Knowl Inf Syst (2017) 52:709–750
DOI 10.1007/s10115-017-1025-5

REGULAR PAPER

Missing value estimation for microarray data through cluster analysis

Soumen Kumar Pati[1] · Asit Kumar Das[2]

Received: 1 April 2016 / Revised: 5 January 2017 / Accepted: 28 January 2017 /
Published online: 13 February 2017

Abstract Microarray datasets with missing values need to impute accurately before analyzing diseases. The proposed method first discretizes the samples and temporarily assigns a value in missing position of a gene by the mean value of all samples in the same class. The frequencies of each gene value in both types of samples for all genes are calculated separately and if the maximum frequency occurs for same expression value in both types, then the whole gene is entered into a subset; otherwise, each portion of the gene of respective sample type (i.e., normal or disease) is entered into two separate subsets. Thus, for each gene expression value, maximum three different clusters of genes are formed. Each gene subset is further partitioned into a stable number of clusters using proposed *splitting and merging* clustering algorithm that overcomes the weakness of Euclidian distance metric used in high-dimensional space. Finally, similarity between a gene with missing values and centroids of the clusters are measured and the missing values are estimated by corresponding expression values of a centroid having maximum similarity. The method is compared with various statistical, cluster-based and regression-based methods with respect to statistical and biological metrics using microarray datasets to measure its effectiveness.

Keywords DNA Microarray data · High-dimensional space · Clustering algorithm · Similarity measurement · Missing value imputation

✉ Soumen Kumar Pati
soumenkrpati@gmail.com

Asit Kumar Das
akdas@cs.iiests.ac.in

[1] Department of Information Technology, St. Thomas' College of Engineering and Technology, Kolkata, India

[2] Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India

1 Introduction

DNA microarray techniques give an overall outlook of gene expression, observing the mRNA levels of thousands or more number of genes. Microarray dataset [7,15] is basically a large matrix of expression levels of observations or genes under different experimental conditions or samples. The datasets generally contain missing values in time of preparation, such as dust or spotting or scratches on the slide, insufficient resolution, hybridization failures and image corruption [25]. But a good number of algorithms for gene dataset analysis take a complete dataset as input. Therefore, more accurate prediction of missing values is an important preprocessing step to form complete dataset and perform further experiments. Approximately 5% or more cell values of the matrix can often be missing unless extreme care is taken by the organization [36,44].

1.1 Literature review

The researchers propose several methods [8,24,48,49] to deal with missing values. The expensive and more time-consuming method presents in [5], where the original experiment is repeated until dataset without missing values is obtained. On the other hand, the method in [1] ignores missing value-related genes, which usually loses useful information and may bias the result if the remaining genes not capable to present the complete dataset. Some methods [1,38] impute the missing values by a constant such as zero (0), or by the mean of the available sample values, which distort correlations among expression values. Another method [17] considers the relationships among gene expression values. It first measures the similarity between a gene with missing value and genes without missing values before predicting the missing values of the gene by observed values of the most similar genes. In the papers [19,41], a missing value estimation method called singular value decomposition imputation (SVD-impute) is reported where missing values are estimated by identifying the K most significant Eigen genes. The paper [41] proposes a method called weighted KNN-impute that constructs the missing values using a weighted average of K most similar genes. Estimation ability of the method [41] is more robust than others, such as imputation by zero, row average or SVD-impute. In [30], a KNN-based missing data estimation algorithm is proposed based on the temporal and spatial correlation of sensor data. The methods discuss in [30,41] have better performance than prior method, but drawback is that their estimation ability depends on parameter K (i.e., number of neighbor genes used to impute missing value) for which no theoretical way exists to determine them appropriately and thus needs to be specified by the user. In the paper [34], the missing values are estimated using fuzzy c-means clustering algorithm and semantic similarity among gene expression data, but the method requires gene oncology structure among several genes, which depend on biological knowledge, whereas, in the paper [4,25,46,47], cluster-based algorithms have been proposed to deal with missing values which do not need such parameters but microarray dataset is very high dimensional and there exist large number of genes with large number of samples which may degrade the clustering performance. Also performance of these methods depends on number of clusters whose selection becomes very crucial. In the paper [37], the missing values are estimated to preprocess the datasets using fuzzy c-means clustering-based expectation maximization (EM) method.

The KNN-impute [41] method depends on the intuitive assumption that the genes closed to each other are potentially similar. The measurement considers both distance computation between genes and the number of nearest neighbors (K). As the missing gene may contain variant number of missing values, so the genes without missing values may have different lengths and their distances are inaccurate. The paper [21] proposes a sequential KNN imputation method (SKNN-impute) where missing values are estimated sequentially starting with a gene having the smallest missing rate. It uses the genes without any missing values for estimating the missing values of the genes. The SKNN-impute method iteratively estimates the missing values of the genes based on the ascending order of their missing rates. The paper [3] proposes an iterative KNN imputation method (IKNN-impute) which first replaces all missing values of a gene by the row average of all values in it. Then an iterative process is performed to obtain all missing values same as SKNN-impute method. The paper [31] proposes a similarity-based missing value estimation technique (PCM-impute), but it considers only single similar gene for estimating missing value. The improved version [32] (HCS-impute) of paper [32] is also a KNN-based algorithm that selects eight neighbors of a gene, but the demerit of the method is that it needs the value of K in advance.

The Bayesian principal component analysis (BPCA) [28] method estimates the missing values by linear combination of certain principal axis vectors, where the parameters are identified by Bayesian estimation method. On the other hand, local least squares (LLS) [22], sequential local least square (SLLS) [46,47] and iterated local least square (ILLS) [6] methods utilize multiple regression models to impute the missing values from KNN genes of the missing value-related genes. Recently, bi-cluster-based estimation methods (BIC) [18] such as bi-cluster-based least square (bi-iLS) [8] and bi-cluster-based BPCA (bi-BPCA) [27] aim to estimate missing values with some integrated approaches and give acceptable results. Another method [16] estimates missing values with high accuracy using triple imputation strategies (TRIIM) based on BPCA, LLS and EM concepts.

1.2 Contribution and comparison

In the paper, a novel 'Missing Value Estimation Technique through Cluster Analysis' (MVETCA) has been proposed on microarray dataset for imputing missing values that not only overcomes the constraints of the existing methods but also gives significantly better statistical measures like less normalized root mean square error *(NRMSE)* [39], high conserved pairs proportion (*CPP*) [11] and high biomarker list concordance index (BLCI) [29].

The proposed method of missing value estimation consists of the following steps:

i. The dataset is discretized to Z-score using transitional state discrimination (TSD) method [43], and the genes are characterized by N discrete sample values. The samples are divided into two disjoint classes, because they are collected from both normal and tumor patients. So, the frequencies of sample values are calculated separately in each class for each gene.

ii. The whole gene set is partitioned into two clusters; one contains all genes without any missing values termed as '*NOMISS*' and the other contains all genes with missing values termed as '*MISS*.' Then missing values of genes in *MISS* is filled up using the mean gene value for all samples belonging to the same class, either normal or disease class. And finally, all genes in *MISS* are also replicated in *NOMISS*. Thus, all missing values are temporarily estimated.

iii. Based on the frequencies of discrete sample values, the gene set *NOMISS* is partitioned into maximum $3N$ gene subsets, explained in Sect. 2.2. Now, N out of $3N$ clusters contains whole part (i.e., normal and cancerous samples) of the genes while each N of remaining $2N$ clusters contains only one part (i.e., either normal or cancerous samples) of the genes.

iv. Each gene in gene set *MISS* is associated with one of the N subsets containing whole portion, or each of its two portions (i.e., normal and cancerous) is associated with one of the $2N$ respective subsets containing only one portion. The association of gene is

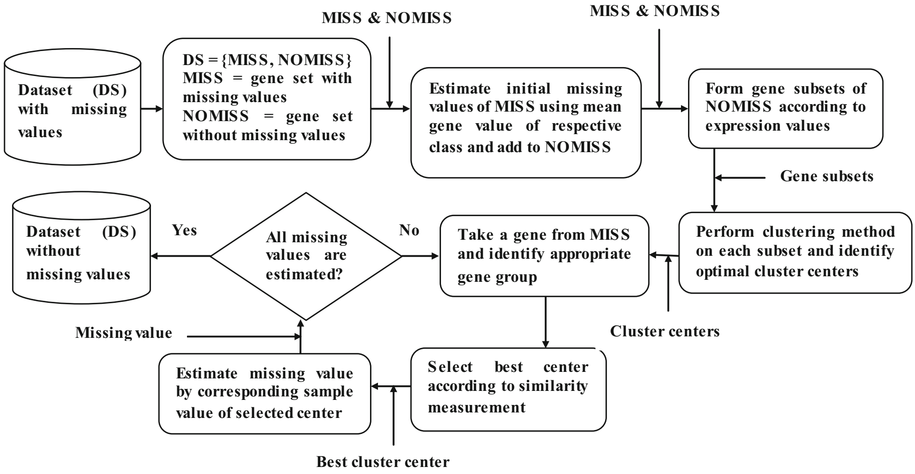

MISS & NOMISS
MISS & NOMISS
Dataset (DS) with missing values
DS = {MISS, NOMISS} MISS = gene set with missing values NOMISS = gene set without missing values
Estimate initial missing values of MISS using mean gene value of respective class and add to NOMISS
Form gene subsets of NOMISS according to expression values
Gene subsets
Dataset (DS) without missing values
Yes
All missing values are estimated?
No
Take a gene from MISS and identify appropriate gene group
Perform clustering method on each subset and identify optimal cluster centers
Missing value
Cluster centers
Estimate missing value by corresponding sample value of selected center
Select best center according to similarity measurement
Best cluster center

Fig. 1 The overall work flow of proposed missing value estimation method

determined by measuring the similarity of the gene expression values with the gene subsets, explained in Sect. 2.3.1.

v. Now the $3N$ gene subsets are partitioned into optimal number of clusters (optimality is determined through validation indices) using similarity-based clustering algorithm where similarity factors are measured between centroid of each partition and associated portion of the missing value-related gene. The missing values of the associated portion of the gene are imputed by the respective values of the centroid with most similar partition. Thus, missing values of each gene are imputed by repeating steps (iv) and (v).

The pictorial representation of the proposed method is shown in Fig. 1.

The HCS-impute method [32] uses eight similar genes, and SKNN-impute [21] and IKNN-impute [3] use twenty or less similar genes to estimate missing values, but the proposed method MVETCA has no such limitation. In KNN-impute [41], PCM-impute [31] and HCS-impute [32], the missing values are estimated only with the help of NOMISS genes. In SKNN-impute [21], sequentially the missing values are estimated according to the non-decreasing order of count of missing values of *MISS* genes and *NOMISS* dataset is updated after estimating the missing values one by one. But IKNN-impute [3], bi-BPCA [27], TRIIM [16] and proposed method use full dataset to estimate missing values. In IKNN-impute [3], row-average method considering all sample values of a gene is used to temporarily estimate the missing values of *MISS* and update the *NOMISS* set accordingly. Similarly, bi-BPCA [27] uses BPCA and TRIIM [16] uses BPCA, LLS and EM method to temporarily estimate the missing values of *MISS* and update the *NOMISS* set accordingly. But MVETCA uses cancerous or normal sample values, which is more logical with respect to expression values for a particular class. The MVETCA uses Hamming distance-based similarity function, whereas other methods use Euclidian distance metric for similarity measurement of genes, which is not so effective especially in case of high-dimensional datasets. The proposed MVETCA method is compared with some other imputation methods with the help of some statistical measures like NRMSE [39], CPP [11] and BLCI [29] considering six publicly available microarray datasets.

The paper is organized into four sections. Section 2 describes the proposed missing value estimation technique through cluster analysis. The experimental results and performances of the proposed method are evaluated for various benchmark microarray datasets in Sect. 3. Finally, the work is concluded in Sect. 4.

2 Missing value estimation

The efficient microarray technology [12] evaluates the gene expressions simultaneously under different experimental conditions. Generally, microarray dataset contains missing values for which almost (5–50%) genes are affected. Missing value is a crucial problem needs to handle before analysis of the microarray data in order to acquire important knowledge. Therefore, missing value imputation is a necessary preprocessing step to estimate proper expression values.

2.1 Gene expression discretization

Initially, dataset DS $= (U, C)$ has some missing values which are temporarily estimated by the mean gene value for all samples belong to the same class (either normal or disease), of a gene, where U the universe of discourse contains g genes and C the condition attribute set contains s samples. The gene subset with missing values is referred as *MISS*, and the initial estimated whole gene set is termed as *NOMISS*. The transitional state discrimination (*TSD*) method [43] is used to discretize *MISS* and *NOMISS*. The discretization factor f_{ij}, based on which the dataset is discretized, is computed for sample $C_j \in C$ of gene $g_i \in U$, using Eq. (1), for $i = 1, 2,\ldots,g$ and $j = 1, 2, \ldots, s$.

$$f_{ij} = \frac{M_i\left[C_j\right] - \mu_i}{\delta_i} \tag{1}$$

where μ_i and δ_i are the mean and standard deviation of gene g_i, respectively, and $M_i[C_j]$ is the sample value C_j in gene g_i. Then, mean (N_i) of negative sample values and mean (P_i) of positive sample values of each gene g_i are calculated and discretized to one of N (here, $N = 5$) fuzzy linguistic terms using Eq. (2).

$$f_{ij} = \begin{cases} 'VL' & |if\ \ f_{ij} \leq N_i \\ 'L' & |if\ \ N_i < f_{ij} < 0 \\ 'Z' & |if\ \ f_{ij} = 0 \\ 'H' & |if\ \ 0 < f_{ij} < P_i \\ 'VH' & |if\ \ f_{ij} \geq P_i \end{cases} \tag{2}$$

2.2 Structure of correlated gene subsets

Based on the frequencies of discrete sample values, the genes of *NOMISS* are partitioned into $3N$ different groups. N out of $3N$ groups contains whole (i.e., normal and cancerous samples) of the genes while each N of remaining $2N$ groups contains only one portion (i.e., either normal or cancerous samples) of the genes.

Either a gene of MISS is associated with one of the N groups containing whole gene or each of its two portions (i.e., normal and cancerous) is associated with one of the $2N$ respective groups containing only one portion.

