Feature Selection with Missing Values in DNA Microarray Gene Expression Data

Nilooofar Rabiei

Hamadan University of Medical Sciences

Ali Reza Soltanian (✉ soltanian@umsha.ac.ir)

Hamadan University of Medical Sciences

Maryam Farhadian

Hamadan University of Medical Sciences

Fatemeh Bahreini

Hamadan University of Medical Sciences

Research Article

Keywords: Entropy, Feature Selection, Gene, Mutual Information, Redundancy

Posted Date: February 23rd, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1354806/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Title: Feature Selection with Missing Values in DNA Microarray Gene Expression Data

Niloofar Rabiei¹, Ali Reza Soltanian^{2*}, Maryam Farhadian³, Fatemeh Bahreini⁴

Affiliations:

1- Niloofar Rabiei, Ph.D. Student, Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. n.rabee@edu.umsha.ac.ir.

ORCID: [0000-0002-4447-2080](https://orcid.org/0000-0002-4447-2080)

2- Ali Reza Soltanian, Ph.D., Modeling of Noncommunicable Diseases Research Center, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.

soltanian@umsha.ac.ir.

ORCID: [0000-0002-7483-3502](https://orcid.org/0000-0002-7483-3502)

3- Maryam Farhadian, Ph.D., Research Center for Health Sciences, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. maryam_farhadian80@yahoo.com.

ORCID: [0000-0002-6054-9850](https://orcid.org/0000-0002-6054-9850)

4- Fatemeh Bahreini, Ph.D., Department of Molecular Medicine and Genetics, School of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran. F.bahreini@umsha.ac.ir.

ORCID: [0000-0003-4444-1206](https://orcid.org/0000-0003-4444-1206)

*Corresponding author: Ali Reza Soltanian, Professor, Street of Shahid Fahmideh, Hamadan University of Medical Sciences, Hamadan, Iran. Tel/Fax: +98(81)38380025/+98(81)38380026, P.O. BOX: 6517838736, E-mail: soltanian@umsha.ac.ir

Abstract

Background

Imputation is one of the strategies for dealing with Missing values (MVs) in microarray data. Employing the best subset of genes for imputation is very important. In this study, we used mutual information gene selection before imputation to select the best subset of genes for imputation MVs and then classified imputed data. Two datasets were used, and we generate MVs with missing rates from 1 to 7 percent. K nearest neighbor, row mean imputation, and the method contains Feature Selection with Missing data by Mutual Information (FSM-MI) were employed. We used Root Mean Square Error (RMSE) for evaluating the performance of the methods. We classified complete and imputed data by random forest classifier and compare them by accuracy.

Results

FSM-MI imputation method with 0.0364 and 0.0083 mean RMSE value in GSE510 and GSE1063 datasets had the best performance, respectively. The classification accuracy of complete and imputed data in GSE510 and GSE1063 datasets were 100 and 80, and 100, 100, 83.3, and 66.7, and 80, 80, 70, and 70 percent in missing rates 1, 3, 5, and 7, respectively.

Conclusion

Feature selection before imputation MVs is as important as the selection of the best imputation method in improving the result of subsequent analyzes.

Keywords: Entropy, Feature Selection, Gene, Mutual Information, Redundancy.

Background

Recently, High dimensional data analysis i.e. microarrays are important issues in bioinformatics, system biological sciences, and machine learning techniques. One of the major obstacles to analyzing microarray data is the presence of MVs. Some of the reasons for MVs in microarray data are insufficient resolution, image corruption, poor hybridization, or contaminants due to dust or scratches on the slide(1). Ordinary microarray data have 1-10% MVs. Since most of the microarray data analysis methods such as biomarker identification, classification, and gene clustering require complete data without any missing, so solving the problem of MVs is one of the most important steps in microarray data analysis(2, 3). To limit the effect of MVs in analyzing microarray data different approaches were employed: (i) repeating experiments which is a time-consuming and costly process (4), (ii) deleting missing observations, but leads to information loss(5), or (iii) imputing missing data using the mode, mean, or zero which is one of the most common approaches, but the results are not optimal(6, 7). Different imputation algorithms such as K-Nearest Neighbor (KNN), singular value decomposition (SVD), EM_array, LSimpute_gene, and LSimpute_array, and bayesian principal component analysis (BPCA) were proposed (8-10). Some of these approaches use different similarity measures. Based on these approaches the similarity between all genes and the gene we intend to impute its MVs (target gene) calculated, and the most similar genes to the target gene selected for imputation (9, 11-13).

