TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 18 - 26

Tensor Factorization with Modified Artificial Bee Colony (MABC) Algorithm for Missing Value Imputation in Breast Cancer Diagnosis

Neeraj Varshney, Narendra Mohan
Department of Computer Engineering & Applications
GLA, University, Mathura, India-281406
neeraj.varshney@gla.ac.in, narendra.mohan@gla.ac.in

Article Info
Volume 82
Page Number: 18 - 26
Publication Issue:
January-February 2020

Article History
Article Received: *14 March 2019*
Revised: *27 May 2019*
Accepted: *16 October 2019*
Publication: *01 January 2020*

Abstract

Cell division and uncontrolled growth caused by changes in cell results in a disease called Cancer. In large data set, patterns are discovered by the process of data mining. Database systems, statistics and machine learning intersections are involved in this method. In clinical diagnosis, machine learning and data mining are commonly used. Missing values are included in this field which reduces the accuracy of diagnosis. Missing data is estimated by an enhanced Reduced Adaptive Particle Swarm Optimization (RAPSO) which is a modified version of tensor factorization. With insufficiency of data and issues in local optima, data's cannot be estimated properly by RAPSO algorithm tensor. So, instead of RAPSO algorithm, Modified Artificial Bee Colony (MABC) is used in tensor factorization method to enhance the accuracy.

Chaotic search is enhanced by Modified Artificial Bee Colony (MABC) algorithm. Process of exploration and exploitation is enhanced by computing all phases of ABC.In employed bee phase, chaotic search based new search is used to enhances onlooker bees probability of finding best solutions. In phase of onlooker bees, new solutions are used to replace worst solutions. Adapt distinctive method for MABC initialization in random manner and it uses Bayesian network. The results of proposed MABC-Bayesian Network (MABC-BN) method are superior to other RAPSO algorithm and BN with respect to specificity Root Mean Square Error(RMSE), sensitivity and accuracy.

Keywords: *Cancer, Modified Artificial Bee Colony (MABC), Root Mean Square Error (RMSE), Bayesian Network (BN), Breast Cancer Diagnosis (BCD)..*

1. Introduction

In the application requiring the usage of all the data, crucial role is played by imputation of missing data. In an extensive data set of real breast cancer, patients recurrence are predicted by using various imputation methods based on machine learning and statistical methods [1]. Imputation corresponds to the process of substituting values to a missing value. Various methods are used in literature for data imputation.

In the prediction of statistics and trends, information which is more useful may be induced by missing values. Observations with missing values are discarded by statisticians for the analysis simplification. For the analysis, size of the sample is reduced by this and may lead to loss of some important information. This results in misleading as well as simplified conclusions [2]. In health care domain, data analysis is depends on imputation.

There are two types of problems in missing value. They are, random missing value and non-random missing value. Compute mean and variance of every group. Between group of observations and group with missing values, mean square error is computed. Missing value is assigned by a group member having mean square error as minimum. Overly simplified solutions may be produced by traditional statistical methods. So by combining statistics and probability, statistical machine learning algorithms are formed to be used in this [3][4].

While taking decisions about, patients health issue, major role is played by a missing data. In process of prediction or estimation, bias is introduced by missing values. Number of available cases are reduced by missing data. Machine learning and data mining methods are used for analyzing the data by researchers [6][7].

In huge set of data, patterns are discovered by data mining techniques. They are used in the intersection of database system, statistics and machine learning. It is a subfield of statistics and computer science. From dataset, information are extracted by this and for further use, they are converted into comprehensive structure. In clinical diagnosis, machine learning and data mining techniques are used commonly. The accuracy of the tensor factorization method, Modified Artificial Bee Colony (MABC) is replaced instead of using RAPSO algorithm.

