from the data sets to extract knowledge and the quality of knowledge are depend on the quality of the data[1]

The presence of missing value is the common problem for all type research. Unfortunately, to acquire the actual data of missing value is often impossible and hence the data analyst must accept with reality that some data is unavailable. The missing value presence in unavoidable problem in many term because most of the data mining algorithm could not work with data sets that consist missing value[2]

2. The Imputation Technique Of Missing Value

Imputation method involve the replacement of missing value with some value that has been estimated that based on data mining of available information in the data set[2]. There are several methods available for handling the missing value which are Case Deletion (CD) method that consist the missing value determination on each of attribute and instance, Parameter Estimation method that use the application of procedures of maximum likelihood procedures for estimation of the parameters of a model defined for the complete data. Expectation-Maximization algorithm is applied in maximum likelihood procedures [3], and imputation method that constitute of class of procedures that solely aims for replacement of missing values with estimated values. The employment of known relationships that can be identified in the valid values of the data set to assist in estimating the missing values is the objective of this method. This papers focus on imputation of missing data.
2.1 Global Approach
In this global approach, Missing value replacement algorithm is based on the algorithm that apply on
global correlation information that originated from derivation of all data matrix. The existence of a
global covariance structure has been assume among all genes or samples in the expression matrix[4].
When this assumption is not appropriate, as example their imputation becomes less precise when the
genes exhibit dominant local similarity structures. The widely used algorithm for this imputation method
category is Bayesian principal component analysis (BPCA) [5] and Singular Vector Decomposition
imputation (SVDimpute) [6]. For the SVDimpute algorithm, the Singular Vector Decomposition is used
to obtain eigengenes that is a set of pattern of mutually orthogonal expression, which can be linearly
combined to approximate the expression of all genes in the data set. The Singular Vector Decomposition
algorithm can only be run on complete matrices[7].First, SVDimpute algorithm regresses the gene
against the most significant eigengenes k most significant and to reconstruct the missing values from a
linear combination of the k eigengenes ,coefficients of the regression is applied.

2.2 Local Approach
In this local approach, for missing value imputation application, the algorithms exploit only local
similarity structure in the data sets. To computes the missing value genes that contain the missing values
are used with the gene subsets that show high correlation. The example of local approach such as the K-
Nearest Neighbour imputation (KNNimpute) and local least square imputation (LLSimpute) are some of the
earliest and well-known algorithms [8]. The k-nearest neighbor imputation ( KNNimpute) [9]
uses pairwise information between the target gene with missing values and the k nearest reference genes
to impute the missing values. The research before have shown that KNNimpute imputation algorithm
performs extremely well when strong correlation exists between genes in the data sets . For imputes
missing values,Local Least Square imputation LLSimpute [10] apply a multiple regression model. This
technique has been proven to be slightly more competitive than KNNimpute and is more complex than
BPCA.

SLLSimpute sequentially performs imputation by starting from the least missing gene with least missing
rates. Sequential LLSimpute (SLLSimpute) [11] is an extension of LLSimpute algorithm and the gene
that has been imputed are then reused for imputation of other genes. It has been proven that SLLSimpute
performs better than LLSimpute because the genes with missing values are reusable in this algorithm.
Another method that widely used in local approach is the MICE-CART, which is the combination of
multiple imputation of classification and regression trees (CART) and chained equations (MICE). It is
a nonparametric approach done by [12] to perform multiple imputations through chained equations using
sequential regression trees as the conditional models. MICE-CART is applied for parameter usage
reduction and this method is able to perform tuning while capturing complex relations of the data. Other
local approach is GMCimpute [13] that is imputation that perform Gaussian mixture clustering. Even
though GMCimpute is a local approach algorithm ,it is able to using more global correlation information.
In this algorithm, by using Expected Maximization , the data is clustered into estimated missing values
components of Gaussian mixtures. A single estimated missing value is computed from each component
and then to obtain the final estimated missing values, the value is averaged. Local correlation
information in the data is use for GMCimpute through the components mixture. The concept of multiple
parallel estimations of missing values are utilized for collateral missing value imputation (CMVE)
technique for improvement of final estimation. Compared to BPCA,KNN and LLSimpute accuracy in
normalized RMS error(NMRSE) on many datasets, CMVE [15] has been able to produce better
accuracy. This calculation are involving data sets ovarian cancer and yeast sporulation time series data.
Ameliorative missing value imputation (AMVI) [16] apply simulation of Monte Carlo for determination
of optimal number of reference gene, therefore AMVI is more advanced than CMVE. The time series
expression profiles used in AMVI have been reported to exhibit strong dependency between
observations. Adaptive Bicluster-Based Approach or ABBA is a missing value estimator for binary
matrices. This algorithm has been made less complex for better understanding and usage, but the amount of parameters tweaking is also higher. [17] verified when the rate of missing values are higher than normal the performance of this algorithm is better than KNN.

