An Evaluation of Machine Learning Algorithms for Missing Values Imputation

KohbalanMoorthy, Mohammed HasanAli, MohdArfianIsmail, Chan Weng Howe, MohdSaberiMohamad, SafaaiDeris

Abstract—In gene expression studies missing values have been a common problem. It has an important consequence on the explanation of the final data. Numerous Bioinformatics examination tools that are used for cancer prediction includes the dataset matrix. Hence, it is necessary to resolve this problem of missing values imputation. Our research paper presents a review of missing values imputation approaches. It represents the research and imputation of missing values in gene expression data. By using the local or global correlation of the data we focus mostly on the contrast of the algorithms. We considered the algorithms in a global, hybrid, local, and knowledge-based technique. Additionally, we presented the different approaches with a suitable assessment. The purpose of our review article is to focus on the developments of current techniques. For scientists rather applying different or newly develop algorithms with the identical functional goal. We want an adaptation of algorithms to the characteristics of the data .

Keywords: Missing Value Imputation, Gene Expression Data, Microarray Data, Cancer Informatics, Computational Intelligence

I. INTRODUCTION

The microarray method has been an essential tool used by several scientists to study the appearance of different genes in a specific organism [3]. Microarray technology is important in the field of genetics because the micro dimension of the chips used in this technology [1] can contain a large number of genes necessary for wide gene expression studies [2]. Microarray technology agrees to sample information development, where a full statistical understanding can be helpful in gene regulation and expression detection. Microarray technology is used in cancer studies for appropriate gene detection and analysis. Furthermore, the effects of cancer drugs can also be detected using this technology [4]. Microarray technology has also found applications in different fields, such as microbiology, virology, and immunology [5][6]. Data investigation is a tedious work Bioinformatics microarray because the dataset contains uncharacterized variable that needs interpretation as well [7]. During the analysis of gene expression, the problem of missing values (MVs) is always encountered; however, MV is considered inconsequential if the rate is < 1%, and controllable when it is between 1-5%. At 5-15%, complex systems are required to handle the imputation and when >15%, the estimation will be strongly affected. Several explanations have been given for the occurrence of MVs and most of them involve the presence of artifacts, hybridization failures, and low-resolution on the microarray [8].

Missing values in a dataset have been shown to severely affect analysis, as well as having a great effect on studies such as genes classification (supervised & unsupervised), detection of differently expressed genes, as well as establishment of gene regulation networks [9]. Additionally, complete dataset with no missing value and low noise is required in most feature selection algorithms for the selection of the target features during feature selection model construction [10]. The performance of algorithms such as SVM, SVD, and PCA is strongly affected by MVs in datasets [11].With the recent advancement in algorithms, the issue of MV can be solved via MVs approximation (MVA) such that the important genes can be preserved. This is important in the initial detection of specific genes for specific classes [12]. Thus, MVA seems to be an alternative low-cost choice. The estimation of MVs in data can be done via computational techniques rather than repeating the whole microarray experiments. As per previous reports, the level of MVs can be increased during inclusive detection or data scrutiny [13]. In the biomedical field, several mathematical models have been proposed and developed [14] for use in different fields of biomedical research [14]. The estimation of the MVs percentage requires a complete dataset with no missing value. Sometimes, the MVs in gene expression data is estimated using incorrect values and this has a serious influence on disease-related genes detection. Thus, there is a high chance of discarding useful genes, especially when there is disturbance in ranking of significant genes [15]. Being that microarray experiments usually involve gene expression data that are newly identified, repeating of the microarray investigation may not provide the correct expectation.

Many cancer patient data exists for the building of cancer prediction algorithms; however, they still depend on pre-processed raw data with MVs using ancient techniques and always reliant on a k-nearest neighbor [40].

Revised Manuscript Received on September 14, 2019.

Kohbalan Moorthy, Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Kuantan, Malaysia
(E-Mail: kohbalan@ump.edu.my)

Mohammed Hasan Ali, Computer Techniques Engineering Department, Faculty of Information Technology, Imam Ja'afar Al-sadiq University, Najaf, Iraq

Mohd Arfian Ismail, Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Kuantan, Malaysia

Chan Weng Howe, Artificial Intelligence and Bioinformatics Group (IAIBG), Faculty of computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

Mohd Saberi Mohamad, Artificial Intelligence and Bioinformatics Group (IAIBG), Faculty of computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

Safaai Deris, Artificial Intelligence and Bioinformatics Group (IAIBG), Faculty of computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

Gene expression profiles replication allows development in the preprocessing stage and this has been the basis for cancer estimation. In the initial methods, rows that contain MV are eliminated, imputing MV with median or regular row values and changes the MVs with zero [16]. It is not appropriate to fill in the MV spots with average values or 0 because it introduces bias since the association of the data is not considered.