2. Literature Review

Jerez et al [1] analyzed various machine learning imputation and statistical methods performance. In extensive dataset of real breast cancer, these method used for predicting patient’s recurrence. Listwise deletion (LD) imputation method’s results are used for making comparison. Multiple imputation, hot-deck and mean are example of statistical methods which are used for imputation. k-nearestneighbour (KNN),self-organization maps (SOM) and multi-layer perceptron (MLP) are machine learning techniques. Artificial neural networks (ANNs) are used to measure prediction accuracy of early cancer relapse. In predicting outcome of a patient, better results are produced by machine learning algorithms based imputation methods.

Purwar and Singh [8] used simple K-means clustering methods to analyze different methods of imputation. They proposed a hybrid prediction model with missing value imputation (HPM-MI) model. For a given data, before the application of classifier, class labels are validated by using K-means clustering. After analyzing eleven imputation methods quantitatively, best imputation technique is used to enhance the quality of data by proposed method. Hepatitis from UCI Repository of Machine Learning, Wisconsin Breast Cancer and Pima Indians Diabetes dataset are used for experimentation.

Vazifehdan et al [9] enhanced the breast cancer recurrence prediction by using hybrid imputation method. It uses dependency betweentype of incomplete attributes and attribute. Support Vector Machine, K-Nearest Neighbor, decision tree are the classifiers used Bayesian network on three different datasets. Bayesian network, Tensor factorization, Weighted K-NN, K-NN, Hot-deck and mean imputation methods are used for making performance comparison.

García-Laencina et al [10] analyzed Institute Portuguese of Oncology of Porto’s real breast cancer dataset. There are high percentage of missing values. KNearest Neighbors imputation, Expectation-Maximization imputation, Mode imputation5-year survival prediction from cleaned dataset and -year survival prediction without imputation scenarios are evaluated. Support Vector Machines, Logistic Regression, Classification Trees and K-Nearest Neighbors are used constructing a model which is used to predict survivability of breast cancer. Nested ten-fold cross-validation procedure is used to perform experimentation.

Schmitt et al [11] introduced an element of ambiguity in the analysis of data. Six various methods of imputation are compared. They are, multiple imputations by chained equations (MICE), bayesian principal component analysis (bPCA), singular value decomposition (SVD),fuzzy K-means (FKM), K-nearest neighbors (KNN) and Mean. Four real dataset with different size are used for making performance comparison. Execution time, supervised classification error (SCE), unsupervised classification error (UCE) and Root mean squared error (RMSE) are used for making comparison.

Kalimatha et al [12] preserved the advantage of using multi-view data in diagnosis by handling the missing data using Iterative Singular Value Decomposition (ISVD) imputation. The performance of this method is analyzed using a parameters like Kappa statistics and accuracy of classification. Various range of missing values are used experimentation. Better performance is shown by proposed ISVD imputation when compared with single view method.

3. Proposed Methodology

Improve the accuracy of the tensor factorization. RAPSO algorithm is replaced by Modified Artificial Bee Colony (MABC) is replaced. Chaotic search is enhanced by Modified Artificial Bee Colony (MABC) algorithm. Process of exploration and exploitation is enhanced by computing all phases of ABC.

1.1.Tensor Factorization

Multidimensional arrangements is exhibited by tensors. Degree of tensor corresponds to number dimensions in it. High-degree tensors corresponds to arrangement having three or more dimension. With missing values, large scale data can be analyzed using tensor methods. In recent days, in various areas, tensor analysis is used. The areas includes, computer vision, bioinformatics, signal processing, chemo metrics and psychology. First-degree tensor item set is used to approximate high degree tensor.