2.3 Hybrid Approach

Local correlation between genes dominates and techniques such as LLSimpute or KN Nimpute perform better compared to BPCA or SVDimpute especially for heterogeneous data sets. This statement prove that the performance of the imputation techniques is affected by correlation structure in the data. Missing value imputation approach such as BPCA or SVDimpute are preferred in term of global correlation and the accuracy of the predicted missing value increase as it will affect the conclusion to the case related to the dataset.

Nevertheless, there are still some hybrid approach for missing value imputation like HPM-MI[18] that significantly improved data quality by use of best imputation technique after quantitative analysis of eleven imputation approaches. HPM-MI combine the K-mean clustering with Multilayer Perceptron and use 11 different missing value imputation techniques and selects the best (compact) clusters among the results. K-means clustering is also used to validate class labels of given data.

Another notable approach is the GA+SVR[19] that reduce the error between the model output and the input. It is also to tackle the issue of difficulty in obtaining optimal value for the vigilance parameters. This approach apply Genetic Algorithm is used to optimize the selections of attributes and Support Vector Regression predicts the decision attributes value. Other hybrid approach is Fuzzy C-means+SVR+GA. This Hybrid use for overcome the issue of no theoretical basis for the selection of the cluster (c) number. It is also to tackle the issue of difficulty in obtaining optimal value for the vigilance parameters.

Knn + Neural Networks[20] is develop to overcome the issue of inadequate input value because before this most of the techniques discovered data as inputs and estimate an output class. This model fails if one or more inputs are missing. Consequently, they cannot be used for decision-making purposes if the data variables are not complete.

Fuzzy C-means based imputation + genetic algorithm[21] that based on inductance loop detector outputs. By utilizing the weekly similarity among data, the conventional vector-based data structure is firstly transformed into the matrix-based data pattern. Then, the GA is applied to optimize the membership functions and centroids in the FCM model.

MIGEC[22] is approach of imputation that create a method that minimizing the bias iteratively and less requirement on prior knowledge to the distributions that estimate missing values by multiple operations based on clustering. The non-missing data instances are separated into several clusters. Then, the imputed value is obtained after multiple calculations by utilizing the information entropy of the proximal category for each incomplete instance in terms of the similarity metric based on Gray System Theory (GST).

Multi-layered artificial immune system and genetic algorithm hybrid[23] is for effectively partial impute missing data in high-dimensional spaces. The multi-layered artificial immune system creates and stores antibodies that bind to and annihilate an antigen. The genetic algorithm optimises the learning process of a stimulated antibody. The evaluation of the imputation is performed using the RIPPER, k-nearest neighbour, naïve Bayes and logistic discriminant classifiers.

HHONC[24] hybrid approach that solve the issue of bias if there are systematic differences between the observed and the unobserved data by uses a higher correlation of input neurons for better fitting properties which often leads to a higher number of learning parameters (weights). Then It finds the closest k-neighbours from the unknown instance and returns the average value of the real-valued labels of the neighbours. The distance of the attribute is zero when the two instances miss the values of the same attribute, but when only one has a missing value, a maximal distance is assigned.
HPF[25] hybrid approach is develop to avoid intervals determined by different cluster information. The hybrid algorithm utilizes the global optimization capability of particle swarm to search the imputations of missing attributes and the cluster prototypes and incorporates the gradient-based FCM to calculate membership information, which is helpful in improving the clustering performance and the computational efficiency.