Let the samples of the dataset are collected together from s_1 normal and s_2 cancerous patients and so each gene contains s_1 normal and s_2 cancerous samples. Let each gene $g_i \in NOMISS$ is represented as $g_i = \{g_{i1}^n, g_{i2}^n, , \ldots g_{is_1}^n, g_{i1}^c, g_{i2}^c, \ldots, g_{is_2}^c\}$ where g_{ij}^n for

$j = 1, 2, \ldots, s_1$ are normal samples and g_{ik}^{c} for $k = 1, 2, \ldots, s_2$ are cancerous samples. Frequencies of discrete expression values for samples $\{g_{i1}^{n}, g_{i2}^{n}, \ldots, g_{is_1}^{n}\}$ and $\{g_{i1}^{c}, g_{i2}^{c}, \ldots, g_{is_2}^{c}\}$ of gene g_i are computed as $\left\{f_{VL}^{ni}, f_{L}^{ni}, f_{Z}^{ni}, f_{H}^{ni}, f_{VH}^{ni}\right\}$ and $\left\{f_{VL}^{ci}, f_{L}^{ci}, f_{Z}^{ci}, f_{H}^{ci}, f_{VH}^{ci}\right\}$, respectively, where f_{VL}^{ni} is the frequency of expression value '*VL*' in normal samples of gene g_i, similar meaning of other terms. Let $f_{\max}^{ni} = \max\left\{f_{VL}^{ni}, f_{L}^{ni}, f_{Z}^{ni}, f_{H}^{ni}, f_{VH}^{ni}\right\}$ and $f_{\max}^{ci} = \max\left\{f_{VL}^{ci}, f_{L}^{ci}, f_{Z}^{ci}, f_{H}^{ci}, f_{VH}^{ci}\right\}$. The gene subsets are formed as follows:

If $f_{\max}^{ni}$ and $f_{\max}^{ci}$ are computed from:

(i) Same discrete expression value say '*VL*' then the gene $g_i = \{g_{i1}^{n}, g_{i2}^{n}, \ldots, g_{is_1}^{n}, g_{i1}^{c}, g_{i2}^{c}, \ldots, g_{is_2}^{c}\}$ is placed in subset GENE_WHOLE$_{\mathrm{VL}}$ (abbreviated as GW$_{\mathrm{VL}}$, subsequently used throughout the paper). Similarly, considering other discrete values, total of five subsets GW$_{\mathrm{VL}}$, GW$_{\mathrm{L}}$, GW$_{\mathrm{Z}}$, GW$_{\mathrm{H}}$ and GW$_{\mathrm{VH}}$ are formed. Each of these five subsets contains genes of *NOMISS,* where maximum frequency of discrete value occurs for same discrete value in both normal and cancerous samples.

(ii) Different discrete expression value say $f_{\max}^{ni}$ occurs for '*VL*' and $f_{\max}^{ci}$ occurs for '*VH*.' In this case, the normal part $\{g_{i1}^{n}, g_{i2}^{n}, \ldots, g_{id_1}^{n}\}$ of g_i is placed in subset GENE_NORMAL$_{\mathrm{VL}}$ (abbreviated as GN$_{\mathrm{VL}}$), and same treatment takes place for other discrete values. And cancerous part $\{g_{i1}^{c}, g_{i2}^{c}, \ldots, g_{id_2}^{c}\}$ of g_i is placed in subset GENE_CANCER$_{\mathrm{VH}}$ (abbreviated as GC$_{\mathrm{VH}}$), same situation occurs for other discrete values. Thus, gene subsets GN$_{\mathrm{VL}}$, GN$_{\mathrm{L}}$,GN$_{\mathrm{Z}}$, GN$_{\mathrm{H}}$ and GN$_{\mathrm{VH}}$ are formed, each of which contains normal samples of genes whose maximum frequency discrete value differs from that of cancerous samples. Similarly, gene subsets containing only cancerous samples are formed which are GC$_{\mathrm{VL}}$,GC$_{\mathrm{L}}$,GC$_{\mathrm{Z}}$, GC$_{\mathrm{H}}$ and GC$_{\mathrm{VH}}$.

Thus, fifteen subsets are formed for the genes of *NOMISS*. These subsets are created according to the gene expression values of the dataset, and each subset contains similar nature of expression values.

2.3 Gene clustering and analysis

The ideal input for a clustering algorithm is a dataset without any noise. When the experimental data deviates from this property, it poses different problems for different types of clustering algorithms. Missing values in the dataset used during cluster analysis are a very crucial problem handled carefully for high-dimensional data.

2.3.1 Gene subset selection

The set NOMISS is partitioned into fifteen subsets without missing values, and each subset contains genes with similar nature according to their expression values. On the other hand, the set MISS contains genes with missing values need to be estimated prior to gene data analysis. Each gene $g_j \in MISS$ is denoted by $g_j = \left\{g_{j1}^{n}, g_{j2}^{n}, \ldots, g_{js_1}^{n}, g_{j1}^{c}, g_{j2}^{c}, \ldots, g_{js_2}^{c}\right\}$, where some missing normal and cancerous samples g_{jk}^{n} and g_{jl}^{c}, for $k = 1, 2, \ldots, s_1$ and $l = 1, 2, \ldots, s_2$ need to be estimated. The method computes the frequency of discrete expression values in both normal and cancerous samples of gene $g_j \in MISS$. If maximum frequency occurs in both types of samples for same expression value, say '*VH*', then g_j is related to subset GW$_{\mathrm{VH}}$. But if maximum frequency occurs for different expression values, say '*VL*' and '*VH*' for normal class and cancerous class, respectively, then normal samples $\{g_{j1}^{n}, g_{j2}^{n}, \ldots, g_{js_1}^{n}\}$ of g_j is associated with GN$_{\mathrm{VL}}$ and cancerous samples $\{g_{j1}^{c}, g_{j2}^{c}, \ldots, g_{js_2}^{c}\}$ of gene g_j is associated with GC$_{\mathrm{VH}}$. Thus each gene $g_j \in MISS$ is either

(i) Linked with any one subset of the set GW = {GW$_{\mathrm{VL}}$, GW$_{\mathrm{L}}$, GW$_{\mathrm{Z}}$, GW$_{\mathrm{H}}$, GW$_{\mathrm{VH}}$} or

(ii) Normal portion of it is linked with any one of GN = {$\mathrm{GN_{VL}}$, $\mathrm{GN_L}$, $\mathrm{GN_Z}$, $\mathrm{GN_H}$, $\mathrm{GN_{VH}}$} and cancerous portion of it is associated with any one of GC = {$\mathrm{GC_{VL}}$, $\mathrm{GC_L}$, $\mathrm{GC_Z}$, $\mathrm{GC_H}$, $\mathrm{GC_{VH}}$}.

2.3.2 Clustering of gene subset by splitting and merging

The most common distance metric used for distance measure is Euclidean distance. Though it is very useful in low dimensions, it does not work well in high-dimensional cancer dataset. It is observed that Euclidean distance does not capture the similarity of high-dimensional objects. Rather, bitwise similarity measurement, i.e., Hamming distance is more powerful for missing value estimation of genes. Here, gene subset associated with missing gene $g_j \in MISS$ is partitioned using similarity-based proposed clustering algorithm which provides set of K-clusters. If g_j is associated with a subset of the set GW, then only that subset of genes is clustered to impute the missing values of g_j. And if missing values of g_j arise both in normal portion and cancerous portion, the corresponding subsets of both the set GN and GC are clustered; otherwise, clustering algorithm is applied only on the corresponding subset of either GN or GC.

The gene subset is grouped into significant clusters capture the normal structure of the data to find genes with related functionality. Cluster analysis has been applied in different fields, like information science to social sciences [23] and biological science [33]. The relationships of genes are established from available information using clustering algorithms [13,35,40]. The aim of clustering algorithms is to collect the similar nature genes in a cluster and dissimilar genes in different clusters according to their expression values. The validity indices [2,14] proposed by the researchers measure the goodness of clusters by checking entropies of the partition, membership distributions, clusters separation and compactness. The proposed method introduces merging and splitting procedure [9] on initial set of clusters, validates newly generated clusters using cluster validation indices and finally produces optimal set of clusters used to impute the missing values of the genes. Some notations and terms used in the algorithm are described below.

A. Notation and Definition

For convenience, following notations are used in proposed clustering algorithm:

$\boldsymbol{C_i}$: It is the ith cluster, where i = 1, 2, …, k.
$\boldsymbol{x_k}$: It is the kth object of any cluster.
$\boldsymbol{n_i}$: Number of objects of ith cluster.
$\boldsymbol{S(x_k, C_i)}$: It is the similarity function which computes bitwise matching of discrete sample values of ith cluster center to the kth object x_k of ith cluster.
$\boldsymbol{S(C_i, C_j)}$: It is the similarity function which computes bitwise matching of the ith cluster center with jth cluster center.

Few terminologies are defined below for understanding the proposed clustering algorithm:

Definition 1 (*Combine Center*) Combine center $CCcomb_{ij}$ of two clusters C_i and C_j ($i, j = 1, 2, \ldots, k$ and $i \neq j$) is the weighted mean of centers of two clusters C_i and C_j, computed using Eq. (3) and is considered as the center of combined cluster C_ij, where $\overline{C_i}$ and $\overline{C_j}$ are the mean of C_i and C_j, respectively.

$$CCcomb_{ij} = \frac{\left(n_i \times \overline{C_i}\right) + \left(n_j \times \overline{C_j}\right)}{n_i + n_j} \tag{3}$$

Definition 2 (*Similarity Factor*) Similarity factor between two clusters C_{i} and C_{j} $(i, j = 1, 2, \ldots, k$ and $i \neq j)$ is denoted by CSF_{ij} and is defined by Eq. (4).

$$\mathrm{CSF}_{ij} = \frac{\sum_{x_k \in C_i} S\left(x_k, CCcomb_{ij}\right) + \sum_{x_k \in C_j} S\left(x_k, CCcomb_{ij}\right)}{n_i + n_j} + S(C_i, C_j) \tag{4}$$

Definition 3 (*Merging factor*) Let $C_1, C_2, \ldots, C_k$ be the k clusters. The merging factor MF is the cluster validation index for k clusters computed using Eq. (5) that is used to evaluate the clusters to obtain optimal set of clusters.

$$\mathrm{MF} = \frac{\sum_{i=1}^{k} \sum_{x_j \in C_i} S\left(x_j, C_i\right)}{\sum_{i=1}^{k} \left(\min\limits_{1 \leq j \leq k, i \neq j} S(C_i, C_j) \right)} \tag{5}$$

Definition 4 (*Splitting Factor*) It is the ratio of intra-cluster similarity to the minimum inter-cluster similarity of ith cluster to all other clusters. The Split Factor $\mathrm{SF_i}$ is calculated using Eq. (6) for each individual cluster.

$$\mathrm{SF}_i = \frac{\sum_{x_k \in C_i} S\left(x_k, C_i\right)}{\min\limits_{1 \leq j \leq k, i \neq j} S(C_i, C_j)} \tag{6}$$

Definition 5 (*Mean of Split Factor*) It is an average split factor of all clusters used as a threshold value (γ), calculated using Eq. (7) to split a cluster.

$$\gamma = \frac{1}{k} \sum_{i=1}^{k} \mathrm{SF}_i \tag{7}$$

B. Merging of Clusters

Let there are n genes in a subset which need to be clustered. Initially, each gene is considered as a separate singleton cluster, and thus, n clusters C_1, C_2, …, C_n are formed. For any two clusters C_i and C_j $(i, j = 1, 2, \ldots, n$ and $i \neq j)$ similarity factor CSF_{ij} is computed using Eq. (4) with the implication that, if the similarity of the combine cluster (C_{ij}) is high and at the same time the similarity of the two respective cluster centers is high, then the similarity factor (CSF_{ij}) between the cluster is high and the clusters are much similar to each other according to the nature of objects of clusters. This implies that higher the similarity factor between the pair, more similar the clusters are. In Eq. (4), the weighted mean $(CCcomb_{ij})$ representing the center of combined cluster (C_{ij}) is computed using Eq. (3) and original gene expression values of C_i and C_j are used, not the discrete values.

Thus, a similarity matrix $S = (\mathrm{CSF}_{ij})_{n \times n}$ is created using Eq. (4) which is a symmetric matrix with empty diagonal entries, as the similarity of a cluster with itself is not required. All the $\frac{n(n-1)}{2}$ similarity factors reside above the leading diagonal of S stored the information based on which clusters are merged. In every iteration, only the cluster pair with maximum similarity are merged and reducing the number of clusters by one.

Initially, the merging factor (MF) is computed using Eq. (5) and it is used while the merging process is over to measure the better clustering phenomenon that finally helps to form the stable set of clusters. In Eq. (5), if the numerator value is high (i.e., the points of each cluster are much more similar) and denominator is low (i.e., the cluster centers are more dissimilar), then the cluster qualities are good enough. High value of MF corresponds to clusters that are formed with more similar nature points and centers are more dissimilar with each other.

So the larger the MF value better is the clustering. The merging procedure continues if MF value increases; otherwise, splitting procedure resumes, explained in Sect. 2.3.2C.

Let $C_1, C_2, \ldots, C_n$ be the n number of clusters. After first merging, $(n-1)$ clusters are obtained whose MF value MF_{n-1} is computed using Eq. (5). The process terminates if MF_{n-1} is larger than MF_n by a threshold value (set experimentally) and the system is rolled back to the previous state to preserve the previous set of n clusters; otherwise, same process is repeated with $(n-1)$ clusters.

The algorithm of merging process is given in detail below:

```
Algorithm: Merging_Clusters (Clus, n)
Input: Clus = {$C_1$, $C_2$, …, $C_n$} of n clusters for n genes.
Output: The set of modified clusters in Clus.
Begin
    $MF_{old}$ = MF value of Clus using Eq. (5)
    For i = 1 to n do
        For j = i+1 to n do
      $CSF_{ij}$ = Similarity factor between $C_i$ and $C_j$ in Clus using Eq. (4)
    End
End
max = $S_{12}$                    /*compute maximum similarity factor*/
For i = 1 to n do
    For j = i+1 to n do
        If ($CSF_{ij}$ > max) then
            max = $SF_{ij}$
            p = i
            q = j
        End
    End
End
$C_{pq} = C_p \cup C_q$
Clus = Clus ∪ {$C_{pq}$} – {$C_p$} – {$C_q$}
$MF_{new}$ = MF value of Clus using Eq. (5)
If (($MF_{new} > MF_{old}$) | | ((|$MF_{new}$ - $MF_{old}$|) < δ ))) then     /* δ > 0, a small threshold value*/
    n = n – 1
    Merging_Clusters (Clus, n - 1)
End
Else Clus = Clus ∪ {$C_q$} ∪ {$C_p$} – {$C_{pq}$}   /*roll back to obtain previous clusters*/
Return Clus
End.
```

C. Splitting of clusters

There is some possibility that the objects are situated in dissimilar manner within the clusters. Such clusters are known as dissimilar clusters which need to be split into two or more clusters. In splitting process, the intra-cluster (within cluster) similarities are measured for each individual cluster. The measurement is performed with the comparison between cluster center and objects of corresponding cluster. Also inter-cluster (between clusters) similarity between each pair of clusters is measured comparing respective cluster centers. Now the split factor is calculated using Eq. (6) for each individual cluster. In Eq. (6), if numerator value is large and denominator value is small, then the split factor is high which implies that the cluster is compact with respect to its objects and other cluster centers; otherwise, the cluster is scattered. Now a threshold value (γ) is calculated using Eq. (7) to measure the scattering of each individual cluster.