The main goal of this work is to provide an overview of the most recent MVI approaches based on gene expression data. Moreover, this work provides analysis of different algorithms as considered from different perspectives, such as global, hybrid, local, and knowledge-based techniques. The main purpose of this review is to help researchers by providing analysis of the current techniques and algorithms with respect to data characteristics. This work is organized as follows; Section 2 provides an overview of MV imputation algorithms. The evaluation of algorithms in terms of performance is provided in Section 3 while the summary of the limitations is provided in Section 4. Section 5 presents the conclusion..

II. REVIEW MISSING VALUES IMPUTATION ALGORITHMS

The software in the initial MVI techniques works in microarray data imputation, with the data based on statistical scrutiny while data's natural evidence is not considered. However, imputation procedures in the current imputation methods are designed based on information sourced from microarray data [17]. The missing value imputation (MVI) approaches are categorized into three, which are parameter estimation technique, pairwise deletion technique, and imputation techniques [18]. The issue of MVs can be resolved using four different [19], including KNN imputation (KNNI), mean imputation (MI), median imputation (MDI), and case deletion (CD). Regarding CD, it works by eliminating all instances that have MVs. It exchanges the MV of any instance by computing the mean of all the known values of that attribute. The MDI uses the median of the attribute values to determine the consistency rather than the mean due to the influence of outliers on the mean. For the KNNI method, a distance function is used to determine the similarity of 2 instances; the MV is imputed by establishing the furthermost illustrations parallel to the instance of interest [20]. The MVI approaches are classified into 2 basic types as "generic statistical (GS) approaches & application-particular alterations". The GS technique includes methods such as mean imputation hot deck imputation, model-based imputation, as well as multiple & cold deck imputation. For explanation of gene expression data imputation, quality matters are normally considered. The existing imputation algorithms are categorized into 4 types based on the information type used. These categories include global, local, hybrid, & knowledge-based techniques. Table 1 presents a list of the categories of imputation algorithms.

Table 1. List of Missing Values Imputation Based on the Categories

Algorithms	Year	Category
SVDimpute	2001	Global
BPCA	2003	Global
MICE-CART	2010	Local
KNNimpute	2001	Local
GMCimpute	2004	Local
LLSimpute	2005	Local
SLLSimpute	2008	Local
CMVE	2005	Local
AMVI	2008	Local
ABBA	2010	Local
LinCmb	2005	Hybrid
HPM-MI	2015	Hybrid
FCM+GA	2015	Hybrid
HPF	2015	Hybrid
AR-ANN	2015	Hybrid
POCSimpute	2006	Knowledge
GOimpute	2006	Knowledge
HAIimpute	2008	Knowledge

2.1. Global Approache

In the following segment, the algorithms execute missing values approximation founded on global correlation material resulting by the whole data matrix. The imputation becomes less accurate when the algorithms take up the existence of a global covariance organization in all genes, samples and also if the genes show identical structures. Global approach based algorithms are singular value decomposition (SVD) [21]and Bayesian principal component analysis [22].

2.2. Local Approach

In this approach, within the data set the algorithms shows the only local correspondence arrangement for computing missing values imputation. The subgroups of genes that show a high correlation through the genes are used to calculate the missing values in the gene. The well-known and initial algorithms in this procedure are local least square imputation (LL Simpute) and K nearest neighbor (KNN). MICE-CART is familiar as multiple imputations by chained equations (MICE) and classification and regression trees (CART) is a nonparametric method done by [23]. One of the related work has decided to perform multiple imputations through chained equations using sequential regression trees as the conditional models[24]. While capturing complex relations of the data it is applied to decrease the practice of limitation and imputing tuning. Among the target gene with missing values and the k nearest reference genes to impute the missing values, KNN imputation uses pairwise information.