Other hybrid approach for imputation missing value is AR-ANN[26] that develop for imputing missing values and jointly handling the nonlinearity problem. AR model was used only for determining the input layer structure of ANN.

ANN-CBR[27] hybrid approach is developed to integrate artificial neural network with case-based reasoning. The hybrid algorithm utilizes ANN’s strength that focus on feature extraction and pattern recognition and feature from CBR that identify same cases despite missing value.

FCM-MLP[28] is hybrid imputation approach that utilize fuzzy c-means and multilayer perceptron. Fuzzy c-means clustering is used as 1st stage and cluster’s center for each attribute is generated for complete datasets. For 2nd stage, MLP is utilized for network modelling.

PCA-NF+GA[29] is hybrid imputation approach that integrate principal component analysis and neural fuzzy with genetic algorithm. The data sets are orthogonalize by PCA to ensure the model is better trained then the compressed data is trained by neuro fuzzy model follow by GA to generate missing value and error between missing value and generated by GA is used as evaluation for comparison. If the error is not minimal, new missing values are generated by GA until error is minimized.

RF-AANN-GA[30] is hybrid based imputation that integrates Random forest and Auto-associative Neural Network Hybrid with genetic algorithm. For prediction of missing values in data sets, random forest are utilized the prediction of missing value is recorded. The prediction missing value by RF are used as limit for the search space for GA in AANN-GA system.

AANN-GA-RF[30] is hybrid imputation approach that combine AANN-GA system in cascade with Random Forest. Data are divided into four sets of training, validation, testing and experimental to generate artificially generate incomplete data sets then propagated through the AANN-GA for obtaining missing value. The imputed data and testing data are made to form complete datasets.

K-mean+MLP[31] is hybrid imputation approach that involving local learning and global approximation. In stage 1, missing value replaced by using k-mean algorithm then refine the resultant approximant values by using multilayer perceptron. The complete data set is trained by MLP with attribute having missing values as target variable one at a time.

ANN-NN[32] is develop based on observed hydrological data in Malaysia. It utilize the principle of Artificial Neural network combined with Nearest Neighbor. The proposed model is able to produce pattern of imputed data which closely resembles that of the observed values where peaks of the observed hydrographs are mostly attainable proves that the proposed model is able to capture daily maximum rainfall series.

DMI[33] is hybrid imputation approach that utilize decision tree and expectation maximization. (DMI) applies EMI on various horizontal segments (of a data set) where correlations among attributes are high. EM based imputation techniques rely on the correlations among the attributes in a data set.

2.4 Knowledge Assisted Approach

This knowledge assisted approach is the missing values imputation process that have integration of external information or domain knowledge. This approach is significantly increase the imputation’s accuracy using domain knowledge especially for data sets with small number of samples which are noisy or have high missing rate.

One of the example of knowledge assisted approach is the Projection Onto Convex Set (POCS) [34] that exploits the biological occurrence of synchronization loss and correlation information between genes and arrays, which is known as flexible set theoretic framework. POCS executes local least square regression to performs PCA imputation to capture array-wise correlation, capture gene-wise correlation
and restricts the squared power of the expressions profiles for synchronization loss capturing. Whether local or global correlation structure prevails in the, by using POCS the optimal solution can be achieved in the data. This is because final solution is always dominated by the smallest yet most reliable constraint set to satisfy larger yet less reliable constraint sets. Gene ontology (GO) has three independent ontologies that explain gene products in terms of associated biological processes (BP), cellular components, and molecular functions (MF) [35] and also known as a well-accepted standard for gene function categorization and GO increase the accuracy of the imputation as proven in experimental results where the proportion of annotated genes was large at higher rates of missing values. By combines histone acetylation information into KNImpute and LLSimpute, Histone Acetylation Information Aided Imputation (HAIimpute) is developed. HAIimpute develop to increase of missing value estimation accuracy [36]. To form the pattern expressions, HAIimpute used the mean expression of genes from each of the clusters. By obtained by fitting a linear regression model between the gene and pattern expressions, the missing values are obtained. The final missing values estimation given by a convex combination of linear regression imputations and secondary imputation using both KNImpute and LLSimpute. Indicated by improved correlation between imputed genes and original complete genes, it has also been proven that HAIimpute has consistently improved the KNImpute or LLSimpute. The Table 1 below shows the grouping of imputation algorithms considered in this review.