If split factor of any cluster is less than the γ value, then the corresponding cluster is split into three clusters considering centroid of the cluster and two most dissimilar objects within the cluster as their centers. The other objects of the cluster are placed in one of the three newly formed clusters to which they are more similar. The algorithm of splitting process is described in detail below:

```
Algorithm: Splitting_Clusters ($Clus$, $n$)
Input: $Clus = \{C_1, C_{2,} \ldots ,C_n\}$ of $n$ clusters obtained after merging.
Output: Set of clusters in $Clus$.
Begin
    For $i = 1$ to $n$ do
        Calculate $SF_i$ using Eq. (6)
    End
    Calculate $\gamma$ value using Eq. (7)
    /* if split factor of cluster $C_i$ is less than threshold $\gamma$, it splits into three clusters $Cl_1$, $Cl_2$ and $Cl_3$ */
    For $i = 1$ to $n$ do
        If ($SF_i < \gamma$) then
            Let, $\mu_i$ is the center of cluster $C_i$.
            Let, $X_p$ and $X_q$ are the most dissimilar objects of $C_i$ based on bit wise matching.
            $Cl_1 = \{\mu_i\}$
            $Cl_2 = \{X_p\}$
            $Cl_3 = \{X_q\}$
            For $j = 1$ to $|C_i|$ do
                If ($X_j$ is the most similar with $\mu_i$) then
                    $Cl_1 = Cl_1 \cup \{X_j\}$
                Else if ($X_j$ is the most similar with $X_p$) then
                    $Cl_2 = Cl_2 \cup \{X_j\}$
                Else $Cl_3 = Cl_3 \cup \{X_j\}$
                End
                $CLUS = CLUS - \{C_i\} \cup \{Cl_1\} \cup \{Cl_2\} \cup \{Cl_3\}$
        End
    End
End.
```

After splitting procedure, suppose n number of clusters is formed. The proposed method computes some cluster validity indices, such as (i) DB-index, (ii) Dunn-index, (iii) H-index and (iv) SC-index using Eqs. (8), (9), (12) and (13), respectively, of the obtained clusters. If two or more validity indices dominate the previous index values, then merging and splitting procedures are continued; otherwise, current k clusters are considered as stable clusters. The above-mentioned validity indices and their values for cluster validation are described below:

(i) DB-index: The Davies–Bouldin (DB) [10] index is defined in Eq. (8).

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{1\leq i\leq k\,\text{and}\,i\neq j} \left\{ \frac{\sigma_i + \sigma_j}{\mathrm{d}S\left(C_i, C_j\right)} \right\} \tag{8}$$

where k is the number of clusters, σ_i is the average dissimilarity of all patterns in cluster i to their cluster center C_i, σ_j is the average dissimilarity of all patterns in cluster j to their cluster center C_j, and $dS\left(C_i, C_j\right)$ is the dissimilarity of cluster centers C_i and C_j. Small values of DB correspond to clusters that are compact, and whose centers are less similar from each other. The minimum of the DB-index determines the actual number of clusters.

(ii) Dunn's index: The Dunn's index (DN) [26] is defined in Eq. (9).

$$\mathrm{DN} = \min_{1\leq i\leq k} \left\{ \min_{i+1\leq j\leq k} \left\{ \frac{\delta\left(C_i, C_j\right)}{\max\limits_{1\leq l\leq n} \partial\left(C_l\right)} \right\} \right\} \tag{9}$$

where $\delta\left(C_i, C_j\right) = \max\left\{\mathrm{d}S\left(C_i, C_j\right) | x_i \in C_i, x_j \in C_j\right\}$ and $\partial\left(C_l\right) = \min\left\{\mathrm{d}S\left(C_i, C_j\right) | x_i, x_j \in C_i\right\}$. The $S\left(C_i, C_j\right)$ is the dissimilarity of two points of C_i and C_j. The maximum of the DN-index determines the actual number of clusters.

(iii) H-index: The similarity within cluster (SSW) and similarity between clusters (SSB) are defined using Eqs. (10) and (11), respectively.

$$\mathrm{SSW} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j\in c_i} S\left(x_j, C_j\right) \tag{10}$$

where k is the total number of clusters, C_j is the respective cluster centers and n is the total number of objects.

$$\mathrm{SSB} = \frac{1}{n}\sum_{i=1}^{k} n_i S\left(C_i, \bar{X}\right) \tag{11}$$

The Hartigan (H) [50] index is defined by Eq. (12).

$$H = -\log\left(\frac{\mathrm{SSW}}{\mathrm{SSB}}\right) \tag{12}$$

The minimum value of the H-index is determined as the desire number of clusters.

(iv) SC-index: The Silhouette Coefficient (SC) [45] is defined by Eq. (13).

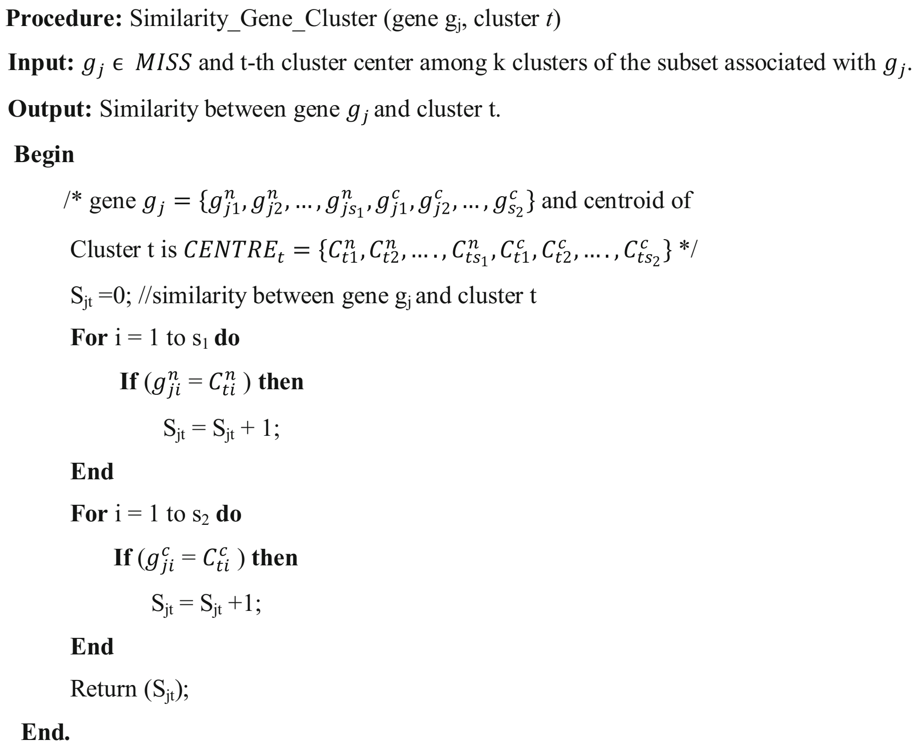
$$SC = \frac{1}{k}\sum_{i=1}^{k} \frac{(b_i - a_i)}{\max\left(a_i, b_i\right)} \tag{13}$$

where a_i is the average dissimilarity of ith objects with all other objects within the same cluster. b_i is the lowest average dissimilarity of ith object to any other cluster. The maximum of the SC-index is determined as the desire number of clusters.

The overall flow diagram of cluster validation method is shown in Fig. 2:

2.4 Similarity measurement

As each gene in the dataset is of s-tuple (i.e., samples), so the centroids of k clusters are also of $s = (s_1 + s_2)$-tuples. Let the centroids of cluster t are $\mathrm{CENTRE}_t = \{C^n_{t1}, C^n_{t2}, \ldots, C^n_{ts_1}, C^c_{t1}, C^c_{t2}, \ldots, C^c_{ts_2}\}$, for $t = 1, 2, \ldots, k$, where C^n_{tj} is the centroid of jth normal samples in cluster t, for $j = 1, 2, \ldots, s_1$ and C^c_{tj} is the centroid of jth cancerous samples of cluster t, for $j = 1, 2, \ldots, s_2$. Now the similarity S_{jt} of gene $g_j \in MISS$ with cluster t is the number of samples matching the values to that of centroid of t. The procedure to measure the similarity of a gene with a cluster is described below:



```
Procedure: Similarity_Gene_Cluster (gene g_j, cluster t)
Input: $g_j \in MISS$ and t-th cluster center among k clusters of the subset associated with $g_j$.
Output: Similarity between gene $g_j$ and cluster t.
 Begin
       /* gene $g_j = \{g^n_{j1}, g^n_{j2}, \ldots, g^n_{js_1}, g^c_{j1}, g^c_{j2}, \ldots, g^c_{s_2}\}$ and centroid of
        Cluster t is $CENTRE_t = \{C^n_{t1}, C^n_{t2}, \ldots., C^n_{ts_1}, C^c_{t1}, C^c_{t2}, \ldots., C^c_{ts_2}\}$ */
        $S_{jt} = 0$; //similarity between gene $g_j$ and cluster t
        For i = 1 to $s_1$ do
            If ($g^n_{ji} = C^n_{ti}$ ) then
                $S_{jt} = S_{jt} + 1$;
        End
        For i = 1 to $s_2$ do
            If ($g^c_{ji} = C^c_{ti}$ ) then
                $S_{jt} = S_{jt} + 1$;
        End
        Return ($S_{jt}$);
  End.
```

Thus, similarity of g_j with all k clusters is obtained and if S_{jP} is maximum for $1 \leq P \leq k$, then the missing g^n_{jq} will be predicted by C^n_{pq}, $1 \leq q \leq s_1$ and missing g^c_{jr}will be estimated by C^c_{pr}, $1 \leq r \leq s_2$. Thus, the missing values are estimated for each gene g_j. The overall algorithm for missing value estimation through cluster analysis (MVETCA) is given below:

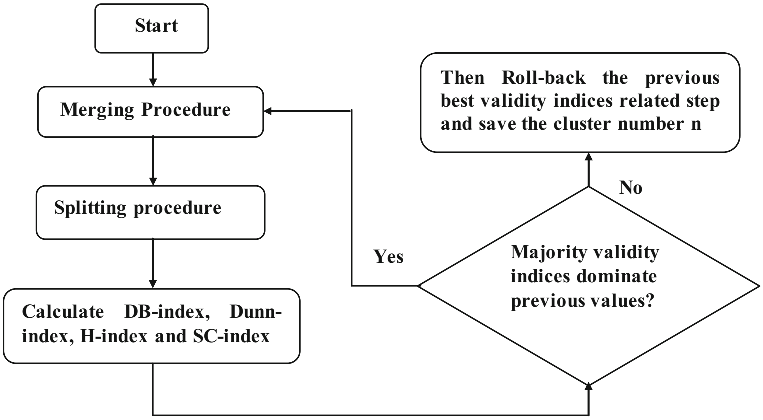

Start
Merging Procedure
Splitting procedure
Calculate DB-index, Dunn-index, H-index and SC-index
Majority validity indices dominate previous values?
Yes
No
Then Roll-back the previous best validity indices related step and save the cluster number n

Fig. 2 Overall flow diagram of cluster validation method

```
Algorithm: MVETCA (MISS, NOMISS)
Input: MISS = Set of genes with missing values; NOMISS = Set of genes without missing values
Output: Complete dataset with estimated missing values
Begin
  The NOMISS is discretized with N number of the discrete values, using Eq. (1) and Eq. (2)
   For each gene in NOMISS do
      Compute $F_1$ = Frequency of discrete value which is the maximum in normal samples
      Compute $F_2$ = Frequency of discrete value which is the maximum in cancerous samples
      If ($f_1$ and $f_2$ occurs for same discrete value) then
            Put whole gene into one of N gene subsets associated with respective discrete value.
      Else, put normal and cancerous part of gene separately into two subsets of 2N gene subsets
   End
  Perform proposed clustering algorithm to find optimal number of clusters for each of 3N gene subsets
  For each gene g in MISS do
     Determine its associated set among 3N gene subsets
     Select cluster center to which gene g has maximum similarity
     Impute missing value of the sample in g by corresponding sample value of the selected center.
 End
End.
```

3 Experimental results and performance evaluation

Experimental studies presented provide an evidence of effectiveness of the proposed MVETCA method on gene expression datasets. Experiments are carried out on different kinds of microarray datasets [20]. Each dataset contains two or more types of samples, nor-

Table 1 Summary of gene expression dataset

Dataset	#Genes	Class name	#Samples (class1/class2)
Leukemia1	7129	ALL/AML	38 (27/11)
Lung cancer	12533	MPM/ADCA	32 (16/16)
Prostate cancer	12600	Tumor/normal	102 (52/50)
Breast cancer	24481	Relapse/non-relapse	78 (34/44)
DLBCL data	6817	DLBCL/FL	77 (58/19)
Colon cancer	2000	Negative/positive	62 (40/22)
Leukemia2	12582	ALL/MLL/AML	57 (20/17/20)

mal and cancerous. The numbers of genes, classes and samples contained in the various datasets are listed in Table 1.

3.1 Efficiency of cluster analysis

To impute missing values, first the correlated gene subsets are identified, then the clustering algorithm is applied on appropriate gene subset and finally the missing values are estimated. To prove the efficiency of clustering algorithm, the algorithm applied on subsets of each dataset, optimal number of clusters are obtained and comparison is made with well-known K-means [35,40] and Fuzzy C-means [13,35] clustering techniques with same number of clusters by some validity indices [10,26,45,50].

3.1.1 Comparative study

In the proposed method, initially all points are treated as individual clusters and after successive iteration of merging and subsequent splitting process, discussed in Sect. 3, optimal number of clusters is obtained. Considering same number of clusters obtained by the proposed method, K-means [35,40] and Fuzzy C-means [13,35] algorithm are applied on the same subsets and a comparison is made on six mentioned gene datasets (for some initial clusters) as listed in Tables 2, 3, 4, 5, 6, 7 and 8. The results show that DB, DN, SC and H indices produced by the proposed method are better than that produced by other methods in most of the cases (show in bolt and shaded font), which confirms the potentiality and superiority of the proposed method.

3.1.2 Optimal cluster selection

The proposed method modifies the clusters iteratively, and finally, based on the validity indices optimal number of clusters is obtained. To demonstrate how the work finds the optimal number of clusters, one group of data (such as GW_Z) for each dataset is selected and graph is plotted against ‘Cluster number’ and ‘validity indices.’ The following figures (Figs. 3, 4, 5, 6, 7, 8, 9) show the validity indices for different number of clusters. In figures, we have considered one group from each dataset and the validity indices are computed for different number of clusters obtained iteratively by our proposed clustering algorithm. It is noted that initially there are large number of clusters, which gradually become stable by splitting and merging procedures.