It is recognized that KNN impute achieves tremendously well when robust correlation founds between genes in the data. Gaussian mixture clustering (GMC) is capable to practice extra global correlation information even though it is considered as local approach algorithm [25]. In this algorithm, the data is clustered into S components. Gaussian mixtures using the S estimates of the missing values and EM

Retrieval Number: L108110812S219/2019©BEIESP
DOI: 10.35940/ijitee.L1081.10812S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

algorithm are component that obtains the final estimated missing values. In the data, GMC impute utilizes the local correlation information through a mixture of components. L Local least squares imputation practices multiple regression models to impute missing values[26]. This method has been recognized to be marginally competitive associated with KNN impute and considerable further difficult than BPCA. Sequential LL Simpute (SLL Simpute) is an extension from LL[27].Simpute algorithm where this method implements imputation sequentially by establishment from the gene with minimum missing rates and the imputed genes are then recycled for imputation of other genes. Due to the fact of the reusability of the genes with missing values SLLS impute shows better performance than LL Simpute. To improve the final estimation collateral missing value imputation (CMVE) technique uses the idea of numerous similar estimations of missing values [28].On many datasets involving ovarian cancer and yeast sporulation time series data CMVE[28] . has been able to produce better accuracy in normalized RMS error (NRMSE) related to BPCA, KNN and L Simpute. By using Monte Carlo simulation for the determination of ideal number reference genes K Ameliorative missing value imputation (AMVI) has enhanced over CMVE [29]. Strong dependency among observations is displayed by time-series expression profiles. For binary matrices, Adaptive Bicluster-Based Approach or ABBA is a missing values estimator[30]. For better understanding and practice the complication of the algorithm itself has been decreased yet the amount of parameters alteration that can be achieved is lifted. However, when the rate of missing values is much higher than normal the algorithm has been verified by [21] and shows much better than KNN. **2.3. Hybrid Approach** For heterogeneous data sets, the local correlation among genes are dominant and local imputation methods such as KNN impute or LL Simpute performs better compared to BPCA or SVD impute. This displays that the correlation structure in the data affects the performance of the imputation systems. Although for homogeneous data, global methods such as BPCA or SVD impute would implement improved by capturing global correlation material in the data. There are numerous hybrid methods for missing values imputation like HPM-MI. By using best imputation methods, it considerably enhanced data quality. It uses 11 different missing values imputation methods. HPM-MI combines the K-mean clustering with Multilayer Perceptron and selects the best clusters among the results. Class labels of given data are validated by using K-means clustering. This is a hybrid prediction model for medical data. After extensive examination of eleven imputation methods uses the best imputation technique [28].For the expression of genes, we use specific linear topology. This topology has a state of transition and self -transition to the next state. At the start of each chain, we have a special start state. We interpret the results as follows expression close to background level is shown by expression value near zero, the value above zero shows overexpression or up-regulation and below zero value shows under expression or down-regulation. For example, a linear HMM with two emitting states the first one with a zero mean emission and the second is with a one mean emission, typically shows up-regulation prototypical behavior [31]. FCM+GA are an abbreviation of Fuzzy C-means based imputation + genetic algorithm. This is based on inductance loop detector outputs. Using the weekly resemblance among data vector-based data structure transformed into the matrix-based data pattern. Hence the genetic algorithm is used to optimize the membership functions and play a central role in the FC model. The fuzzy C is a hybrid method. Fuzzy C imputation method is combined with the genetic algorithm to estimate the missing values in data [32]. This method is mostly used in pattern recognition. This method has the advantage of giving the best modeling results. The scientists fix weighting exponent (m) to a conventional value of two which is not suitable for all applications.The iterative search approach used to get an optimal number of clusters, find the optimal single-output Sugeno type. Fuzzy Inference System (FIS) model is used to gives a minimum least square error by improving the limits of the subtractive clustering algorithm among the actual data and the Sugeno fuzzy model. The two methods are proposed when the number of clusters was optimized weighting exponent (m). These two approaches are namely, the iterative search approach and genetic algorithms. These methods were tested on the data from the original function. The fuzzy models are found with least error among the real data and fuzzy model [32]. FCM allows feature vector to belong to all clusters with a fuzzy truth value (between 0 and 1). The algorithm assigns feature vector to clusters based on maximum feature vector weight over other clusters. HPF is a hybrid approach. To avoid the intervals determined by altered cluster material this hybrid method is used. This is helpful in improving the clustering performance[33]. This algorithm uses the global optimization. To calculate membership information, the cluster prototypes and the gradient-based FCM is used. HPF Hybrid approach utilizes the global optimization capability of particle swarm. Another method used for missing values imputation. It handles the nonlinearity issues. The input layer structure of ANN was determined by using the AR model. The hybrid AR-ANN method used to analyze the imputation of missing values [34].Before AR modeling to deal with missing values in wind speed time series data sets deletion will be used. This also shows the nonlinearity of wind speed data. Use of the multilayer feed-forward back propagation neural network for time series forecasting was supported by the ANN toolbox in MATLAB software. To create the most appropriate ANN structure, types of hidden and output layers and other requirements were necessary. Hence determine the training functions and the transfer functions. Transfer functions are tan-sigmoid that generates nonlinear outputs among -1 and +1. The log-sigmoid generates nonlinear outputs between 0 and 1 and linear whichgenerates linear outputs between -1 and +1. It is important to select a suitable function to obtain the best results.The best training functions for backpropagation algorithms are Levenberg-Marquardt and Bayesian regularization. To create an applicable ANN the number of neurons must be measured correctly in hidden layer [35].