Matrix factors are computed by analyzing CP. Kronecker product is known as tensor product is applied on those matrix factors. From various types of huge dataset, information can be discovered by using this matrix factor. Issues with imbalance of data and insufficiency of data are realized by tensor factors. This paper proposes a MABC algorithm to address this issues. In the following manner, third-degree tensor analysis is done. Main matrix is analyzed to compute matrices A, B and C.

$$X \approx \sum_{r=1}^{R} A(:,r)\,{}^{\circ}B(:,r){}^{\circ}C(:,r)$$

The $\mathrm{A} \otimes \mathrm{B}$ represents, matrices $\in R^{I \times J}$ and $B \in R^{K \times L}$'s Kronecker product. Matrix with dimension $\mathrm{Ik} \times \mathrm{JL}$ is produced by this product. Following represents this multiplication.

$$\mathrm{A} \otimes \mathrm{B} \begin{bmatrix} \mathrm{a}_{11}\mathrm{B} & \cdots & \mathrm{a}_{1\mathrm{n}}\mathrm{B} \\ \vdots & \ddots & \vdots \\ \mathrm{a}_{\mathrm{m}1}\mathrm{B} & \cdots & \mathrm{a}_{\mathrm{mn}}\mathrm{B} \end{bmatrix}$$

Linear combination of other features are used by tensor to reconstruct missing values. Assume, three-rank tensor as x with size $\mathrm{I} \times \mathrm{J} \times \mathrm{K}$ and rank of tensor or number of broken matrices as R.

Factor matrices A, B and C are used to create CP factorization. Size of these matrices are given by $\mathrm{I} \times \mathrm{R}, \mathrm{J} \times \mathrm{Rand}\ \mathrm{K} \times \mathrm{R}$. For all values $\mathrm{i} = 1...\mathrm{I}, \mathrm{j} = 1...\mathrm{J}$ and $\mathrm{k} = 1...\mathrm{K}$, following condition is holded.

$$x_{ijk} = \sum_{r=1}^{R} a_{ir}\, b_{jr}\, c_{kr}$$

Primary tensor's error rate of reconstruction is minimized by CP factorization. This makes sum of one-rank tensors with least difference with original tensor. Low value is contained by f function.

$$f(A,B,C)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk}-\sum_{r=1}^{R}a_{ir}\,b_{jr}\,c_{kr}\right)^2$$

With high missing ness percentage, missing data cannot be imputed by CP factorization. Missing values are imputed by a proposed Weighted CP algorithm (CP-WOPT) which is improved version of CP model. In which, size of original tensor equals the weight tensor. F function can be computed as,

$$f(A,B,C)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left\{w_{ijk}\left(x_{ijk}-\sum_{r=1}^{R}a_{ir}\,b_{jr}\,c_{kr}\right)\right\}^2$$

Where, non-negative weight tensor is represented as w, and for all $\mathrm{i}=1\ldots\mathrm{I},\mathrm{j}=1\ldots\mathrm{Jand}=1\ldots\mathrm{K}$, it is initialized as:

$$w_{ijk}=\begin{cases}1\ if\ x_{i,j,k}\ is\ known\\0\ if\ x_{i,j,k}\ is\ unknown\end{cases}$$

Dataset has to converted to interval from numerical values, before the application of Bayesian network on data. Impute missing data by applying Bayesian network on data. Complete the dataset by this method. Data-deficient classes is added with data by MABC. Next section describes the method used to compensate data insufficiency.

1.2.Bayesian Network

Probabilistic graphic model family includes Bayesian network and it is a belief network. Directed acyclic graph (DAG) is contained by this network. Attributes are associated with nodes. over a collection of discrete variables of X dataset, joint probability distribution (Pr_M) and dependency between attributes are represented by nodes. A triple $M=(g,X,P)$, is used to represent a Bayesian network, where $g=(V_g,E_g)$. It represents a X variable's dependency graph and includes a m nodes set with edges. Dependencies between variables are represented by edges.

Conditional probabilities are given by P. $\mathrm{Pr}_M(X_i|PA_i)$. Where nodes are represented by PA_i and it defines X_i. Markov structure is a major capability of Bayesian network. While having parent (pa_i) , non descendants conditional independency exist between every attribute X_i. Following equation shows the joint probability distribution of Bayesian network.

$$Pr_M(X_1\ldots.X_M)=\prod_i Pr_M(X_i|PA_i)$$

Between random variables, complex relationship is learned by Bayesian networks. Classification and approximations can be done using this relationship. Vivid nature of structure is exhibited in some network, in modeling problems which are based on Bayesian network. Computing complete data's probable structure and understanding of graph structure is a biggest difficulty in Bayesian network. Complete dataset is needed by Bayesian network's learning algorithms. This includes, Bound and collapse (BC), Data Augmentation (DA) and Expectation Maximization (EM).