Microarrays data has more than thousands of genes with small samples and poses a significant challenge to researchers. Many of these genes may be completely irrelevant to the task, or redundant in the context of others. Learning in this situation has several problems such as computational burden(14). Therefore, to overcome this problem, a suitable subset of genes must be selected, so that the correct results can be extracted(15). In general, a good set of features is a

set where the features are individually relevant (the most correlation with an output), but should not be redundant (the least correlation between themselves) to each other. For this purpose, many mathematical and statistical models and methods have been introduced, which each of them has a different performance. One of these methods is mutual information, introduced by Shannon, which is the amount of information shared by two features (16). Several studies incorporate mutual information for considering relevancy and redundancy between features (17-22).

In this paper, to select the best subset of genes that are used in the imputation process of MVs in the target gene, the two relevance-redundancy methods are employed to select the subset of genes with the most relevance individually and the least redundancy to each other. Also to emphasize the effect of different imputation methods on the accuracy of classification, after imputation, a random forest classifier was employed to classify the complete data and imputed data based on the imputation methods used in this study, and classification accuracies were compared.

Methods

Data sets

In this study, we used two datasets from Gene Expression Omnibus (GEO) database with accession numbers GSE510 and GSE1063. Genes with at least one MV were removed from the gene expression datasets, and reference gene expression data was created (complete dataset). The first reference gene expression matrix contains 146 genes measured in 3 *Mus Musculus* samples in each control and case group. Second reference gene expression data contains 216 genes measured in 5 *Rattus norvegicus* samples in two groups (vitamin E supplemented diet, vitamin E deficient diet).

Missing Simulation

MVs were generated with missing rates of 1, 3, 5, and 7 percentage from reference gene expression data with a completely at random (CAR) mechanism, and simulated data were created (incomplete dataset).

Let $X^{N \times p}$ be expression matrix contain N gene in p experiment. The elements of matrix define as x_{ij} $i = 1, \dots, N, j = 1, \dots, p$. We define an indicator matrix $R \in \{0,1\}^{N \times p}$, and call it the missingness pattern of the expression matrix $X^{N \times p}$, and it's elements define as r_{ij} $i = 1, \dots, N, j = 1, \dots, p$. $r_{ij} = 1$ if x_{ij} is observed and $r_{ij} = 0$ otherwise. Based on this pattern, the expression matrix $X^{N \times p}$ can partition into observe and missing matrix as $X^{N \times p} = \{X^o, X^m\}$. Both X and R define as random variables with probability distributions P_X and P_R , respectively. The MVs are said to be CAR if the probability that an observation is missing is independent of the observed and MVs, and defined as $p(R|X^o, X^m) = p(R)$ (23).

Matrices created with different missing rates are called simulated matrices. This process was repeated three times separately and finally three independent simulated matrices were generated for each missing rate.

KNN Imputation Method

In the KNN method, the MVs are imputed by the corresponding values of the k nearest genes. In this method, at first, the optimal number of nearest neighbors (k) is set based on experimental tests. Then similarity between the target gene (the gene with MVs) and the genes without any missing calculated. There are different metrics for measuring similarity (Pearson correlation Coefficient, Manhattan distance, Euclidean distance, and variance minimization). In the third step, the top k similar genes selected and the MV imputed by an average of values in these k genes in the same position with target gene(9).

115 **Mean Imputation Method**

116 Substituting the mean or imputation the arithmetic mean, also known as the unconditional mean
117 imputation, is a strategy in which the mean for a gene is calculated based on all the observations
118 that are not missing. This means is then used instead of any missing value on that gene.

119 **Feature Selection with Missing data by Mutual Information (FSM-MI)**

120 As mentioned earlier, a good subset of genes should be individually relevant and also not be
121 redundant concerning each other. To achieve this purpose, many entropy-based methods have been
122 proposed. The Mutual Information-Based Feature Selection (MIFS) scheme was the first of many
123 criteria that attempted to capture the relevance-redundancy trade off in feature sets(17) and defined
124 as follows,

$$125 \quad J_{Mifs} = I(X_N; Y) - \beta \sum_{k=1}^{N-1} I(X_N; X_k), \quad (1)$$

126 Where X_N is N^{th} gene in $X^{N \times p}$ expression matrix, Y is target gene, $I(X_N; Y)$ is the term for gene
127 relevance and $\sum_{k=1}^{N-1} I(X_N; X_k)$ is for redundancy with penalty parameter equal to β . the penalty in
128 MIFS approach must be set experimentally and based on inter-feature dependencies.