Table 2 Leukemia1 subsets with 10% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
$GW_L(4)$	**0.450**	**0.387**	**0.539**	**0.058**	0.869	0.319	0.410	0.192	0.629	0.331	0.447	0.171
$GW_Z(6)$	**0.352**	**0.394**	**0.509**	**0.029**	0.676	0.202	0.390	0.076	0.581	0.320	0.399	0.046
$GW_H(7)$	**0.213**	**0.406**	**0.714**	**0.079**	0.544	0.271	0.489	0.216	0.563	0.358	0.583	0.172
$GN_L(7)$	**0.346**	**0.276**	**0.490**	**0.066**	0.753	0.262	0.306	0.100	0.637	0.271	0.375	0.901
$GN_Z(4)$	**0.233**	0.116	**0.398**	0.041	0.561	0.119	0.211	0.088	0.433	**0.125**	0.271	**0.040**
$GN_H(8)$	**0.271**	**0.227**	**0.441**	**0.066**	0.447	0.105	0.391	0.073	0.401	0.197	0.362	0.382
$GC_L(3)$	**0.188**	0.065	**0.227**	0.039	0.280	**0.094**	0.208	0.053	0.205	0.074	0.210	**0.038**
$GC_Z(3)$	0.205	**0.058**	0.251	**0.061**	**0.203**	0.034	0.278	0.098	0.209	0.046	**0.281**	0.065
$GC_H(5)$	**0.176**	**0.106**	**0.303**	0.132	0.197	0.102	0.227	0.134	0.210	0.092	0.292	**0.130**

Table 3 Lung cancer subsets with 20% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
GW_L(3)	**0.582**	**0.275**	**0.590**	**−0.006**	0.621	0.116	0.339	0.221	0.587	0.271	0.404	0.107
GW_Z(12)	**0.131**	**0.526**	**0.883**	**0.002**	0.164	0.394	0.712	0.192	0.162	0.447	0.754	0.062
GW_H(6)	**0.290**	**0.373**	**0.648**	**0.053**	0.439	0.264	0.490	0.093	0.387	0.305	0.507	0.071
GN_L(7)	**0.257**	**0.302**	**0.580**	**0.090**	0.375	0.295	0.419	0.371	0.348	0.289	0.472	0.307
GN_Z(4)	**0.211**	**0.129**	**0.397**	**0.026**	0.338	0.110	0.201	0.373	0.277	0.113	0.291	0.178
GN_H(9)	**0.159**	**0.514**	**0.816**	**0.051**	0.202	0.467	0.621	0.284	0.182	0.504	0.601	0.296
GC_L(8)	0.249	**0.356**	0.567	**−0.250**	**0.248**	0.250	0.553	0.150	0.260	0.283	**0.569**	0.059
GC_Z(9)	**0.198**	0.487	**0.751**	**0.016**	0.199	**0.496**	0.643	0.041	0.210	0.421	0.728	0.027
GC_H(8)	**0.264**	**0.319**	**0.627**	**0.051**	0.302	0.310	0.300	0.082	0.294	0.313	0.410	0.056

Table 4 Prostate cancer subsets with 30% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
GW_L(3)	0.525	0.301	0.631	−**0.003**	0.725	0.221	0.553	0.073	**0.516**	**0.315**	**0.630**	0.063
GW_Z(10)	**0.176**	**0.446**	**0.748**	**0.002**	0.389	0.389	0.539	0.050	0.338	0.402	0.661	0.047
GW_H(5)	**0.269**	**0.413**	**0.661**	**0.005**	0.399	0.402	0.514	0.068	0.401	0.406	0.529	0.032
GN_L(4)	**0.309**	**0.213**	**0.486**	−**0.012**	0.412	0.195	0.390	0.043	0.409	0.200	0.401	0.002
GN_Z(3)	**0.427**	0.305	**0.530**	**0.006**	0.464	0.264	0.428	0.092	0.436	**0.371**	0.500	0.071
GN_H(3)	**0.253**	0.149	**0.332**	−**0.01**	0.361	0.116	0.310	0.011	0.370	**0.141**	0.326	0.001
GC_L(3)	0.367	**0.300**	**0.407**	**0.0017**	0.377	0.174	0.229	0.037	**0.358**	0.276	0.372	0.016
GC_Z(4)	**0.356**	0.313	0.563	**0.021**	0.393	0.300	0.533	0.041	0.382	**0.320**	**0.566**	0.036
GC_H(4)	0.225	**0.399**	**0.608**	**0.002**	**0.224**	0.388	0.591	0.064	0.228	0.379	0.603	0.008

Table 5 Breast cancer subsets with 15% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
GW_L(5)	**0.610**	**0.631**	**0.734**	**0.036**	0.732	0.422	0.555	0.109	0.701	0.501	0.580	0.083
GW_Z(7)	**0.684**	**0.581**	**0.702**	**0.018**	0.733	0.371	0.632	0.053	0.690	0.445	0.640	0.048
GW_H(9)	**0.743**	**0.720**	0.629	0.132	0.806	0.583	0.584	0.266	0.788	0.589	**0.703**	**0.125**
GN_L(8)	**0.528**	**0.457**	**0.594**	**0.092**	0.791	0.299	0.485	0.243	0.715	0.382	0.466	0.105
GN_Z(5)	0.493	0.391	0.760	**0.005**	0.529	**0.420**	0.721	0.017	**0.483**	0.381	**0780**.	**0.005**
GN_H(11)	**0.732**	**0.562**	0.800	**0.201**	0.739	0.410	0.655	0.308	0.786	0.433	**0.803**	0.252
GC_L(3)	0.470	0.692	0.640	**0.073**	**0.350**	**0.696**	**0.711**	0.076	0.402	0.650	0.651	0.101
GC_Z(4)	0.447	0.364	0.558	**0.008**	**0.445**	**0.511**	**0.620**	0.063	0.471	0.483	0.581	0.081
GC_H(6)	**0.528**	**0.588**	**0.722**	**0.107**	0.682	0.517	0.643	0.200	0.580	0.500	0.705	0.191

Table 6 DLBCL subsets with 10% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
$GW_L(5)$	**0.332**	0.049	**0.407**	**0.004**	0.632	**0.068**	0.340	0.073	0.459	0.057	0.389	0.057
$GW_Z(11)$	**0.266**	**0.113**	**0.591**	**0.009**	0.539	0.092	0.390	0.095	0.551	0.106	0.418	0.076
$GW_H(5)$	**0.139**	0.061	**0.391**	**0.031**	0.380	0.062	0.212	0.050	0.309	**0.080**	0.229	0.044
$GN_L(4)$	**0.205**	0.055	**0.383**	**0.033**	0.333	**0.074**	0.266	0.073	0.224	0.063	0.294	0.058
$GN_Z(3)$	**0.207**	**0.116**	**0.474**	**0.021**	0.378	0.112	0.372	0.090	0.284	0.111	0.283	0.069
$GN_H(3)$	0.202	**0.064**	**0.311**	**0.031**	0.308	0.052	0.252	0.063	**0.200**	0.059	0.303	0.056
$GC_L(3)$	0.290	**0.060**	0.331	**0.020**	0.301	0.035	0.201	0.038	**0.283**	0.046	**0.334**	**0.020**
$GC_Z(4)$	**0.111**	0.063	0.202	**0.011**	0.137	0.071	**0.212**	0.031	0.142	**0.082**	0.208	0.035
$GC_H(4)$	**0.104**	0.057	**0.228**	**0.024**	0.156	0.054	0.221	0.057	0.133	**0.060**	0.225	0.048

Table 7 Colon subsets with 20% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
GW_L(3)	**0.201**	**0.720**	**0.544**	**0.037**	0.572	0.543	0.309	0.106	0.462	0.630	0.401	0.102
GW_Z(8)	**0.351**	**0.638**	**0.600**	**0.107**	0.476	0.210	0.591	0.322	0.483	0.522	0.550	0.203
GW_H(9)	**0.422**	**0.649**	**0.731**	0.119	0.491	0.549	0.509	0.132	0.473	0.622	0.501	**0.110**
GN_L(7)	**0.299**	**0.822**	0.730	**0.274**	0.433	0.581	0.640	0.382	0.421	0.788	**0.738**	0.288
GN_Z(4)	0.400	0.717	**0.739**	**0.352**	0.389	0.476	0.633	**0.352**	**0.372**	**0.781**	**0.739**	0.355
GN_H(7)	**0.511**	**0.530**	**0.629**	**0.222**	0.683	0.770	0.538	0.300	0.621	0.511	0.599	0.300
GC_L(3)	0.250	**0.493**	**0.582**	0.316	0.290	0.263	0.499	0.417	**0.249**	0.428	**0.582**	**0.310**
GC_Z(4)	0.278	0.642	**0.721**	**0.042**	0.370	0.374	0.666	0.092	**0.270**	**0.691**	0.633	0.151
GC_H(6)	**0.390**	**0.582**	**0.501**	**0.061**	0.461	0.439	0.395	0.118	0.400	0.550	0.489	0.114

Table 8 Leukemia2 subsets with 5% missing values

Dataset (#cluster)	Merging-splitting algorithm				K-means				C-means			
	DB	DN	SC	H	DB	DN	SC	H	DB	DN	SC	H
GW $_L$(4)	**0.404**	**0.759**	**0.630**	**0.083**	0.618	0.696	0.583	0.174	0.522	0.721	0.607	0.117
GW $_Z$(7)	**0.483**	**0.620**	**0.609**	**0.194**	0.529	0.518	0.437	0.222	0.502	0.580	0.482	0.206
GW $_H$(3)	**0.270**	**0.738**	**0.492**	**0.008**	0.397	0.669	0.448	0.096	0.299	0.725	0.460	0.026
GC1$_L$(5)	0.533	0.505	**0.595**	**0.065**	0.529	0.381	0.503	0.194	**0.526**	**0.510**	0.583	0.142
GC1$_Z$(4)	**0.218**	**0.561**	**0.674**	**0.027**	0.384	0.480	0.585	0.107	0.306	0.538	0.580	0.084
GC1$_H$(3)	**0.095**	0.706	0.478	**0.111**	0.227	0.677	0.439	0.121	0.284	**0.710**	**0.481**	0.130
GC2$_L$(7)	**0.290**	**0.693**	**0.538**	0.174	0.366	0.605	0.521	0.190	0.328	0.640	0.525	**0.170**
GC2$_Z$(9)	**0.185**	**0.500**	**0.444**	**0.106**	0.196	0.420	0.376	0.242	0.191	0.417	0.417	0.182
GC2$_H$(5)	**0.200**	**0.718**	**0.389**	**0.087**	0.287	0.653	0.360	0.184	0.256	0.630	0.370	0.110
GC3$_L$(3)	**0.420**	**0.675**	**0.641**	**0.170**	0.464	0.607	0.512	0.233	0.427	0.629	0.559	0.227
GC3$_Z$(4)	**0.386**	0.492	0.350	**0.253**	0.406	**0.503**	**0.354**	0.271	0.397	0.475	0.339	0.262
GC3$_H$(5)	**0.302**	**0.539**	**0.511**	**0.206**	0.385	0.501	0.473	0.294	0.337	0.521	0.491	0.240

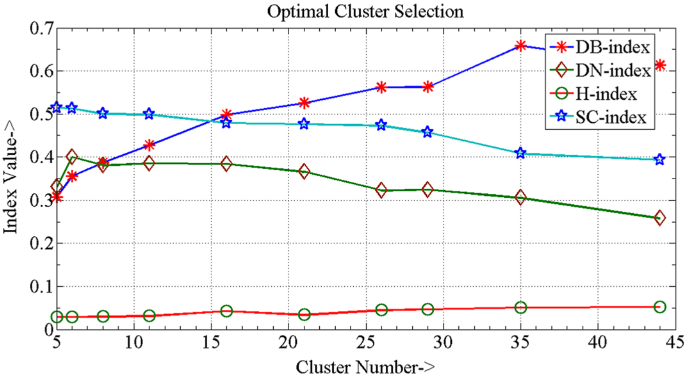


Fig. 3 Optimal no. of cluster for GW_Z set of Leukemia1 dataset

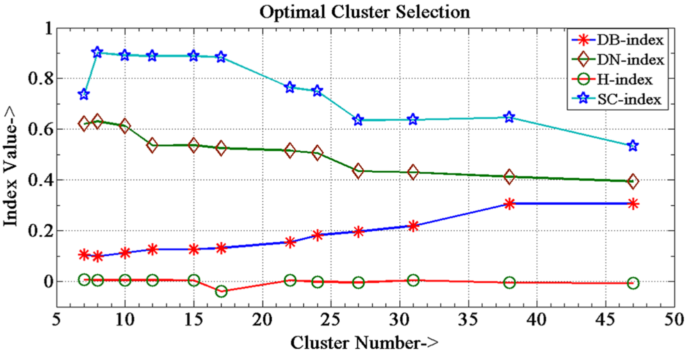


Fig. 4 Optimal no. of cluster for GC_L set of Lung cancer dataset

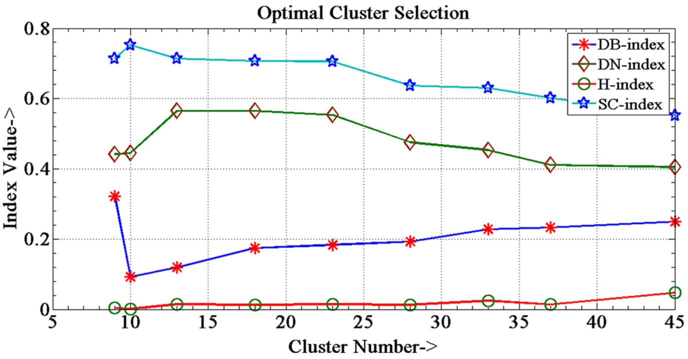


Fig. 5 Optimal no. of cluster for GW_Z set Prostate cancer dataset

In all diagrams (Figs. 3, 4, 5, 6, 7, 8, 9), it is observed that DB and H value decreases and DN and SC value increases up to a certain number of clusters. According to proposed clustering algorithm, in any iteration if two or more validity indices dominate their previous values, then procedure is continued, and otherwise, the clusters used in this iteration are considered as the optimal clusters. For example, in case of GW_Z set of Leukemia1 dataset, DN-index, H-index and SC-index dominate their previous values (as shown in Fig. 3) in the