The parameters and structure of incomplete set is adapted by EM algorithm. It requires high amount computation. Averaging and iterations are used for the approximation of missing data. The convergence of DA algorithm is doubtful because density function which is predetermined is used by this and it uses Monte Carlo sampling method for randomly selecting samples. Multiple density functions are used in BC method and averaging is performed in a weighted manner. Without iteration, it is not possible to obtain, precise approximation distribution in Bayesian network.

1.3.Modified Artificial Bee Colony (MABC)

A new nature-inspired optimization algorithm is Artificial bee colony algorithm. Behavior of honeybee swarms is inspired by this algorithm. Imputation of missing bees done by join working of three population based bees. Imputation of missing data is done by employed bees, which communicate with each other. Instead of old position, new position is memorized by employed bee, if it finds a position of a new better missing data.

Employed bees information is used by onlooker bees for making decisions about choosing missing data in exploration. Onlooker bees will memorize the position of new best solution if it founds. Scout bees are formed from employed bee with missing data abandoned for a period of time. In next cycle of search, ne positions are generated by scout bees. Following describes the ABC algorithm.

1. *Initialization*: The initial i^{th} missing data $\mathrm{x_i} = (\mathrm{x_{i,1}}, \mathrm{x_{i,2}}, \ldots, \mathrm{x_{i,D}})$ which is associated with i^{th} bee is generated as,

   $$x_{i,j} = L_j + \phi_{ij}(U_j - L_j)$$

   for $i = 1, 2, 3, \ldots, NP$ and $j = 1, 2, 3, \ldots, D$, where NP is the number of bees and number of variables or dimension is represented by D, random number is given by ϕ_{ij} and it lies between (0,1), andfor the dimension j, lower bound is represented by L_j and upper bound is represented byU_j.
2. *Employed bee phase*: New food source v_i is generated by sharing an information piece byi^{th} bee with k^{th} bee (missing data imputed value). Following expression is used for the same,

   $$v_{i,j} = \begin{cases} x_{i,j} + \phi_{ij}\left(x_{i,j} - x_{k,j}\right); j = j^*, \\ x_{i,j}; \qquad\qquad j \neq j^*, \end{cases}$$

   Where, randomly choose k which range between 1 to NP such that $k \neq i$, randomly choosej^*which range between 1 to D, and ϕ_{ij} is a random number in [−1, 1]. Note that , at j^*th component, v_i is different from x_i . Evaluated the new posiition and compare with old x_i . If $f(v_i) < f(x_i)$, then replace x_i by v_i ; otherwise, hold x_i and $trial(i) = trial(i) + 1$is set, where counter number of unimproved trials is represented as $trial$.
3. *Onlooker bee* phase: Missing data imputation is selected by Onlooker bees based on their quality using probability values (i). It is computed as,

   $$p(i) = \frac{fit(x_i)}{\sum_{j=1}^{NP} fit(x_j)}$$

   Where

   $$fit(x_i) = \begin{cases} \dfrac{1}{1 + f(x_i)}; f(x_i) \geq 0, \\ 1 + |f(x_i)|; otherwise \end{cases}$$

   New v_i is generated, if $\mathrm{rand}(0,1) < \mathrm{p(i)}$. If $f(v_i) < f(x_i)$, then replace x_i by v_i ;otherwise,retain x_i and $trial(i) = trial(i) + 1$ is set.
4. *Scout bee phase*: For x_i, new position is generated, if food source x_i (missing data imputed value) is not improved through limitated number of trails $(limit)$.
5. Find the best position x_{best} and best value f_{best}.
6. Until reaching stopping criterion, steps (2)–(5) are repeated.