| Title | Algorithm | Year | Class | Reference | Remark |
|-------|-----------|------|-------|-----------|--------|
| Missing value estimation methods for DNA microarrays | SVDImpute | 2001 | Global | 5 | Singular Vector Decomposition is used to obtain eigengenes that is a set of pattern of mutually orthogonal expression |
| A Bayesian missing value estimation method for gene expression profile data | BPCA | 2003 | Global | 6 | Estimate arbitrary missing variables is new in terms of statistical methodology. |
| Improving cluster-based missing value estimation of DNA microarray data | MICE-CART | 2010 | Local | 12 | combination of multiple imputation of classification and regression trees (CART) and chained equations (MICE) |
| Missing value imputation for gene expression data: computational techniques to recover missing data from available information | KNImpute | 2001 | Local | 8 | uses pairwise information between the target gene with missing values and the k nearest reference genes to impute the missing values |
| Gaussian mixture clustering and imputation of microarray data | GMCimpute | 2004 | Local | 13 | perform Gaussian mixture clustering |
| Method                                                                 | Year | Type  | Number | Description                                                                                                                                 |
|-----------------------------------------------------------------------|------|-------|--------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Missing value estimation for DNA microarray gene expression data: local least squares imputation | 2005 | Local | 10     | Apply a multiple regression model                                                                                                         |
| Sequential local least squares imputation estimating missing value of microarray data | 2008 | Local | 11     | Performs imputation by starting from the least missing gene with least missing rates                                                     |
| Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data | 2005 | Local | 15     | Apply the concept of multiple parallel estimations of missing values are utilized for collateral missing value imputation                    |
| Ameliorative missing value imputation for robust biological knowledge inference | 2008 | Local | 16     | Apply simulation of Monte Carlo for determination of optimal number of reference gene                                                      |
| ABBA: adaptive bicluster-based approach to impute missing values in binary matrices | 2010 | Local | 17     | Missing value estimator for binary matrices                                                                                               |
| Hybrid prediction model with missing value imputation for medical data | 2015 | Hybrid | 18     | Use the best imputation technique after quantitative analysis of eleven imputation approaches                                             |
| A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm | 2013 | Hybrid | 19     | Tackle the issue of difficulty in obtaining optimal value for the vigilance parameters                                                     |
| A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm | 2013 | Hybrid | 19     | Use for overcome the issue of no theoretical basis for the selection of the cluster (c) number                                             |
| Method                                                                 | Year | Type   | Iterations | Description                                                                                                                                 |
|-----------------------------------------------------------------------|------|--------|------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks | KNN+NN | Hybrid | 20         | overcome the issue of inadequate input value because before this most of the techniques discovered data as inputs and estimate an output class |
| A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation | FCM+GA | Hybrid | 21         | based on inductance loop detector outputs                                                                                                                                                           |
| Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering | MIGEC | Hybrid | 22         | create a method that minimizing the bias iteratively and less requirement on prior knowledge to the distributions that estimate missing values by multiple operations based on clustering | 
| Partial imputation of unseen records to improve classification using a hybrid multi-layered artificial immune system and genetic algorithm | MAIS+GA | Hybrid | 23         | for effectively partial impute missing data in high-dimensional spaces                                                                                                                               |
| Missing Value Imputation Using Hybrid Higher Order Neural Classifier   | HHONC | Hybrid | 24         | solve the issue of bias if there are systematic differences between the observed and the unobserved data                                                                                               |
| A Global Clustering Approach Using Hybrid Optimization for Incomplete Data Based on Interval Reconstruction of Missing Value | HPF  | Hybrid | 25         | utilizes the global optimization capability of particle swarm                                                                                                                                       |
| Imputation of Missing Values in Daily Wind Speed Data Using Hybrid AR-ANN Method | AR-ANN | Hybrid | 26         | imputing missing values and jointly handling the nonlinearity problem                                                                                                                               |
| Imputation of missing values by integrating neural networks and case-based reasoning | ANN-CBR | Hybrid | 27         | integrate artificial neural network with case-based reasoning                                                                                                                                      |
| Description                                                                 | Method          | Year | Type    | Value | Note                                                                 |
|-----------------------------------------------------------------------------|-----------------|------|---------|-------|----------------------------------------------------------------------|
| Hybrid model for data imputation: using fuzzy c means and multi layer perceptron | FCM-MLP         | 2014 | hybrid  | 28    | utilize fuzzy c-means and multilayer perceptron                      |
| Imputation of missing data using PCA, neuro-fuzzy and genetic algorithms    | PCA-NF+GA       | 2008 | Hybrid  | 29    | integrate principal component analysis and neural fuzzy with genetic algorithm |
| Missing data imputation through the use of the Random Forest Algorithm      | RF-AANN-GA      | 2009 | Hybrid  | 30    | integrates Random forest and Auto-associative Neural Network Hybrid with genetic algorithm |
| Missing data imputation through the use of the Random Forest Algorithm      | AANN-GA-RF      | 2009 | Hybrid  | 30    | combine AANN-GA system in cascade with Random Forest                 |
| A novel soft computing hybrid for data imputation                           | K-mean+MLP      | 2011 | Hybrid  | 31    | involving local learning and global approximation                    |
| Imputation of time series data via Kohonen self organizing maps in the presence of missing data | ANN-NN          | 2008 | Hybrid  | 32    | utilize the principle of Artificial Neural network combined with Nearest Neighbor |
| A decision tree-based missing value imputation technique for data pre-processing | DMI             | 2011 | Hybrid  | 33    | Utilize decision tree and expectation maximization. (DMI) applies EMI on various horizontal segments (of a data set) |
| Microarray missing data imputation based on a set theoretic framework and biological knowledge | POCSimpute      | 2006 | Knowledge | 34    | exploits the biological occurrence of synchronization loss and correlation information between genes and arrays |
| Improving missing value estimation in microarray data with gene ontology    | GOimpute        | 2006 | Knowledge | 35    | increase the accuracy of the imputation as proven in experimental results where the proportion of annotated genes was large at higher rates of missing values |
| Missing value imputation for microarray gene expression data using histone acetylation information | HAimpute        | 2008 | Knowledge | 36    | used the mean expression of genes from each of the clusters          |
3. Imputation Result Validation