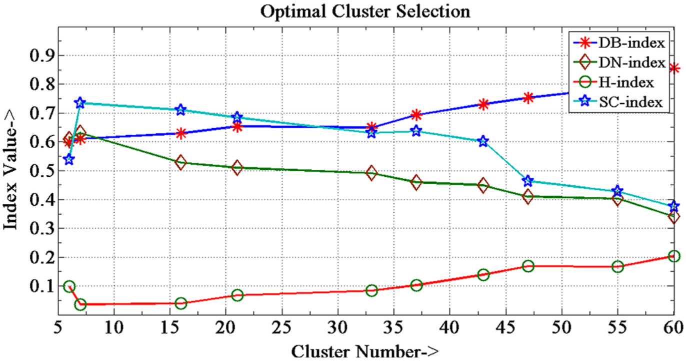


Fig. 6 Optimal no. of cluster for GW_Z set of Breast cancer dataset

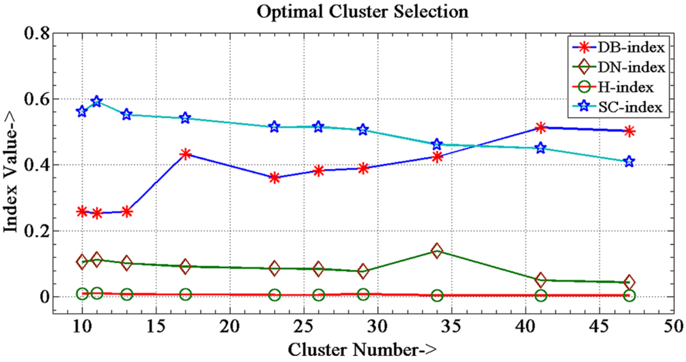


Fig. 7 Optimal no. of cluster GW_Z set for DLBCL dataset

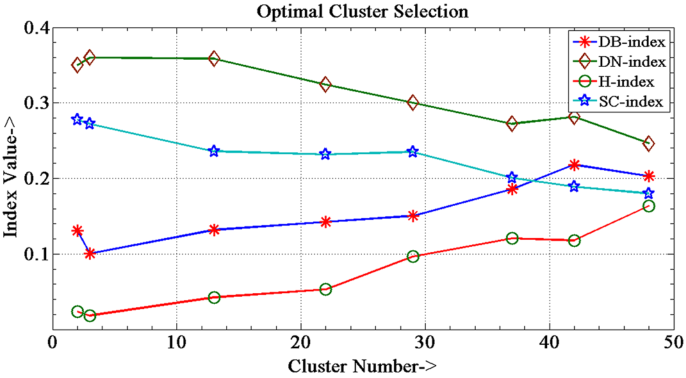


Fig. 8 Optimal no. of cluster for GW_L set of Colon cancer dataset

iteration with 35 clusters, and so the process is continued. Finally the process terminates with six clusters when DN-index, H-index and SC-index dominate their previous values. Similar analysis is made for other datasets using Figs. 4, 5, 6, 7, 8, 9 to visualize optimal number of clusters obtained by the algorithm, as given in Tables 2, 3, 4, 5, 6, 7 and 8.

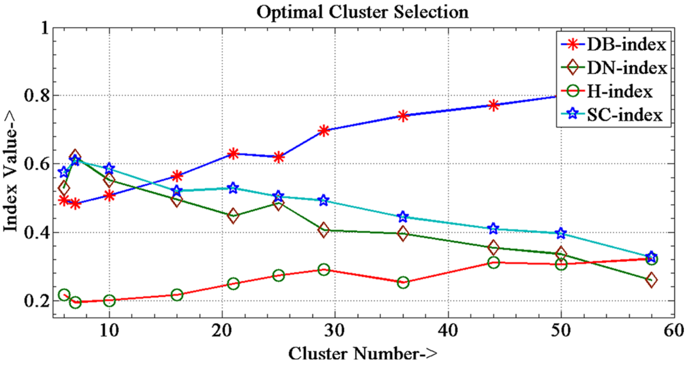


Fig. 9 Optimal no. of cluster for GW_L set of Leukemia2 dataset

3.2 Performance analysis

Experimental results presented here provide an evidence of effectiveness of the proposed MVETCA algorithm on publicly available benchmark microarray datasets. The microarray dataset is divided into two subsets and in one set randomly missing values are provided at any random positions. Thus, two datasets *NOMISS* and *MISS* are formed from a given dataset. Then the missing values of genes in *MISS* are imputed by the proposed method using the genes in *NOMISS*.

In cluster-based methods, the number of nearest neighbors, K, must be selected. In KNN-impute, SKNN-impute and IKNN-impute, value of K is set by a value in between 10 and 20 and the best value of K is obtained. Here, we have decided to perform multiple estimation tests incorporating 5, 10, 15, 20, 25 and 30% missing values in the datasets.

3.2.1 Evaluation criteria

The performance of our missing values imputation algorithm is evaluated by three metrics such as 'Normalized Root Mean Squared Error,' 'Conserved Pair Proportions' and 'Biomarker List Concordance Index.'

(a) Statistical index

We use the normalized root mean squared error (NRMSE) [39], a statistical index to evaluate the performance of the proposed and existing missing values estimation techniques, defined in Eq. (14). Lower the value of the NRMSE, better the method performs.

$$\text{NRMSE} = \frac{1}{\text{std\_dev}\left(X_{\text{known}}\right)}\sqrt{\frac{\sum_{i=1}^{n}\left(X_{\text{predict}} - X_{\text{known}}\right)^2}{X}} \tag{14}$$

where X_{known} is the original gene expression value and X_{predict} is the estimated value obtained by the proposed algorithm, std_dev (X_{known}) is the standard deviation of original expression values and X is the total number of missing values. The number X is set randomly as 5, 10, 15, 20, 25 and 30% of total genes, and *NRMSE* is computed in all methods.

(b) Clustering index

Conserved pairs proportion *(CPP)* [11] one of the most promising clustering indices is used to evaluate the stability of the clusters against the missing values present in the genes. Here, the proposed ‘splitting and merging’ clustering algorithm is applied on the original dataset and the obtained clusters are say, OC $= \{C_{\mathrm{O}1}, C_{\mathrm{O}2}, \ldots, C_{\mathrm{Ok}}\}$. Also, the same clustering algorithm is applied on dataset with generating and estimating the missing values and let the obtained clusters are, EC $= \{C_{\mathrm{E}1}, C_{\mathrm{E}2}, \ldots, C_{\mathrm{El}}\}$. The CPP index computed using Eq. (15) is used to compute the percentage of genes obtained associated with the clusters in OC and EC.

$$\mathrm{CPP} = \frac{\sum_{i=1}^{K}\left(\max\limits_{1\le j\le l}\left(\sum_{x\in C_{Oi}}\sum_{y\in C_{Ej}}\delta_{xy}\right)\right)}{n} \tag{15}$$

where n is the total number of genes, $\delta_{xy} = 1$, if the genes x and y are identical and 0 otherwise. Higher the value of the CPP index, better the method performs.

(c) Differentially expressed genes index

We have used the biologically meaningful metric presented in [42] to express the biological impact of missing value imputation in gene datasets. The method [42] identifies differentially expressed genes for the original dataset and the imputed dataset. Then the biomarker list concordance index (BLCI) [29], defined in Eq. (16), is computed to evaluate the performance of different imputing methods.

$$\mathrm{BLCI} = \frac{n\left(D_{\mathrm{OD}} \cap D_{\mathrm{ID}}\right)}{n\left(D_{\mathrm{OD}}\right)} + \frac{n\left(D_{\mathrm{OD}}^{C} \cap D_{\mathrm{ID}}^{C}\right)}{n\left(D_{\mathrm{OD}}^{C}\right)} - 1 \tag{16}$$

where D_{OD} and D_{ID} are the significantly differentially expressed genes in the *original dataset* (*OD*) and the *imputed dataset* (ID). D_{OD}^{C} and D_{ID}^{C} are the complement set of the D_{OD} and D_{ID}, respectively, and $n\,(*)$ is the number of genes. A high BLCI value indicates that the list of the significantly differentially expressed genes of the OD is similar to that of the ID and the method is of high performance.

3.2.2 Comparative study

The performance of MVETCA is compared with statistical methods (such as Zero-impute [1], Row-average [38], SVD-impute [19] and BPCA-impute [28]), cluster-based methods (such as KNN-impute [41], SKNN-impute [21], IKNN-impute [3] and HCS-impute [32], bi-BPCA-impute [27]), and regression-based methods (LLS-impute [22], SLLS- impute [46,47], ILLS-impute [6] and TRIIM [16]).

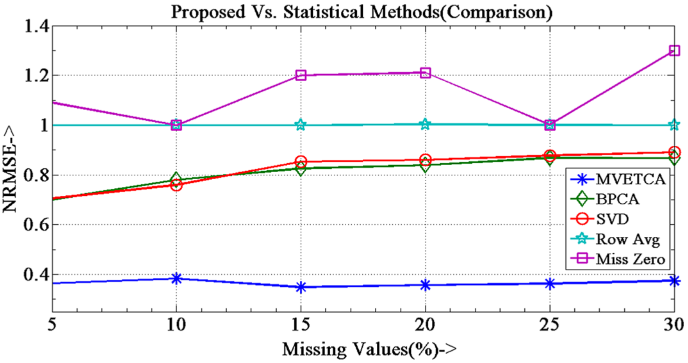


Fig. 10 NRMSE values for Leukemia1 dataset

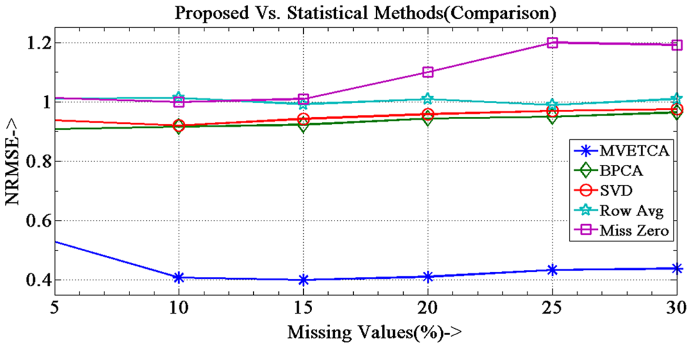


Fig. 11 NRMSE values for Lung cancer dataset

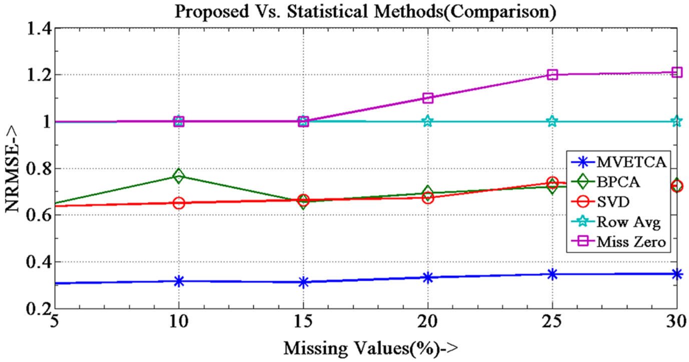


Fig. 12 NRMSE values for Prostate cancer dataset

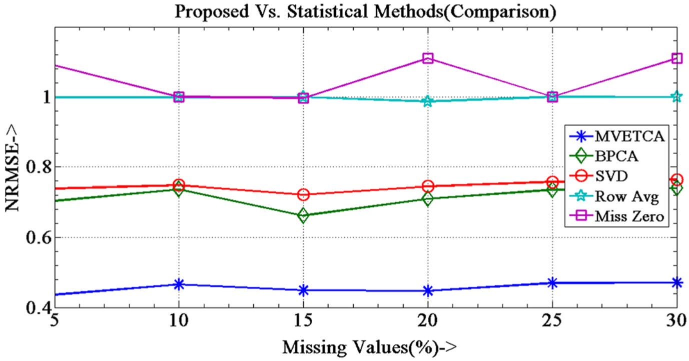


Fig. 13 NRMSE values for Breast cancer dataset

(a) Statistical methods versus MVETCA
In Figs. 10, 11, 12, 13, 14, 15 and 16, the NRMSE values are plotted for various imputed datasets obtained by the proposed method MVETCA and well-known statistical methods, such as Zero-impute, Row-average, SVD and BPCA-impute.

From the figures (Figs. 10, 11, 12, 13, 14, 15, 16), it is observed that MVETCA gives better results (i.e., minimum NRMSE) compare to other methods, which confirms the potentiality

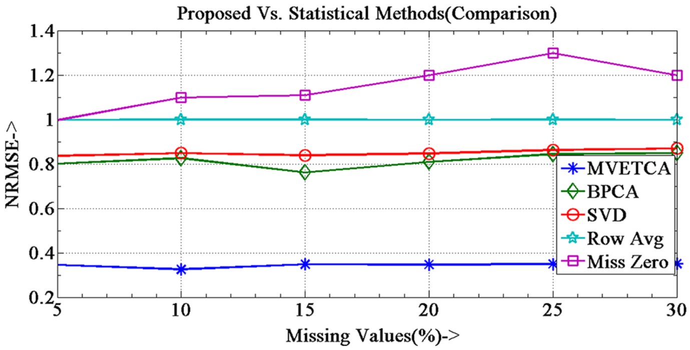


Fig. 14 NRMSE values for DLBCL dataset

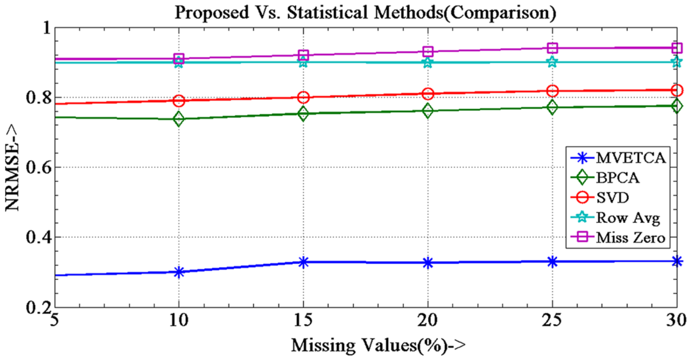


Fig. 15 NRMSE values for Colon cancer dataset

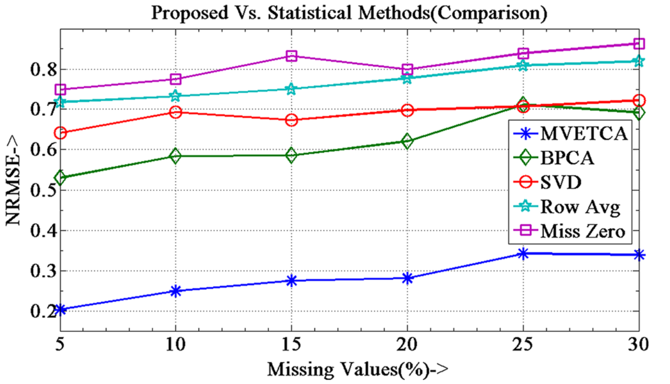


Fig. 16 NRMSE values for Leukemia2 dataset

and superiority of the proposed method. Also, two popular biologically significant metrics such as CPP and BLCI are computed for all the considered datasets, as listed in Table 9. The result shows that for different percentage of missing values imputation, the proposed method always outperforms the others.