2.4. Knowledge Assisted Approach

In this method, the incorporation of domain information or outside data into the missing values imputation procedure is present. The imputation accuracy is expressively improved with the use of domain knowledge associated with data-driven method particularly for data sets with high missing rate, a small number of samples and noisy. The correlation information between genes and arrays are exploited when the missing values imputation for Projection onto convex sets (POCS) is a flexible set-theoretic structure [36]. POC Simpute performs local least square regression to capture gene-wise correlation. It also performs PCA imputation to capture array-wise correlation, capture synchronization loss restricts the squared power of the expressions profiles.Using POC Simpute best solution can be obtained irrespective of global or local correlation structure prevalent in the data. This is due to the final solution constantly dominated by smallest yet furthermost consistent constraint set though adequate larger yet less reliable constraint sets. Functionally correlated genes are probable to express in a modular fashion through a higher degree of concerted responses to certain stimuli [37].For gene function classification Gene ontology (GO) is a well-accepted standard[38]. It has three independent ontologies that describe gene product in terms of related biological processes, molecular functions (MF) and cellular components [39].GO increases the imputation accuracy. This is proven in the investigation that when the number of experimental situations is small, or the ratio of annotated genes is large and at higher rates of missing values.To improve the precision of missing values approximation Histone acetylation information aided imputation (HAIimpute) combines histone acetylation information into KNNimpute and LLSimpute [28]. HAIimpute uses the mean expression of genes from all the clusters to form the pattern expression. By fitting a linear regression model, the missing values are then found among the gene and pattern expressions. The final estimations of the missing values are assumed by a convex combination of linear regression imputations and secondary imputation, which is using KNNimpute and LLSimpute. The author proves by the experimental results that HAIimpute reliably develops the KNNimpute or LLSimpute that indicates the imputed genes shows a better correlation with the original complete genes.

III.PERFORMANCE EVALUATION & RESULTS

Assessment of the algorithm imputation results is a critical step for algorithm reliability, performance, and accurate comparison. For missing values imptation there are two main validation types. These are internal validation and external validation. The performance indices are computed among the imputed and the known original values for the validation of the algorithms in internal validation. This validation also uses information from the dataset. While, for external validation, the validation is prepared by following biological analysis for assessing the imputation effects. External validation usually uses the knowledge assembled externally rather than internal data information. The lists of comparative validation methods are shown in table 2.

Table 2. Performance Evaluation Methods for Missing Values Imputation Algorithms

Validation Type
Internal Validation
NRMSE or Variants of it
Pearson Correlation
Preservation of differentially expressed genes
Preservation of prediction / classification problem
External Validation
GO enrichment
Presence of biologically relevant genes

3.1. Internal Validation

By computing the normalized root mean square error (NRMSE) the missing values imputation algorithms are done. For the lower values with NRMSE, the imputation algorithm is more precise.

The NRMSE is defined as:

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik} - \tilde{g}_{ik})^2}{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik})^2}} \qquad (1)$$

If the kth experiment is for gene gi, and g the gik denotes the value and $\hat{g}$ denote the true value and imputed value correspondingly.