One of the important steps in demonstrating the reliability and accuracy of imputation algorithms is the validation of imputation results. Firstly, validation methods can be categorized into two: external and internal validation. Validation of imputation methods is done by using internal validation that uses only information in the data sets. Internal validation is done by computing performance indices of imputed data with original data [37]. External validation is done by assessing the imputation effect on subsequent biological analysis. External validation applies external knowledge rather than information’s in the data sets.

3.1 Internal Validation

The most widely used method for analyzing the accuracy of missing value imputation algorithms is to compute normalized root mean square error (NRMSE). The accuracy of the missing value imputation algorithm is determined by the lower the NMRSE, the accuracy is higher [37]. The equation of NMRSE is as follows:

\[
NMRSE = \sqrt{\frac{\sum_{i=1}^{m} \sum_{k=1}^{n}(g_{ik} - \hat{g}_{ik})^2}{\sum_{i=1}^{m} \sum_{k=1}^{n}(g_{ik})^2}}
\]  

(1)

Where \(g_{ik}\) represent the \(k\)th experiment for dataset value \(g_i\), and \(g\) and \(\hat{g}\) represent the true value and imputed value respectively.

3.1 External Validation

The validation of missing value imputation algorithms by using external validation is by examining the effect of external information such as functional annotation of pathway information to downstream biological analysis. In microarrays experiments, the objective is to detect gene sets with similar functional roles and when clustered, GO term that are significant are searched among gene of each cluster and one way to analyze the performance of missing value imputation algorithm is by checking for significant enrichment of GO term in gene cluster[37]. The calculation of the enrichment \(P\)-value for each GO term, \(t\) and for each cluster is as follows:

\[
p = \sum_{i=k}^{\min(b,T)} \frac{(\frac{t}{b} - \frac{b}{T})}{\frac{b}{T}}
\]

(2)

Where \(b\)=number of gene in cluster,
\(K\)=number of gene in cluster with GO term(\(t\))
\(B\)=number of gene in the dataset

4. Limitation

Each algorithm has its own advantages and disadvantages, so does the datasets being used for each missing values technique. Several studies have shown that the performance of missing values imputation algorithms is significantly affected by factors such as correlation structure in the data, the missing data mechanism, the distribution of missing entries in the data, and the percentage of missing values in the data [2]. Selecting the right algorithm may significantly boost the accuracy of imputation results since there is no single imputation algorithm that performs the best in every situation. Global methods such as SVDimpute and BPCA perform better on data sets with low entropy, whereas local methods such as LLSimpute and KNNimpute perform better with high entropy data sets.