129 The mutual information or interdependency between genes X and Y is defined as:

$$130 \quad I(X; Y) = H(X) - H(X|Y), \quad (2)$$

131 Where $H(X|Y)$ is conditional entropy and is defined as:

$$132 \quad H(X|Y) = H(X, Y) - H(Y), \quad (3)$$

133 $H(X, Y)$ is the joint entropy of X and Y and is defined as:

$$134 \quad H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y), \quad (4)$$

135 Where χ and $\mathcal{Y}$ are a subset of possible values of gene X and Y, respectively. $p(x, y)$ is joint
136 probabilities.

137 $H(X)$ is entropy and defined as:

138
$$H(X) = -\sum_{x \in \chi} p(x) \log p(x), \quad (5)$$

139 Where χ is a subset of possible values of gene X with corresponding probabilities(24, 25).

140 The pseudo code for FSM-MI method is given below:

- 141 (I) Split dataset into complete and incomplete gene set which complete gene set contain
142 genes with no missing and incomplete gene set contain genes with at least one missing.
- 143 (II) Missing rates for the genes in incomplete set computed as follows:
144 $r_i = \frac{m_i}{p}$, which m_i is the number of MVs in gene i , and p is the number of experiments.
- 145 (III) All MVs imputed by row mean method.
- 146 (IV) The gene with minimum missing rate choose as a target gene, and from complete gene
147 set the subset of genes with maximum relevance and minimum redundancy choose.
- 148 (V) Impute MVs in target gene by KNN using the genes selected in pervious step.
- 149 (VI) The imputed gene goes to the complete gene set.
- 150 (VII) Repeat steps II-VI.

151 Figure 1**Error! Reference source not found.** illustrates the proposed method.

152 [Please insert Figure 1: The proposed mutual information imputation method]

153 As it said the MIFS used the β as a penalty and it must be set experimentally. Another criterion for
154 capturing relevancy and redundancy is Maximum-Relevance Minimum-Redundancy (MRMR) and
155 is equivalent to MIFS with $\beta = \frac{1}{n-1}$ (18).

In step IV we used MIFS and MRMR criteria for selecting the subset of genes with maximum relevance and minimum redundancy. We also use KNN and Row mean imputation methods and compare them by RMSE, which is used to measure the numerical similarity between the values imputed in the simulated matrix and the original values in the reference gene expression matrix. The RMSE is defined as the following:

$$(6)RMSE = \sqrt{\frac{\sum_{i=1}^M (G_{i,Orig} - G_{i,Imp})^2}{M}},$$

The term $G_{i,Orig}$ is the original expression value, and $G_{i,Imp}$ is the expression value of an imputed gene for a given imputation algorithm, and M is the number of MVs. The smaller the RMSE is, the better the imputation method is.

Classification

To compare the effect of the missing imputation methods on classification, after imputation of the MVs, we select the best and the worst imputation method based on RMSE, and a Random Forest (RF) classifier was employed for classification. We also classified a reference gene expression matrix and the results of the classification were compared based on 3-fold classification accuracy. Before starting the classification process, the dataset is split into train and test sets. In the first step of the classification process, set the number of tree (*ntree*) bootstrap sampling and the number of variables randomly sampled as candidates at each split (*mtry*) from the train set and grow a tree for each bootstrap dataset. Each tree grows until there are no fewer than *nodesize* in each terminal node. In the second step, we used majority voting for aggregating data from *ntree* trees. In the last step, we calculated the accuracy of classification by a test set(26).

Results

Figure 2 showed the result of missing imputation for GSE510 dataset. RMSE value for all methods increased as missing rate increased. Among imputation methods used, FSM-MI method had the minimum RMSE value. KNN imputation method with different k values employed and when missing rate was 1%, 3%, and 7%, k = 10 had minimum RMSE, when missing rate was 5%, k = 20 had minimum RMSE.

[Please insert Figure 2: RMSE values for GSE510 dataset (a) 1%; (b) 3 %; (c) 5%; (d) 7% missing]

Figure 3 showed the result of missing imputation for GSE1063 dataset. RMSE value for all methods increased as missing rate increased. Among imputation methods used, FSM-MI method had the minimum RMSE value. KNN imputation method with different k values employed and when missing rate was 1%, 3%, 5%, and 7% k = 10 had minimum RMSE.