3.2. External Validation

The validity of the imputation algorithms is determined by external information such as pathway information and functional annotations. The biologist researches about the term GO that is considerably enriched between the genes. It is then useful to characterize the functional roles of the genes under the research. For each GO terms t the enrichment P-value is calculated using this formula for each cluster:

$$p = \sum_{i=K}^{\min(b,T)} \frac{\binom{T}{i}\binom{B-T}{b-i}}{\binom{B}{b}} \qquad (2)$$

The number of genes in the cluster represented by b, with the GO term t the cluster of genes is represented by K. In dataset a number of genes are B, the number of genes with GO term t represented by T..

IV. LIMITATIONS

There are many advantages and disadvantages of an algorithm, so does the datasets being used for each missing values technique. The performance of missing values imputation algorithms is considerably affected by different factors such as the missing data mechanism, correlationstructure in the data, the percentage of missing values in the data and distribution of missing entries in the data By selecting the right algorithm we may significantly boost the accuracy of the imputation results. However, there is no one imputation algorithm that shows the best results in every condition.With low entropy data sets, Global methods such as SVD impute and BPCA perform better. While with high entropy data sets local methods such as LL Simpute and KNN impute perform better.Most studies showed that missing values in the microarray are randomly missing through missing values imputation. In practice, the missing values also tend to arise in a systematic manner. In the data matrix the distribution of missing entries such as missing data pattern influence the imputations performance and it will need to be considered in data analysis or algorithm design. The data matrix that contains a non-random distribution of missing values may result from different experimental conditions across the columns. In a microarray data matrix, each column comes from one experiment.

V. CONCLUSION

In many studies, it is reported that missing values imputation is a common problem because the microarray gene expression data may include missing values and that effects the conclusion. Due to various experimental causes, the gene expression profiling techniques such as cDNA microarray technology suffers from missing values problem. In microarray data analysis missing values imputation is an essential pre-processing step. Many analyses require the complete data set. If the assessment of existing algorithms were matched and recognized instead of establishing new algorithms that also based on an existing technique than it would be a great contribution. For the optimal execution of the algorithm, assessment of suitable parameters and operating platform could be additional recognized and examined by using different algorithms. This is used to avoid the development of new missing values imputation algorithms. These algorithms perform similarly as the full functional capability of previously existing ones are not fully explored.

VI. FUTURE WORKS

Many algorithms are found for missing values imputation. The Algorithms that gives accurate results and adapt the features of data sets were needed. An adaptive technique that can capture both the local and global method correlation information would be useful in many situations. More experimental data from different kinds of domains becomes available hence new imputation algorithms that can handle categorical data and mixed domain data sets with missing continuous information are required.

ACKNOWLEDGEMENTS

We would like to thank Universiti Malaysia Pahang for supporting this work under the RDU Grant, Grant number: RDU180344 and RDU190113..

REFERENCES

[1] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015;33(5):495.

[2] Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371-85. e18.

[3] Fehrmann RS, Karjalainen JM, Krajewska M, Westra H-J, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nature genetics. 2015;47(2):115.

[4] Lima-Tenório MK, Pineda EAG, Ahmad NM, Fessi H, Elaissari A. Magnetic nanoparticles: In vivo cancer diagnosis and therapy. International journal of pharmaceutics. 2015;493(1-2):313-27.

[5] Criscuolo E, Spadini S, Lamanna J, Ferro M, Burioni R. Bacteriophages and Their Immunological Applications against Infectious Threats. Journal of immunology research. 2017;2017.

[6] Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing. 2017;50:124-34.

[7] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics.Bioinformatics. 2007;23(19):2507-17.

[8] Lai H-H, Chuang T-H, Wong L-K, Lee M-J, Hsieh C-L, Wang H-L, et al. Identification of mosaic and segmental aneuploidies by next-generation sequencing in preimplantation genetic screening can improve clinical outcomes compared to array-comparative genomic hybridization. Molecular cytogenetics. 2017;10(1):14.

[9] Danaee P, Ghaeini R, Hendrix DA, editors. A deep learning approach for cancer detection and relevant gene identification. PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017; 2017: World Scientific.

[10] Lan YD. A Hybrid Feature Selection based on Mutual Information and Genetic Algorithm. Indonesian Journal of Electrical Engineering and Computer Science. 2017;7(1):214-225.

[11] Larose DT, Larose CD. Discovering knowledge in data: an introduction to data mining: John Wiley & Sons; 2014.

[12] Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nature Reviews Genetics. 2016;17(1):47.