Compute all phases of ABC in Modified Artificial Bee Colony (MABC) Algorithm. Search space division (SSD) is used generate the population in initialization [14]. Best solution's information is used to enhance the equation of search gradually in employed bee phase. Search is accelerated by this. For additional search moves, select 25% of employed bees with same probability in onlooker phase.

Current best solution's information is used to construct a new position which replaces the 5% of worst positions. In multimodal functions, scaling factor corresponds to number of best solutions (M). In many best solutions, moves of long distance is provided by this factor. Following shows the proposed MABC algorithm.

1. *Initialization*: The i^{th} food source x_i is generated in order to produce initial solutions with high quality, (missing data imputed value). Search space division is used for this.

$$x_{i,j} = L_j + \frac{(\phi_{ij} + 2i - 1)(U_j - L_j)}{2NP}$$

for$i$ = 1, 2, 3, . . . , NP and j = 1, 2, 3, . . . ,D, where random number is represented by ϕ_{ij} which lies between [−1, 1].

2. *Employed bee phase*: A new food source $\mathrm{v_i}$is generated by using best position x_{best}(missing data imputed value) by the following equation:

$$v_{i,j} = \begin{cases} x_{best,j} + \phi_{ij}\ (\mathrm{x_{i,j}} - \mathrm{x_{k,j}}; \mathrm{j} = \mathrm{j}^*,) \\ x_{i,j}; \qquad\qquad j \neq j^*, \end{cases}$$

Where, randomly chosen number is given by k is and which is selected between 1 to NP such that $k \neq i,$ randomly chosen number is given by j^* and which is selected between 1 to D and ϕ_{ij} is a random number in [−1, 1]. If $f(v_i) < f(x_i)$, then replace x_i by $\mathrm{v_i}$; otherwise, hold x_i and $trial(i) = trial(i) + 1$ is set, where counter number of unimproved trials is represented as $trial$.

3. *Onlooker bee phase*: Constant probability values $p(i) = 0.25$, is used by onlooker bees to make a decision, new position v_i is generated by using $rand(0, 1)$, if $rand(0, 1) < p(i)$;. If $f(v_i) < f(x_i)$,then replace x_i by v_i ; otherwise, retain x_i and $trial(i) = trial(i) + 1$ is set. Below expression is used to replace the worst positions by a new one.

$$x_{z_t} = M[x_{best} + \phi_t(x_{z_t} - x_{r_1}) + \omega_t(x_{best} - x_{r_2})]$$

Where,indexes of 5% worst positions is given by z_t, t = 1, 2, . . . ,⌊0.05NP⌋, randomly chosen indexes are given by r_1 and r_2 are and they range between 1 to NP such that for all t, $r_1 \neq r_2 \neq z_t$, random numbers are represented as ϕ_t and ω_t and they lies between [−1, 1], and M is the number of the best positions obtained from previous generation.

4. *Scout bee phase*: Following firefly algorithm's strategy is used for generating scout bee x'snew position from unupdated position.

$$x_i = x_i + e^{-r_{iq}^2}(x_q - x_i) + (rand(0,1) - 0.5)$$

where$q$ is the first index such that $f(x_q) < f(x_i)$.

Instead of randomness, initial population is used as output in BN in the proposed method in order to prevent local convergence. MABC algorithm is used to generate minority class data, to compensate the problem of data insufficiency of classes. Chaos theory is used for generating initial population in MABC algorithm. Disordered and chaotic appearances is shown by the systems described by this criterion. Apparently random data's order or sequence is computed by developing chaos theory.

Lot of attention is attracted by chaos theory in recent days. Use of chaos signal is shown by the results of experimentation. In comparison with random sequences, algorithm's indicators efficiency is enhanced by chaos sequence. For random process, enhanced power in searching process is exhibited by chaos-based method. Because of this, it is used various random generator applications. For different problems, it can be adapted easily. Various mappers are used for generating chaos.