[Please insert Figure 3: RMSE values for GSE1063 dataset (a) 1%; (b) 3 %; (c) 5%; (d) 7% missing]

Table 1 showed mean RMSE and SE for two datasets used. In GSE510 dataset, FSM-MI method and row mean imputation method had the minimum and maximum mean RMSE value, respectively. While in GSE1063 dataset, FSM-MI method and KNN imputation method had the minimum and maximum mean RMSE value, respectively. For KNN imputation method, k = 10 had the minimum mean RMSE value.

197

Table 1. Mean RMSE values for GSE510 and GSE1063 datasets

Imputation Method	Dataset	
	GSE510	GSE1063
	Mean(SE)	Mean(SE)
KNN($k^*=10$)	0.0466(0.014)	0.0169(0.003)
KNN($k^*=20$)	0.0469(0.013)	0.0220(0.004)
KNN($k^*=30$)	0.0485(0.014)	0.0267(0.004)
KNN($k^*=40$)	0.0508(0.014)	0.0317(0.005)
KNN($k^*=50$)	0.0535(0.015)	0.0356(0.006)
FSM-MI(MRMR [€])	0.0367(0.011)	0.0083(0.001)
FSM-MI (MIFS [†])	0.0364(0.011)	0.0083(0.001)
Row Mean	0.0495(0.013)	0.0088(0.002)

198

SE: Standard Error; k: The number of nearest neighbor; €: Maximum-

199

Relevance Minimum-Redundancy criteria; †: Mutual Information-Based

200

Feature Selection criteria.

201

Based on the results reported in Figure 1 and Figure 2, we chose the best and the worst imputation

202

method in each missing rate and classified imputed data with RF classifier. The accuracy of

203

classification reported in Table 2. The classification accuracy of complete data in GSE510 and

204

GSE1063 was 100 and 80 percent, respectively. Classification accuracy of imputed data decreased

205

as missing rate increased. In both dataset, the imputed data with FSM-MI method using MRMR

206

and MIFS criteria had better accuracy comparing to KNN($k=10$) and Row mean imputation

207

method.

208

Table 2: Classification Accuracy for GSE510 and GSE1063 dataset after imputation

Missing Rate (%)	GSE510		GSE1063	
	Imputation method	Accuracy (%)	Imputation method	Accuracy (%)
0	None	100.0	None	80.0
1	FSM-MI (MIFS [†])	100.0	FSM-MI (MRMR [€])	80.0
	Row Mean	83.3	KNN($k=10$)	60.0
3	FSM-MI (MIFS [†])	100.0	FSM-MI (MIFS [†])	80.0
	KNN($k=10$)	66.7	KNN($k=10$)	60.0
5	FSM-MI (MRMR [€])	83.3	FSM-MI (MIFS [†])	70.0
	Row Mean	66.7	KNN($k=10$)	50.0
7	FSM-MI (MRMR [€])	66.7	FSM-MI (MIFS [†])	70.0
	Row Mean	50.0	KNN($k=10$)	50.0

209

k: The number of nearest neighbor; €: Maximum-Relevance Minimum-Redundancy criteria; †: Mutual Information-Based Feature Selection criteria

210

Discussion

Biological experiments such as Microarray data contain MVs, and they can influence the subsequent analysis. One of the approaches to dealing with MVs is missing imputation. There is no best imputation method and the best method should be selected according to the type of data. In this study, we used two datasets to investigate the performance of the KNN and row mean imputation method. We also proposed an approach that selects the subset of genes with maximum relevancy and minimum redundancy with the target gene before imputation and then imputed MVs based on this subset.

KNN imputation method employed with k value 10 to 50 by steps 5. As shown in the result section, increasing k values do not reduce the RMSE values and the minimum RMSE value was obtained in $k = 10$. In the GSE1063 dataset, for missing rates used in this study, the KNN imputation method had the maximum RMSE value. In the GSE510 dataset, for all missing rates used in this study, the row mean imputation method had the maximum RMSE value. For both datasets the FSM-MI method had the minimum RMSE value.