Table 9 CPP and LCI values for experimental datasets

Dataset	Methods	CPP						BLCI					
		Missing values (%)						Missing values (%)					
		5	10	15	20	25	30	5	10	15	20	25	30
Leukemia1	MissZ	0.5569	0.5639	0.5395	0.5344	0.5486	0.5022	0.6009	0.5743	0.5492	0.5264	0.5432	0.5826
	RowA	0.6060	0.5854	0.5618	0.5551	0.5273	0.5175	0.6080	0.5849	0.5729	0.5390	0.6021	0.5811
	SVD	0.6775	0.6677	0.6593	0.6610	0.6467	0.6373	0.6882	0.6744	0.6818	0.6700	0.6490	0.6527
	BPCA	0.7897	0.7768	0.7728	0.7603	0.7669	0.7462	0.7163	0.7016	0.6538	0.6899	0.7004	0.6755
	MVETCA	**0.9746**	**0.9707**	**0.9679**	**0.9603**	**0.9588**	**0.9442**	**0.8462**	**0.8057**	**0.8247**	**0.8340**	**0.8102**	**0.8011**
Lung	MissZ	0.5951	0.5983	0.5746	0.5782	0.5737	0.5551	0.5906	0.6182	0.5720	0.5748	0.5690	0.5773
	RowA	0.6103	0.6065	0.6015	0.5976	0.5842	0.5725	0.6352	0.6471	0.6290	0.6400	0.6210	0.6038
	SVD	0.6880	0.6658	0.6534	0.6135	0.6092	0.6007	0.7268	0.7109	0.7187	0.7035	0.6930	0.6913
	BPCA	0.6970	0.6940	0.6880	0.6840	0.6711	0.6794	0.7465	0.7381	0.7267	0.7328	0.7184	0.7277
	MVETCA	**0.9540**	**0.9505**	**0.9451**	**0.9294**	**0.9163**	**0.9053**	**0.8630**	**0.8476**	**0.8590**	**0.8411**	**0.8362**	**0.8246**
Prostate	MissZ	0.5794	0.5779	0.5786	0.5722	0.5710	0.5556	0.5120	0.5216	0.5028	0.4908	0.5050	0.4929
	RowA	0.6253	0.6217	0.6185	0.6183	0.6027	0.6079	0.6151	0.6108	0.6050	0.5837	0.5639	0.5724
	SVD	0.8309	0.8164	0.8231	0.8111	0.8056	0.8042	0.6234	0.6200	0.6145	0.6046	0.6199	0.6132
	BPCA	0.8537	0.8372	0.8217	0.8190	0.8121	0.8067	0.6210	0.6134	0.6259	0.6290	0.6069	0.6004
	MVETCA	**0.9828**	**0.9649**	**0.9420**	**0.9387**	**0.9311**	**0.9174**	**0.8092**	**0.8043**	**0.7958**	**0.7960**	**0.8024**	**0.7868**
Breast	MissZ	0.6134	0.6101	0.6017	0.5782	0.5823	0.5681	0.6794	0.6672	0.6487	0.6409	0.6388	0.6319
	RowA	0.6672	0.6643	0.6613	0.6427	0.6490	0.6323	0.6825	0.6834	0.6730	0.6512	0.6602	0.6420
	SVD	0.7548	0.7582	0.7329	0.7097	0.7122	0.7153	0.7170	0.7053	0.7148	0.6903	0.6935	0.6922
	BPCA	0.8231	0.8163	0.8177	0.8043	0.7955	0.7901	0.7784	0.7690	0.7538	0.7521	0.7540	0.7483
	MVETCA	**0.9633**	**0.9630**	**0.9609**	**0.9563**	**0.9503**	**0.9471**	**0.9039**	**0.9027**	**0.8827**	**0.8740**	**0.8640**	**0.8701**

Table 9 continued

Dataset	Methods	CPP						BLCI					
		Missing values (%)						Missing values (%)					
		5	10	15	20	25	30	5	10	15	20	25	30
DLBCL	MissZ	0.6890	0.6811	0.6509	0.6573	0.6441	0.6306	0.6453	0.6373	0.6389	0.6218	0.6231	0.6148
	RowA	0.7048	0.7021	0.6940	0.6764	0.6710	0.6744	0.6653	0.6732	0.6540	0.6444	0.6503	0.6417
	SVD	0.8964	0.8561	0.8601	0.8533	0.8216	0.8390	0.7690	0.7560	0.7738	0.7519	0.7309	0.7344
	BPCA	0.9010	0.9027	0.8869	0.8602	0.8566	0.8261	0.7564	0.7431	0.7579	0.7364	0.7400	0.7310
	MVETCA	**0.9846**	**0.9821**	**0.9754**	**0.9658**	**0.9611**	**0.9539**	**0.8255**	**0.8046**	**0.8143**	**0.8128**	**0.8080**	**0.8001**
Colon	MissZ	0.7241	0.7010	0.7190	0.6917	0.6549	0.6391	0.6861	0.6754	0.6721	0.6634	0.6290	0.6271
	RowA	0.7650	0.7543	0.7472	0.7103	0.7058	0.7070	0.7243	0.7266	0.7109	0.7200	0.7122	0.7007
	SVD	0.8659	0.8590	0.8354	0.8301	0.8297	0.8164	0.8253	0.8197	0.8135	0.8105	0.8079	0.8014
	BPCA	0.9143	0.9105	0.9042	0.9070	0.8720	0.8633	0.8524	0.8629	0.8502	0.8373	0.8345	0.8319
	MVETCA	**0.9803**	**0.9853**	**0.9766**	**0.9729**	**0.9532**	**0.9468**	**0.9464**	**0.9475**	**0.9384**	**0.9322**	**0.9163**	**0.9110**
Leukemia2	MissZ	0.7032	0.7053	0.6905	0.6734	0.6723	0.6680	0.6731	0.6698	0.6630	0.6604	0.6580	0.6522
	RowA	0.7397	0.7329	0.7300	0.7267	0.7254	0.7054	0.7194	0.7105	0.7080	0.7014	0.7028	0.6876
	SVD	0.8375	0.8390	0.8265	0.8268	0.8219	0.8196	0.8306	0.8153	0.809	0.8080	0.8012	0.8018
	BPCA	0.8564	0.8513	0.8500	0.8483	0.8414	0.8392	0.8234	0.8210	0.8141	0.8167	0.8076	0.7928
	MVETCA	**0.9875**	**0.9821**	**0.9798**	**0.9726**	**0.9621**	**0.9609**	**0.9537**	**0.9429**	**0.9448**	**0.9320**	**0.9303**	**0.9285**

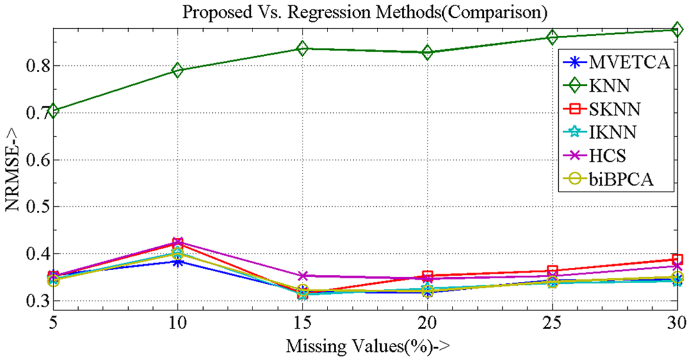


Fig. 17 NRMSE values for Leukemia1 dataset

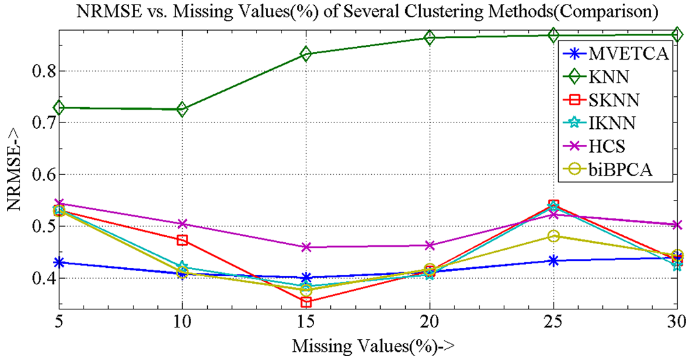


Fig. 18 NRMSE values for Lung cancer dataset

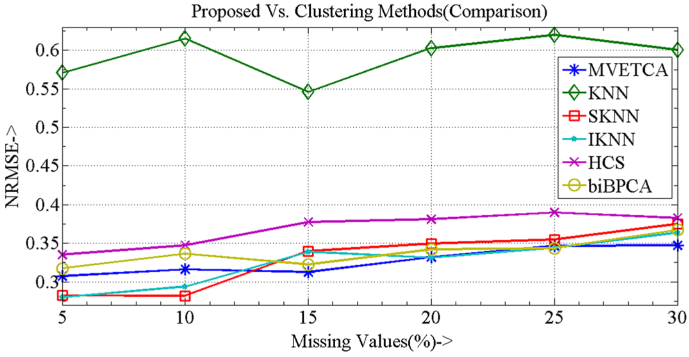


Fig. 19 NRMSE values for Prostate cancer dataset

(b) Cluster-based methods versus MVETCA

In Figs. 17, 18, 19, 20, 21, 22 and 23, the NRMSE values are plotted for various imputed datasets obtained by the proposed method MVETCA and some cluster-based methods, such as KNN-impute, SKNN-impute, IKNN-impute, HCS-impute and bi-BPCA.

From the figures (Figs. 17, 18, 19, 20, 21, 22, 23), it is observed that MVETCA gives better results (i.e., minimum *NRMSE*) compare to other methods in most of the cases, which con-

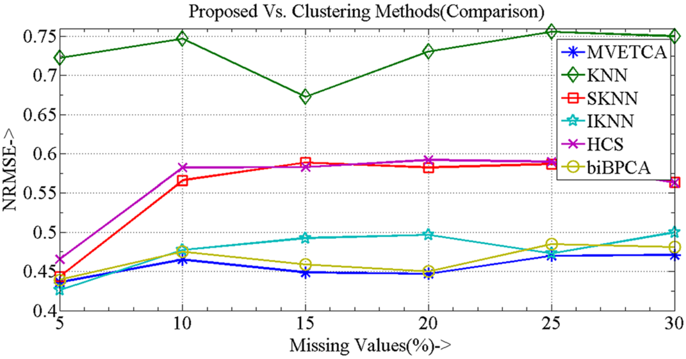


Fig. 20 NRMSE values for Breast cancer dataset

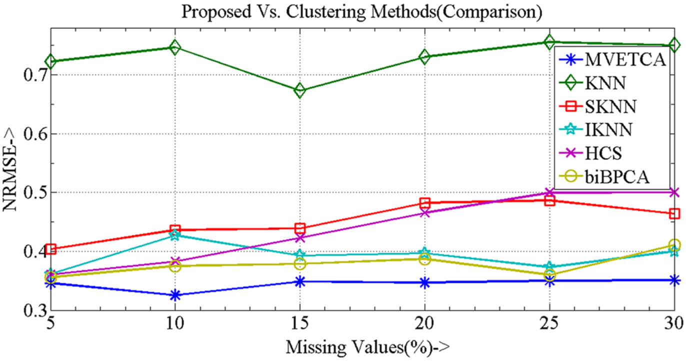


Fig. 21 NRMSE values for DLBCL dataset

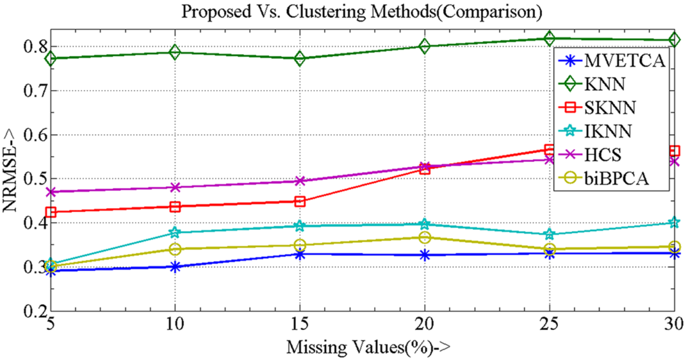


Fig. 22 NRMSE values for Colon cancer dataset

firms the potentiality and superiority of the proposed method. In some cases, like Leukemia1 dataset the IKNN method gives better result for 30% missing values and in Lung cancer dataset the SKNN and IKNN methods give better result for 15% missing values, but for rest of the cases the MVETCA gives the best results. The outstanding estimation ability of MVETCA is due to the accurate grouping of correlated genes, clustering of genes in stable and optimal way. Also, two popular biologically significant metrics such as *CPP* and *BLCI*

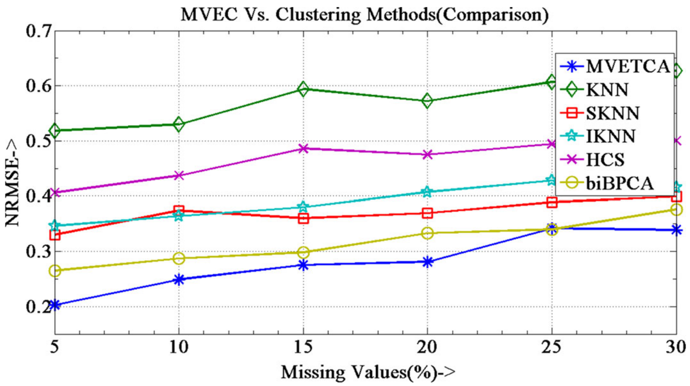


Fig. 23 NRMSE values for Leukemia2 dataset

are computed for all the considered datasets, as listed in Table 10. The result shows that for different percentage of missing values imputation, the proposed method outperforms the others in most of the cases.

(c) Regression-based methods versus MVETCA
In Figs. 24, 25, 26, 27, 28, 29 and 30, the *NRMSE* values are plotted for various imputed datasets obtained by the proposed method MVETCA and some regression-based methods, such as LLS-impute, SLLS-impute, ILLS-impute and TRIIM.

From the figures (Figs. 24, 25, 26, 27, 28, 29, 30), it is observed that MVETCA gives better results (i.e., minimum *NRMSE*) compare to other methods in most of the cases. Also, two popular biologically significant metrics such as *CPP* and *BLCI* are computed for all the considered datasets, as listed in Table 11.

Thus in general, the proposed method is evaluated estimating missing values by various performance evaluation indices and the results obtained by the proposed method outperform other statistical, cluster-based and regression-based methods, which shows the effectiveness of the proposed method.