[13] Grigoriy G, Eric B, S. RA. New Algorithm and Software (BNOmics) for Inferring and Visualizing Bayesian Networks from Heterogeneous Big Biological and Genetic Data.Journal of Computational Biology. 2017;24(4):340-56.

[14] Zomorrodi AR, Segrè D. Synthetic ecology of microbes: mathematical models and applications. Journal of molecular biology. 2016;428(5):837-61.

[15] Hu W, Lin X, Chen K. Integrated analysis of differential gene expression profiles in hippocampi to identify candidate genes involved in Alzheimer's disease. Molecular medicine reports. 2015;12(5):6679-87.

[16] Cressie N. Statistics for spatial data: John Wiley & Sons; 2015.

[17] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods

applied on microarray data. Advances in bioinformatics. 2015;2015.

[18] Lang KM, Little TD. Principled missing data treatments.Prevention Science. 2018;19(3):284-94.

[19] Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. Journal of Statistical Software. 2016;70(1):1-31.

[20] Tsai C-F, Li M-L, Lin W-C. A class center based approach for missing value imputation. Knowledge-Based Systems. 2018.

[21] Garvey C, Meng C, Nagy JG. Singular Value Decomposition Approximation via Kronecker Summations for Imaging Applications.arXiv preprint arXiv:180311525. 2018.

[22] Chatfield C. Introduction to multivariate analysis: Routledge; 2018.

[23] Tran CT, Zhang M, Andreae P, editors. A genetic programming-based imputation method for classification with missing data. European Conference on Genetic Programming; 2016: Springer.

[24] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology. 2014;179(6):764-74.

[25] Bhattacharya S, Rajan V, Anand A. Clustering high dimensional data using gaussian mixture copula model with lasso based regularization. Google Patents; 2017.

[26] Fox J. Applied regression analysis and generalized linear models: Sage Publications; 2015.

[27] van der Loo M. simputation: Simple Imputation. R package version 02. 2017;2.

[28] Armina R, Zain AM, Ali NA, Sallehuddin R, editors. A Review On Missing Value Estimation Using Imputation Algorithm. Journal of Physics: Conference Series; 2017: IOP Publishing.

[29] Rubinstein RY, Kroese DP. Simulation and the Monte Carlo method: John Wiley & Sons; 2016.

[30] Colantonio A, Di Pietro R, Ocello A, Verde NV, editors. ABBA: Adaptive bicluster-based approach to impute missing values in binary matrices. Proceedings of the 2010 ACM Symposium on Applied Computing; 2010: ACM.

[31] Smart Richman L, Blodorn A, Major B.An identity-based motivational model of the effects of perceived discrimination on health-related behaviors. Group Processes & Intergroup Relations. 2016;19(4):415-25.

[32] Naik B, Mahapatra S, Nayak J, Behera H. Fuzzy Clustering with Improved Swarm Optimization and Genetic Algorithm: Hybrid Approach. Computational Intelligence in Data Mining: Springer; 2017. p. 237-47.

[33] Qi S, Schmid F. Hybrid particle-continuum simulations coupling Brownian dynamics and local dynamic density functional theory. Soft matter. 2017;13(43):7938-47.

[34] Shukur OB, Lee MH. Imputation of missing values in daily wind speed data using hybrid AR-ANN method. Modern Applied Science. 2015;9(11):1.

[35] Kayri M. Predictive abilities of bayesian regularization and Levenberg–Marquardt algorithms in artificial neural networks: a comparative empirical study on social data. Mathematical and Computational Applications. 2016;21(2):20.

[36] Gan S, Wang S, Chen Y, Chen X, Huang W, Chen H. Compressive sensing for seismic data reconstruction via fast projection onto convex sets based on seislet transform. Journal of Applied Geophysics. 2016;130:194-208.

[37] van der Loo M, de Jonge E. Statistical data cleaning with applications in R: John Wiley & Sons; 2018.

[38] Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic acids research. 2016;45(D1):D183-D9.

[39] Aziz MF, Caetano-Anollés K, Caetano-Anollés G. The early history and emergence of molecular functions and modular scale-free network behavior.Scientific reports. 2016;6:25058.

[40] J. Z. and Z. X. Qingbo Li, Wenjie Li, “An Improved K-nearest neighbour method to diagniise breast cancer,” Analyst, 2018