4. Results And Discussion

Dataset of Diagnostic Wisconsin Breast Cancer Database is used for evaluating proposed MABC method. There are 569 samples with 32 attributes. Digitized breast mass's fine needle aspirate (FNA) image is used for computing features. In an image, cell nuclei's characters are described by this. In predictive analytics, a confusion matrix (Table 1), involves a Table with two columns and rows that corresponds to the number of true negatives, true positives, false negatives and false positives. This permits more elaborate inspection in comparison to just by accuracy. Accurateness is not regarded to be a persistent metric for real performance of a classifier, since it may lead to confusing results if there is instability in the data set.

Table 1. Confusion Matrix

		Predicted	
		Positive	**Negative**
Actual	**Positive**	TP	FP
	Negative	FN	TN

- True positive (TP) represents amount of correct positive predictions.
- False Negative (FN) represents amount of incorrect negative predictions.
- False Positive (FP) represents amount of incorrect positive predictions.
- True Negative (FN) represents amount of correct negative predictions.

The Sensitivity isratio of positive cases that are correctly got. It is also termed as True Positive Rate (TPR) and it is given by,

$$\text{Sensitivity} = TP/(TP+FN) \qquad (1)$$

Ratio of total number of predictions that were right produces accuracy value. It is given by (2)

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \qquad (2)$$

Ratio of number of correct negative predictions to total number of negatives produced specificity. It is also termed as true negative rate (TNR).

$$\text{Specificity} = TN/(FP + TN) \qquad (3)$$

The overall results of the methods with performance evaluation metrics is discussed in table 2.

Table 2. Performance Comparison Metrics vs. BCD Classification Methods

Methods	Metrics		
	Sensitivity (%)	**Specificity(%)**	**Accuracy (%)**
KNN	95.54	95.24	95.43
RAPSO	96.09	96.62	96.26
MABC	96.15	97.05	96.66

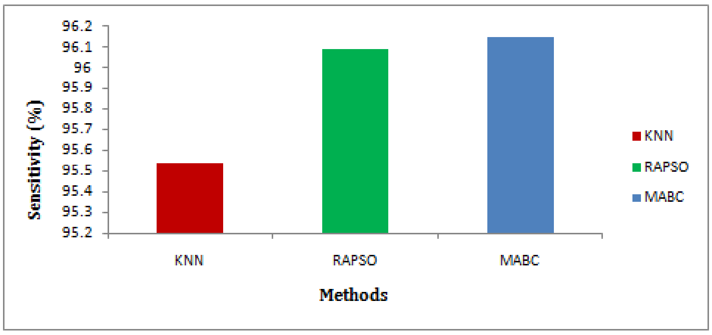

Figure 1. Sensitivity Results Evaluation of BCD Classification Methods

Figure 1 shows the performance results of sensitivity metrics with respect to three classifiers like proposed KNN, RAPSO and MABC. The results demonstrate that the proposed MABC classifier gives higher sensitivity value of 96.15%, the other existing methods such as KNN, RAPSO gives lesser sensitivity value of 95.54%, 96.09% respectively.

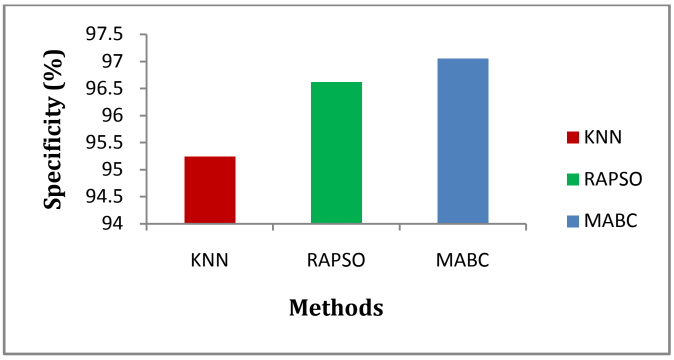

Figure 2. Specificity Results Evaluation of BCD Classification Methods

Figure 2 shows specificityresults metrics with respect to three classifiers like proposed KNN, RAPSO and MABC. It shows that the proposed MABC classifier provides higher specificity value of 97.05%, the existing methods such as KNN, RAPSO gives lesser specificity value of 95.24%, 96.62% respectively.