Another limitation is related to assumptions made on the distribution of missing values. Missing values in microarray data sets are normally assumed to be missing at random in most studies on missing values imputation, even though this is not a realistic assumption as missing values tend to arise in a
systematic manner in practice. This pattern can significantly affect the imputation’s performance and thus need to be properly considered in data analysis or algorithm design [12]. It also depends on the experimental conditions across the columns as variations in conditions may result in a non-random distribution of missing values in the data matrix. This is because each column in a microarray data matrix comes from different experiments.

Another limitation is testing and improving the model for multi-class imbalanced classification problems. Imbalanced classification problem arises due to imbalanced distribution of instances among multiple classes present in the data set. To further improve the imputation accuracy, feature selection and/or dimension reduction techniques can be applied to the training data; for example, principal component analysis methods can be applied before the support vector training to reduce the training time.

5. Conclusion
In large-scale experiment analyses and studies, missing values have become a frequent problem in data mining and this affects the result or conclusion being made. Since many analyses require complete data set, missing values imputation is an essential pre-processing step in data analysis. Instead of introducing new algorithms, variations based on existing imputation approaches are worthy contributions to systematic evaluation of existing algorithms. With so many different algorithms, evaluation of suitable parameters and operating platform has to be performed for optimal execution of the algorithm. This is to avoid overwhelming development of new missing value imputation algorithms before the full functional capability of previous existing ones has been fully explored.

References
[1] Batista, Gustavo EAPA, and Maria Carolina Monard. "A Study of K-Nearest Neighbour as an Imputation Method." HIS 87.251-260 (2002): 48.
[2] Gautam, Chandan, and Vadlamani Ravi. "Data imputation via evolutionary computation, clustering and a neural network." Neurocomputing 156 (2015): 134-142.
[3] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the royal statistical society. Series B (methodological) (1977): 1-38.
[4] Liew, Alan Wee-Chung, Ngai-Fong Law, and Hong Yan. "Missing value imputation for gene expression data: computational techniques to recover missing data from available information." Briefings in bioinformatics 12.5 (2011): 498-513.
[5] Oba S, Sato MA, Takemasa I, etal. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 2003;19(16):2088–96.
[6] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–5.
[7] Troyanskaya, Olga, et al. "Missing value estimation methods for DNA microarrays." Bioinformatics 17.6 (2001): 520-525.
[8] Liew, Alan Wee-Chung, Ngai-Fong Law, and Hong Yan. "Missing value imputation for gene expression data: computational techniques to recover missing data from available information." Briefings in bioinformatics 12.5 (2011): 498-513.
[9] Troyanskaya, Olga, et al. "Missing value estimation methods for DNA microarrays ." Bioinformatics 17.6 (2001): 520-525.
[10] Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 2005; 21: 187-198.
[11] Zhang X, Song X, Wang H, et al. Sequential local least squares imputation estimating missing value of microarray data. Comp Biol Med 2008;38:1112–20.
[12] Bras LP, Menezes JC. Improving cluster-based missing value estimation of DNA microarray data. Bioinformatics Eng 2007; 24:273–82.

[13] Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. Bioinformatics 2004;20(6):917–23.

[14] Cai Z, Heydari M, Lin G. Iterated local least squares microarray missing value imputation. J Bioinfom Comput Biol 2006;4(5):935–57.

[15] Sehgal MSB, Gondal I, Dooley LS. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. Bioinformatics 2005; 21: 2417-2423.

[16] Sehgal MSB, Gondal I, Dooley LS, Coppel R. Ameliorative missing value imputation for robust biological knowledge inference. J Biomed Inform 2008; 41: 499-514.