4 Conclusion

Missing values can bring lots of complications in microarray data analysis because most of the existing methods are designed without any technique for handling them. But missing value estimation is one of the most significant preprocessing steps to deal with the missing values for further experiments. The existing cluster-based methods such as PCM, HCS, KNN, SKNN, IKNN, bi-BPCA and TRIIM estimate more erroneous values compare to the proposed MVETCA method. The PCM and HCS compute similarity between the missing gene and all other *NOMISS* genes and in time of imputation PCM takes information from only one gene, whereas HCS takes from eight neighbor genes. The KNN, SKNN and IKNN use different *K* values and taking best results among them. In time of estimation, all genes are not participated for all the above methods (except IKNN, bi-BPCA and TRIIM) and traditional Euclidian distance metric is used for similarity measurement, which is inefficient in case of high-dimensional datasets.

Table 10 *CPP* and *BLCI* values for experimental datasets

Dataset	Methods	CPP						BLCI					
		Missing values (%)						Missing values (%)					
		5	10	15	20	25	30	5	10	15	20	25	30
Leukemia1	KNN	0.7374	0.7380	0.7256	0.7025	0.7276	0.6927	0.6902	0.6837	0.6810	0.6754	0.6625	0.6601
	SKNN	0.9018	0.9104	0.8867	0.8697	0.8712	0.8666	0.8278	**0.8172**	0.8133	0.8051	0.8002	0.8097
	IKNN	0.8962	0.8886	0.8837	0.8754	0.8587	0.8428	0.8192	0.8134	0.8032	0.7850	0.7938	0.7811
	HCS	0.8809	0.8711	0.8683	0.8193	0.8052	0.7987	0.7827	0.7643	0.7690	0.7718	0.7516	0.7502
	bi-BPCA	**0.9791**	0.9599	0.9546	0.9581	0.9439	0.9186	0.8351	0.8151	0.8119	0.8037	0.8058	**0.8015**
	MVETCA	0.9746	**0.9707**	**0.9679**	**0.9603**	**0.9588**	**0.9442**	**0.8462**	0.8057	**0.8247**	**0.8340**	**0.8102**	0.8011
Lung	KNN	0.6897	0.6809	0.6792	0.6766	0.6631	0.6543	0.7520	0.7439	0.7383	0.7277	0.7315	0.7270
	SKNN	0.8343	0.8324	0.8237	0.8202	0.8139	0.7978	0.8429	0.8452	0.8320	0.8238	0.8212	0.8093
	IKNN	0.8217	0.7963	0.7953	0.7851	0.7819	0.7835	0.8256	0.8109	0.8170	0.8061	0.7968	0.7942
	HCS	0.7692	0.7619	0.7526	0.7516	0.7321	0.7181	0.8071	0.8055	0.7811	0.7950	0.7836	0.7719
	bi-BPCA	0.9503	**0.9515**	**0.9487**	0.9287	0.9078	0.8902	0.8609	**0.8587**	0.8522	0.8397	0.8317	**0.8284**
	MVETCA	**0.9540**	0.9505	0.9451	**0.9294**	**0.9163**	**0.9053**	**0.8630**	0.8476	**0.8590**	**0.8411**	**0.8362**	0.8246
Prostate	KNN	0.8235	0.8211	0.8125	0.8047	0.7932	0.7901	0.6314	0.6306	0.6287	0.6221	0.6298	0.6182
	SKNN	0.9699	0.9662	0.9521	0.9243	0.9210	0.9181	0.7853	0.7790	0.7728	0.7702	0.7535	0.7520
	IKNN	0.9326	0.9216	0.9009	0.9071	0.8957	0.8746	0.7910	0.7819	0.7707	0.7700	0.7614	0.7504
	HCS	0.9376	0.9322	0.9256	0.9273	0.9199	0.9025	0.8003	0.7824	0.7748	0.7804	0.7539	0.7510
	bi-BPCA	0.9811	**0.9727**	0.9416	0.9315	0.9300	0.9085	0.8059	0.8020	**0.7980**	0.7839	0.7803	0.7655
	MVETCA	**0.9828**	0.9649	**0.9420**	**0.9387**	**0.9311**	**0.9174**	**0.8092**	**0.8043**	0.7958	**0.7960**	**0.8024**	**0.7868**
Breast	KNN	0.7832	0.7815	0.7745	0.7722	0.7641	0.7600	0.7420	0.7409	0.7386	0.7327	0.7243	0.7121
	SKNN	0.9032	0.9057	0.8964	0.8902	0.8756	0.8811	0.8733	0.8629	0.8617	0.8600	**0.8650**	0.8563
	IKNN	0.9061	0.8941	0.8848	0.8531	0.8570	0.8417	0.8620	0.8205	0.8173	0.8116	0.8053	0.8044
	HCS	0.9235	0.9103	0.9045	0.9084	0.8756	0.8573	0.7832	0.7616	0.7540	0.7508	0.7500	0.7622

Table 10 continued

Dataset	Methods	CPP						BLCI					
		Missing values (%)						Missing values (%)					
		5	10	15	20	25	30	5	10	15	20	25	30
	bi-BPCA	0.9560	0.9545	0.9531	0.9427	0.9417	0.9320	0.8942	0.8730	0.8791	**0.8761**	0.8664	0.8552
	MVETCA	**0.9633**	**0.9630**	**0.9609**	**0.9563**	**0.9503**	**0.9471**	**0.9039**	**0.9027**	**0.8827**	0.8740	0.8640	**0.8701**
DLBCL	KNN	0.8393	0.8162	0.8255	0.8110	0.8010	0.7846	0.7240	0.7209	0.7185	0.6942	0.7034	0.6821
	SKNN	0.9376	0.9303	0.9158	0.9216	0.9030	0.9004	0.8163	**0.8157**	0.8106	0.8027	0.7929	0.7922
	IKNN	0.9256	0.9254	0.9105	0.9053	0.8807	0.8903	0.7918	0.7889	0.7804	0.7700	0.7742	0.7619
	HCS	0.9265	0.9210	0.9188	0.8902	0.8962	0.8611	0.8085	0.8038	0.7841	0.7720	0.7510	0.7418
	bi-BPCA	0.9749	0.9768	0.9710	0.9582	0.9364	0.9272	**0.8276**	0.8150	**0.8162**	0.8011	0.7999	0.7953
	MVETCA	**0.9846**	**0.9821**	**0.9754**	**0.9658**	**0.9611**	**0.9539**	0.8255	0.8046	0.8143	**0.8128**	**0.8080**	**0.8001**
Colon	KNN	0.8836	0.8671	0.8534	0.8352	0.8460	0.8400	0.8366	0.8202	0.8269	0.8190	0.8111	0.8118
	SKNN	0.9603	0.9549	0.9644	0.9401	0.9348	0.9117	0.8504	0.8562	0.8390	0.8421	0.8322	0.8298
	IKNN	0.9450	0.9381	0.9322	0.9145	0.9102	0.9073	0.8430	0.8274	0.8321	0.8218	0.8200	0.8126
	HCS	0.9535	0.9524	0.9382	0.9291	0.9160	0.9117	0.8379	0.8355	0.8264	0.8212	0.8134	0.8080
	bi-BPCA	**0.9806**	0.9757	0.9639	0.9622	0.9470	0.9441	0.9291	0.9237	0.9254	0.9201	**0.9185**	**0.9122**
	MVETCA	0.9803	**0.9853**	**0.9766**	**0.9729**	**0.9532**	**0.9468**	**0.9464**	**0.9475**	**0.9384**	**0.9322**	0.9163	0.9110
Leukemia2	KNN	0.7842	0.7877	0.7730	0.7664	0.7610	0.7593	0.7970	0.7932	0.7745	0.7721	0.7724	0.7680
	SKNN	0.9418	0.9394	0.9308	0.9243	0.9228	0.9030	0.9264	0.9215	0.9132	0.9134	0.9090	0.9057
	IKNN	0.9454	0.9315	0.9285	0.9249	0.9272	0.9102	0.9146	0.9076	0.9019	0.8965	0.8906	0.8843
	HCS	0.9056	0.9021	0.8843	0.8809	0.8734	0.8719	0.9084	0.9021	0.9000	0.8856	0.8882	0.8805
	bi-BPCA	0.9574	0.9553	0.9520	0.9495	0.9411	0.9245	0.9341	0.9350	0.9248	0.9200	0.9095	0.9049
	MVETCA	**0.9875**	**0.9821**	**0.9798**	**0.9726**	**0.9621**	**0.9609**	**0.9537**	**0.9429**	**0.9448**	**0.9320**	**0.9303**	**0.9285**

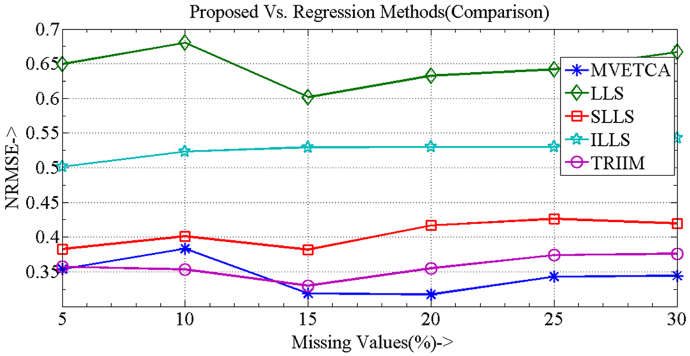


Fig. 24 NRMSE values for Leukemia1 dataset

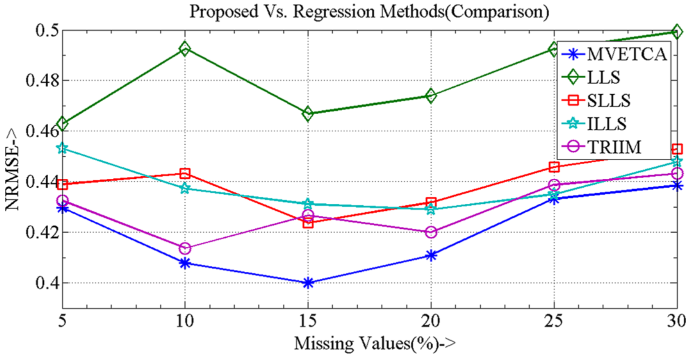


Fig. 25 NRMSE values for Lung cancer dataset

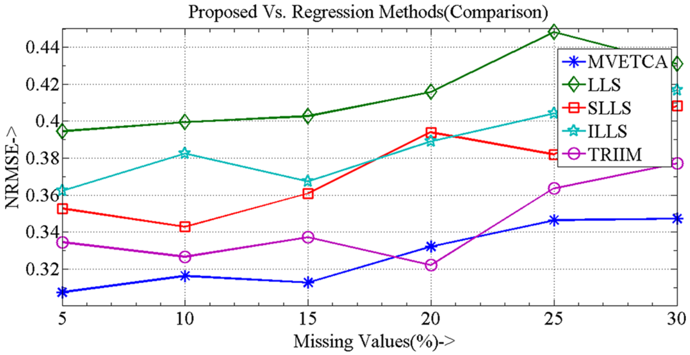


Fig. 26 NRMSE values for Prostate cancer dataset

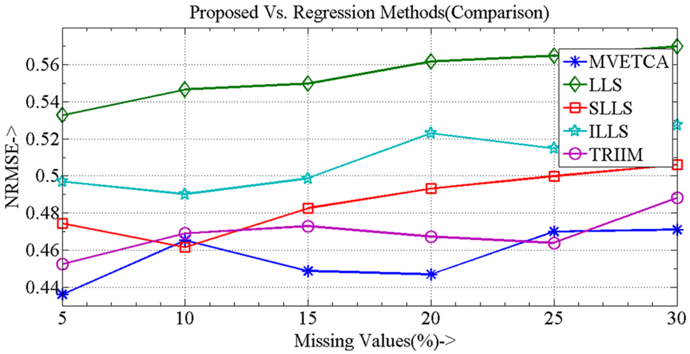


Fig. 27 NRMSE values for Breast cancer dataset

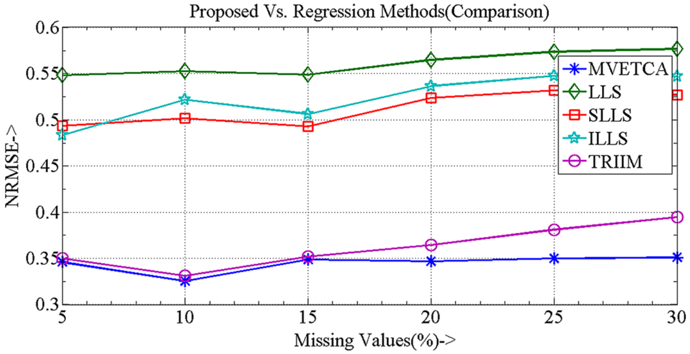


Fig. 28 NRMSE values for DLBCL dataset

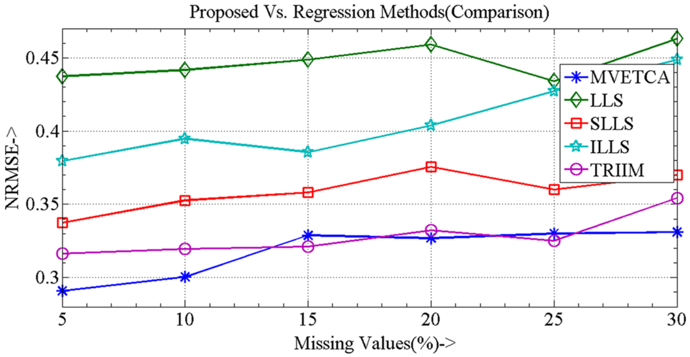


Fig. 29 NRMSE values for Colon cancer dataset

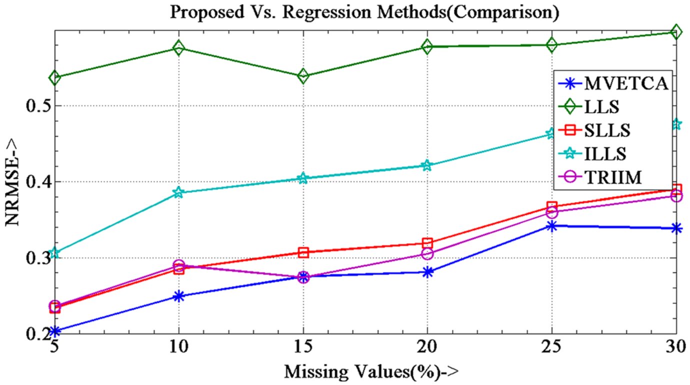


Fig. 30 NRMSE values for Leukemia2 dataset

In this circumstance, MVETCA method totally depends on gene expression values and independent on number of genes and all genes are participated to estimating the missing values and use bitwise similarity matching instead of the Euclidian distance metric. To measure the correlation with respect to expression values between the normal and cancerous samples, the dataset is split into small subsets, which help to estimate the missing values effectively. In this paper, the gene dataset is initially divided into a group of correlated genes and then splitting and merging-based clustering algorithm gives the stable and optimal clusters of gene and finally the missing value of a gene is estimated by comparing it with the centroids of the final clusters of genes. The performance of proposed method is analyzed using publicly available microarray datasets, and the accuracy of the method is compared with some state of the art methods measuring *NRMSE, CPP* and *BLCI* values, which shows the goodness of proposed method. The proposed method is applicable for any dataset with two or more class labels for imputing the missing values. So, though the method is not suitable for a time series dataset of single class, it is equally applicable for multi-class time series microarray datasets.