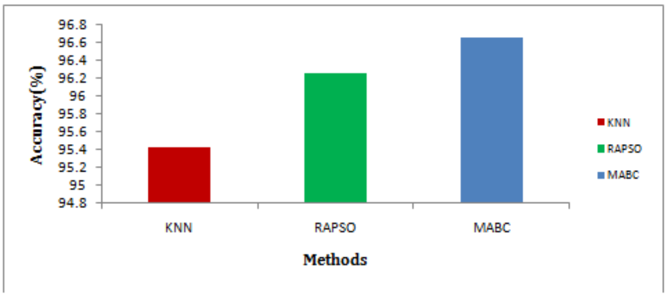

Figure 3. Accuracy Results Evaluation of BCD Classification Methods

Accuracyresults comparison of the three classification methods are shown in figure 3. Those methods are KNN, RAPSO and MABC. The results discloses that the proposed MABC classifier produces 96.66% of accurate results, whileKNN, RAPSOproducing 95.43% and 96.26% of accurate results.

5. Conclusion And Future Work

In medical application, methods to evaluate missing values are progressively increasing. Imputation of missing data is performed by tensor techniques in the previous study. Class imbalance is having serious effect in performance of it. So this class imbalance problem is rectified by using tensor factorization based enhanced missing data imputation technique. It used MABC. In medical decisions, significant role is played by this and it minimizes the level of error. In this paper, Diagnostic Wisconsin Breast Cancer Database is used to evaluate the performance of proposed MABC method. Specificity, sensitivity and accuracy are used as a performance evaluating measures. Better performance is shown by proposed method when compared with existing methods. In future, performance degradation by the presence of noise can be addressed.

References

1. J.M. Jerez, I. Molina, P.J. García-Laencina, E. Alba, N. Ribelles, M. Martín and L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artificial intelligence in medicine, Vol.50, No.2, pp.105-115, 2010.
2. R.J. Little and D.B. Rubin, Statistical analysis with missing data (Vol. 793), John Wiley & Sons, 2019.
3. D. Lowd and P. Domingos, Naive bayes models for probability estimation, Proceedings of the 22nd international conference on Machine learning, pp.529-536, 2005.
4. J.M. Jerez, I. Molina, J.L. Subirats and L. Franco, Missing data imputation in breast cancer prognosis, Survival, Vol.8, No.9, pp.10-11, 2006.
5. S. Tirunagari, N. Poh, H. Abdulrahman, N. Nemmour and D. Windridge, Breast cancer data analytics with missing values: A study on ethnic, age and income groups, arXiv preprint arXiv: 1503.03680, 2015.
6. A. Nekouie and M.H. Moattar, Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization, Journal of King Saud University-Computer and Information Sciences, Vol.31, No.3, pp.287-294, 2019.
7. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis and D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal, Vol.13,pp.8-17, 2015.

8. A. Purwar and S.K. Singh, Hybrid prediction model with missing value imputation for medical data, Expert Systems with Applications, Vol.42, No.13, pp.5621-5631, 2015.
9. M. Vazifehdan, M.H. Moattar and M. Jalali, A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction, Journal of King Saud University-Computer and Information Sciences, Vol.31, No.2, pp.175-184, 2019.
10. P.J. García-Laencina, P.H. Abreu, M.H. Abreu and N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Computers in biology and medicine, Vol.59, pp.125-133, 2015.
11. P. Schmitt, J. Mandel and M. Guedj, A comparison of six methods for missing data imputation, Journal of Biometrics & Biostatistics, Vol.6, No.1, pp.1-6, 2015.