[17] Colantonio A, Pietro RD, Ocello A, Verde NV. ABBA: adaptive bicluster-based approach to impute missing values in binary matrices. Proceedings of the 2010 ACM Symposium on Applied Computing; Sierre, Switzerland. 1774304: ACM; 2010. p. 1026-33.

[18] Purwar, Archana, and Sandeep Kumar Singh. "Hybrid prediction model with missing value imputation for medical data." Expert Systems with Applications 42.13 (2015): 5621-5631.

[19] Aydilek, Ibrahim Berkan, and Ahmet Arslan. "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm." Information Sciences 233 (2013): 25-35.

[20] Aydilek, Ibrahim Berkan, and Ahmet Arslan. "A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks." International Journal of Innovative Computing, Information and Control 7.8 (2012): 4705-4717.

[21] Tang, Jinjun, et al. "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation." Transportation Research Part C: Emerging Technologies 51 (2015): 29-40.

[22] Tian, Jing, et al. "Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering." Applied intelligence 42.2 (2014): 376-388.

[23] Duma, Mlungisi, et al. "Partial imputation of unseen records to improve classification using a hybrid multi-layered artificial immune system and genetic algorithm." Applied Soft Computing 13.12 (2013): 4461-4480.

[24] Panigrahi, Lipismita, Kaberi Das, and Debahuti Mishra. "Missing Value Imputation Using Hybrid Higher Order Neural Classifier." Indian Journal of Science and Technology 7.12 (2014): 2007-2014.

[25] Zhang, Liyong, et al. "A Global Clustering Approach Using Hybrid Optimization for Incomplete Data Based on Interval Reconstruction of Missing Value." International Journal of Intelligent Systems (2015).

[26] Shukur, Osamah Basheer, and Muhammad Hisyam Lee. "Imputation of Missing Values in Daily Wind Speed Data Using Hybrid AR-ANN Method." Modern Applied Science 9.11 (2015): 1.

[27] Ennett, Colleen M., Monique Frize, and Robin C. Walker. "Imputation of missing values by integrating neural networks and case-based reasoning." Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE. IEEE, 2008.

[28] Azim, Shambeel, and Suhas Aggarwal. "Hybrid model for data imputation: using fuzzy c means and multi layer perceptron." Advance Computing Conference (IACC), 2014 IEEE International. IEEE, 2014.

[29] Hlalele, Nthabiseng, Fulufhelo Nelwamondo, and Tshilidzi Marwala. "Imputation of missing data using PCA, neuro-fuzzy and genetic algorithms." Advances in neuro-information processing. Springer Berlin Heidelberg, 2008. 485-492.

[30] Pantanowitz, Adam, and Tshilidzi Marwala. "Missing data imputation through the use of the Random Forest Algorithm." Advances in Computational Intelligence. Springer Berlin Heidelberg, 2009. 53-62.

[31] Ankaiah, Naravula, and Vadlamani Ravi. "A novel soft computing hybrid for data imputation." Proceedings of the 7th international conference on data mining (DMIN). 2011.
[32] Malek, M. A., et al. "Imputation of time series data via Kohonen self organizing maps in the presence of missing data." Engineering and Technology 41 (2008): 501-506.
[33] Rahman, Geaur, and Zahidul Islam. "A decision tree-based missing value imputation technique for data pre-processing." Proceedings of the Ninth Australasian Data Mining Conference Volume 121. Australian Computer Society, Inc., 2011.
[34] Gan, Xiangchao, Alan Wee-Chung Liew, and Hong Yan. "Microarray missing data imputation based on a set theoretic framework and biological knowledge." Nucleic Acids Research 34.5 (2006): 1608-1619.
[35] Tuikkala, Johannes, et al. "Improving missing value estimation in microarray data with gene ontology." Bioinformatics 22.5 (2006): 566-572.
[36] Xiang, Qian, et al. "Missing value imputation for microarray gene expression data using histone acetylation information." BMC bioinformatics 9.1 (2008): 252.
[37] Moorthy, Kohbalan, Mohd Saberi Mohamad, and Safaai Deris. "A review on missing value imputation algorithms for microarray gene expression data." Current Bioinformatics 9.1 (2014): 18-22.