Table 11 *CPP* and *BLCI* values for experimental datasets

Dataset	Methods	CPP						BLCI					
		Missing values (%)						Missing values (%)					
		5	10	15	20	25	30	5	10	15	20	25	30
Leukemia1	LLS	0.8683	0.8502	0.8683	0.8206	0.8153	0.8074	0.7218	0.7172	0.7102	0.7058	0.6990	0.7013
	SLLS	0.9617	0.9602	0.9539	0.9449	0.9461	0.9394	0.8298	**0.8173**	0.8108	0.8056	0.8022	0.8017
	ILLS	0.9644	0.9446	0.9327	0.9335	0.9163	0.9115	0.8153	0.8129	0.8109	0.8065	0.7951	0.7745
	TRIIM	0.9742	0.9694	**0.9686**	0.9446	0.9369	0.9285	0.8328	0.8154	0.8157	0.8145	**0.8109**	0.8004
	MVETCA	**0.9746**	**0.9707**	0.9679	**0.9603**	**0.9588**	**0.9442**	**0.8462**	0.8057	**0.8247**	**0.8340**	0.8102	**0.8011**
Lung	LLS	0.7173	0.7085	0.6978	0.6942	0.6889	0.6871	0.7120	0.7219	0.7052	0.7007	0.6926	0.6941
	SLLS	0.8999	0.8865	0.8727	0.8700	0.8664	0.8569	0.8572	0.8475	0.8310	0.8281	0.8042	0.8011
	ILLS	0.8891	0.8777	0.8654	0.8695	0.8477	0.8476	0.8462	0.8410	0.8393	0.8128	0.8081	0.7967
	TRIIM	0.9461	**0.9516**	0.9367	0.9264	**0.9214**	0.8992	0.8573	**0.8490**	0.8428	**0.8414**	0.8293	0.8251
	MVETCA	**0.9540**	0.9505	**0.9451**	**0.9294**	0.9163	**0.9053**	**0.8630**	0.8476	**0.8590**	0.8411	**0.8362**	**0.8246**
Prostate	LLS	0.8638	0.8318	0.8371	0.8296	0.8139	0.8160	0.7452	0.7320	0.7338	0.7284	0.7169	0.7142
	SLLS	0.9780	**0.9687**	**0.9528**	0.9254	0.9217	0.9126	0.7933	0.7912	0.7827	0.7802	0.7742	0.7700
	ILLS	0.9452	0.9311	0.9203	0.9213	0.9029	0.9018	0.7727	0.7704	0.7684	0.7720	0.7616	0.7563
	TRIIM	0.9789	0.9726	0.9436	0.9254	0.9300	**0.9222**	**0.8109**	0.8014	0.7918	0.7872	0.7824	0.7741
	MVETCA	**0.9828**	0.9649	0.9420	**0.9387**	**0.9311**	0.9174	0.8092	**0.8043**	**0.7958**	**0.7960**	**0.8024**	**0.7868**
Breast	LLS	0.7482	0.7320	0.7300	0.7281	0.7162	0.7005	0.7652	0.7320	0.7427	0.7218	0.7090	0.7027
	SLLS	0.9451	0.9375	0.9309	0.9133	0.9028	0.8759	0.8837	0.8730	0.8801	0.8732	0.8522	0.8504
	ILLS	0.9400	0.9361	0.9388	0.9253	0.9004	0.9017	0.8720	0.8519	0.8538	0.8390	0.8318	0.8257
	TRIIM	**0.9662**	0.9610	0.9484	0.9410	0.9389	0.9162	0.8948	0.8783	0.8777	0.8710	0.8544	0.8502
	MVETCA	0.9633	**0.9630**	**0.9609**	**0.9563**	**0.9503**	**0.9471**	**0.9039**	**0.9027**	**0.8827**	**0.8740**	**0.8640**	**0.8701**
DLBCL	LLS	0.8276	0.8190	0.8214	0.8052	0.7904	0.7960	0.7053	0.7080	0.7028	0.6849	0.6836	0.6731
	SLLS	0.9452	0.9412	0.9344	0.9390	0.9157	0.9100	**0.8271**	0.8110	0.8031	0.7946	0.7937	0.7971
	ILLS	0.9527	0.9352	0.9270	0.9263	0.9017	0.9008	0.8030	0.8007	0.7928	0.7891	0.8009	0.7811

Table 11 continued

Dataset	Methods	CPP						BLCI					
		Missing values (%)						Missing values (%)					
		5	10	15	20	25	30	5	10	15	20	25	30
	TRIIM	0.9742	0.9730	**0.9793**	0.9620	0.9382	0.9450	0.8247	**0.8139**	0.8076	0.8012	0.7959	0.7933
	MVETCA	**0.9846**	**0.9821**	0.9754	**0.9658**	**0.9611**	**0.9539**	0.8255	0.8046	**0.8143**	**0.8128**	**0.8080**	**0.8001**
Colon	LLS	0.8745	0.8835	0.8310	0.8427	0.8166	0.8120	0.8464	0.8273	0.8299	0.8252	0.8145	0.8132
	SLLS	0.9802	0.9749	0.9638	0.9633	0.9570	**0.9437**	0.9073	0.9051	0.8937	0.8720	0.8547	0.8390
	ILLS	0.9453	0.9212	0.9252	0.9130	0.9017	0.9090	0.8837	0.8816	0.8630	0.8699	0.8542	0.8507
	TRIIM	**0.9831**	0.9810	**0.9771**	0.9689	0.9452	0.9411	0.9254	0.9218	0.9107	0.9068	0.9154	0.9013
	MVETCA	0.9803	**0.9853**	0.9766	**0.9729**	**0.9532**	0.9424	**0.9464**	**0.9475**	**0.9384**	**0.9322**	**0.9163**	**0.9110**
Leukemia2	LLS	0.7620	0.7602	0.7564	0.7510	0.7470	0.7426	0.7596	0.7608	0.7529	0.7505	0.7469	0.7420
	SLLS	0.9507	0.9515	0.9454	0.9408	0.9375	0.9266	0.9320	0.9276	0.9204	0.9165	0.9134	0.9122
	ILLS	0.9399	0.9332	0.9300	0.9295	0.9254	0.9211	0.9046	0.9019	0.8967	0.8950	0.8842	0.8810
	TRIIM	0.9564	0.9530	0.9512	0.9497	0.9481	0.9429	0.9368	0.9198	0.9167	0.9207	0.9138	0.9120
	MVETCA	**0.9875**	**0.9821**	**0.9798**	**0.9726**	**0.9621**	**0.9609**	**0.9537**	**0.9429**	**0.9448**	**0.9320**	**0.9303**	**0.9285**

Acknowledgements The authors would like to thank anonymous reviewers for their valuable comments.

Compliance with ethical standards

Conflict of interest The authors declare that there are no conflicts of interest in this paper.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Alizadeh AA (2000) Distinct types of diffuse large B-cell Lymphoma identified by gene expression profiling. Nature 403:503–511
2. Bezdek JC, Pal NR (1998) Some new indexes of cluster validity. IEEE Trans Syst Man Cybern 28(3):301–315
3. Bra's LP, Menezes JC (2007) Improving cluster-based missing value estimation of DNA microarray data. Biomol Eng Elsevier 24:273–282
4. Brevern AG, Hazout S, Malpertuy A (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinform. doi:10.1186/1471-2105-5-114
5. Butte AJ, Ye J (2001) Determining significant fold differences in gene expression analysis. Pac Symp Biocomput 6:6–17
6. Cai Z, Heydari M, Lin G (2006) Iterated local least squares microarray missing value imputation. Bioinform Comput Biol 4:935–957
7. Causton HC, Quackenbush J, Brazma A (2004) Microarray gene expression data analysis: a Beginner's guide, vol 21. Blackwell, Oxford, pp 973–974
8. Cheng KO, Law NF, Siu WC (2012) Iterative bicluster-based least square framework for estimation of missing values in micro array gene expression data. Pattern Recognit 45(4):1281–1289
9. Das AK, Sil J (2010) Cluster validation method for stable cluster formation. Can J Artif Intell Mach Learn Pattern Recognit 1(3):26–41
10. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227
11. de Brevern AG, Hazout S, Malpertuy A (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinform. doi:10.1186/1471-2105-5-114
12. DeRisi J (1996) Use of a cDNA microarray to analyze gene expression patterns in human cancer. Nat Genet 14(4):457–460
13. Fu L, Medico E (2007) FLAME: a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinform. doi:10.1186/1471-2105-8-3
14. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17(2–3):107–145
15. Hand DJ, Heard NA (2005) Finding groups in gene expression data. J Biomed Biotechnol 2:215–225
16. He C, Li HH, Zhao C et al (2015) Triple imputation for microarray missing value estimation. IEEE international conference on bioinformatics and biomedicine (BIBM), pp 208–213
17. Huynen M, Snel B, Lathe W et al (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10:1204–1210
18. Ji R, Liu D, Zhou Z (2011) A bicluster-based missing value imputation method for gene expression data. J Comput Inf Syst 7(13):4810–4818
19. Kaur A, Singh SS, Kaur SS (2010) Fuzzy clustering based missing value estimation of gene expression data. Computer engineering technology RIMT, pp 122–126
20. Kent Ridge Bio-medical Dataset. http://datam.i2r.a-star.edu.sg/datasets/krbd
21. Kim KY, Kim BJ, Yi GS (2004) Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinform. doi:10.1186/1471-2105-5-160
22. Kim H, Golub GH, Park H (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21(2):187–198
23. Koopmans R, Schaeffer M (2015) Relational diversity and neighborhood cohesion unpacking variety balance and in-group size. Soc Sci Res Elsevier 53:162–176
24. Luengo J, García S, Herrera F (2011) On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst 32:77–108

25. Luo J, Yang T, Wang Y (2005) Missing value estimation for microarray data based on fuzzy C-means clustering. In: Proceedings of the 8th international conference on high-performance computing in Asia-Pacific region (HPCASIA'05), pp 611–616
26. Maulik U, Bandyopadhyay S (2002) Performance evaluation of some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell 24(12):1650–1654
27. Meng F, Cai C, Yan H (2014) A bicluster-based Bayesian principal component analysis method for microarray missing value estimation. IEEE J Biomed Health Inform 18(3):863–871
28. Oba S, Sato MA, Takemasa I et al (2003) A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19(16):2088–2096
29. Oh S, Kang DD, Brock GN et al (2011) Biological impact of missing-value imputation on downstream analyses of gene expression profiles. Bioinformatics 27(1):78–86
30. Pan L, Li J (2010) K-nearest neighbor based missing data estimation algorithm in wireless sensor networks. Wirel Sens Netw Sci Res 2:115–122
31. Paul A, Sil J (2011) Estimating missing value in microarray gene expression data using fuzzy similarity measure. IEEE international conference on fuzzy systems- Taiwan, pp 27–30
32. Paul A, Sil J (2011) Missing value estimation in microarray data using Co regulation and similarity of genes. World congress on information and communication technologies, pp 705–710
33. P'erez MJ, Romero-Campero FJ (2006) A new computational modeling tool for systems biology. Trans Comput Syst Biol 6:176–197
34. Pourhashem MM, Kelarestaghi M, Pedram MM (2010) Missing value estimation in microarray data using fuzzy clustering and semantic similarity. Global J Comput Sci Technol 10(12):18–22
35. Qu Y, Xu S (2004) Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. Bioinformatics 20:1905–1913
36. Rahman MG, Islam MZ, Bossomaier T, Gao J (2012) Cairad: a co-appearance based analysis for incorrect records and attribute-values detection. IEEE international joint conference on neural networks (IJCNN), pp 1–10. doi:10.1109/IJCNN.2012.6252669
37. Rahman MG, Islam MZ (2016) Missing value imputation using a fuzzy clustering-based EM approach. Knowl Inf Syst 46:389–422
38. Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7(2):147–177
39. Shi F, Zhang D, Chen J et al (2013) Missing value estimation for microarray data by Bayesian principal component analysis and iterative local least squares. Math Probl Eng. doi:10.1155/2013/162938
40. Suresh RM, Dinakaran K, Valarmathie P (2009) Model based modified k-means clustering for microarray data. ICIME 53:271–273
41. Troyanskaya O, Cantor M, Sherlock G et al (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525
42. Tusher VG (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98:5116–5121
43. Velarde CC, Escudero R, Zaliz RR (2008) Boolean networks: a study on microarray data discretization. ESTYLF08, Cuencas Mineras (Mieres-Langreo), pp 17–19
44. Wang H, Wang S (2010) Mining incomplete survey data through classification. Knowl Inf Syst 24(2):221–233
45. Zahid N, Limouri M, Essaid A (1999) A new cluster-validity for fuzzy clustering. Pattern Recogn 32:1089–1097
46. Zhang S, Zhang J, Zhu X, Qin Y, Zhang C (2008) Missing value imputation based on data clustering. Trans Comput Sci 1:128–138
47. Zhang X, Song X, Wang H et al (2008) Sequential local least squares imputation estimating missing value of microarray data. Comput Biol Med 38:1112–1120
48. Zhang S (2011) Shell-neighbor method and its application in missing data imputation. Appl Intell 35(1):123–133
49. Zhang S, Jin Z, Zhu X (2011) Missing data imputation by utilizing information within incomplete instances. Syst Softw 84(3):452–459
50. Zhao O, Fränti P (2014) WB-index: a sum-of-squares based index for cluster validity. Data Knowl Eng Elsevier 92:77–89

Soumen Kumar Pati is working toward his PhD in Computer Science & Technology from Department of Computer Science & Technology, IIEST, Shibpur. He holds Assistant Professor in Department of Information Technology, St. Thomas' College of Engineering & Technology, Kolkata, India. His research interest includes Bioinformatics, Data Mining, Pattern Recognition, Rough Set theory, etc.

Asit Kumar Das received his PhD in Computer Science and Technology from Bengal Engineering and Science University, Shibpur, in 2011. He currently holds Associate Professor Position in the Department of CST, IIEST, Shibpur, India. He has authored numerous scientific research papers in refereed journals and conferences. His research interest includes Data Mining, Pattern Recognition, Social Network, Bioinformatics, etc.