As it shown, in both dataset the FSM-MI method had a stable behavior and the best performance. In contrast the behavior of KNN and row mean imputation methods is unstable, as in the GSE510 dataset the KNN imputation with $k = 10$, and in the GSE1063 dataset row mean imputation method had lower RMSE value. The good performance of the row mean imputation method in the GSE1063 dataset could be due to the low dispersion of gene expression in the experiments. One of the disadvantages of the row mean imputation method is that imputing the row mean instead of the MVs reduces the variance of the genes and leads to unreliable estimates of covariances and correlations in subsequent analyses.

It should be noticed, there is no the best imputation method and it should be determined by the nature of the, and the results of this study are a testament to this sentence. As it shown, Row mean had relatively good performance in one dataset, but its performance is not good in the other dataset. KNN imputation method behaved similarly to the row mean imputation method.

After imputation of MVs with KNN, row mean and FSM-MI method, the best and the worst imputation method employed based on mean RMSE value and classification is done using RF classifier. RF is a popular ensemble method, and can applies to high dimensional data.

The classification accuracy of complete data in GSE510 and GSE1063 dataset was 100 and 80 percent. The classification accuracy of each incomplete dataset with missing rate 1, 3, 5 and 7 percent calculated after imputation with the best and the worst imputation method. With increasing the missing rate the classification accuracy decreased. The classification accuracy of the data that imputed with the best imputation method is higher than the data that imputed with the worst imputation method. As it shown, in both dataset the classification accuracy of imputed data with the best imputation method when missing rate was 1 and 3 percent is exactly similar to accuracy of complete data.

Conclusion

Employing the best subset of genes for imputation MVs is very important issue, and can improve performance of missing imputation method. Also applying the appropriate imputation method reduce the bias of subsequent analyzes and achieve the best results as long as possible.

Declaration

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Two datasets chose from Gene Expression Omnibus (GEO, Pubmed) database with accession numbers GSE510 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE510>) and GSE1063 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1063>).

Competing interest

The authors declare that they have no competing and conflict of interests.

Funding

Not applicable

Author contributions

All authors have contributed to the preparation and expanding the manuscript. Each author has read and approved the manuscript. ARS and NR contributed to the literature review and analysis of the study and drafting the manuscript. FB contributed to clinical and genetics aspects of the manuscript. ARS, NR, MF and FB reviewed the manuscript.

Acknowledgement

This study was a part of Ph.D. thesis at Hamadan University of Medical Sciences with the approval ethical code: IR.UMSHA.REC.1400.212. We would like to thank the Vice-Chancellor of Research and Technology, Hamadan University of Medical Sciences for the approval and support of the study (No 140006024508).

List of abbreviations

276 MVs: Missing values

277 FSM-MI: Feature Selection with Missing data by Mutual Information

278 RMSE: Root Mean Square Error

279 KNN: K-Nearest Neighbor

280 SVD: Singular Value Decomposition

281 BPCA: Bayesian Principal Component Analysis

282 GEO: Gene Expression Omnibus

283 CAR: Completely At Random

284 MIFS: Mutual Information-Based Feature Selection

285 MRMR: Maximum-Relevance Minimum-Redundancy

286 RF: Random Forest

References

1. Xiang Q, Dai X, Deng Y, He C, Wang J, Feng J, et al. Missing value imputation for microarray gene expression data using histone acetylation information. *BMC bioinformatics*. 2008;9(1):1-17.
2. Saha S, Ghosh A, Seal DB, Dey KN. An improved fuzzy based missing value estimation in DNA microarray validated by gene ranking. *Advances in Fuzzy Systems*. 2016;2016.
3. De Brevern AG, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC bioinformatics*. 2004;5(1):1-12.
4. Butte AJ, Ye J, Häring H, Stumvoll M, White M, Kohane I. Determining significant fold differences in gene expression analysis. *Biocomputing 2001: World Scientific*; 2000. p. 6-17.
5. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503-11.
6. Salleh MNM, Samat NA, editors. FCMPPO: An imputation for missing data features in heart disease classification. *Materials Science & Engineering Conference Series*; 2017.
7. Little RJ, Rubin DB. *Statistical analysis with missing data*: John Wiley & Sons; 2019.
8. Bø TH, Dysvik B, Jonassen I. LSImpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic acids research*. 2004;32(3):e34-e.
9. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520-5.
10. Oba S, Sato M-a, Takemasa I, Monden M, Matsubara K-i, Ishii S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*. 2003;19(16):2088-96.
11. Saha S, Singh PK, Dey KN. Missing value estimation in DNA microarrays using linear regression and fuzzy approach. *Computation and Communication Technologies: De Gruyter*; 2016. p. 254-68.
12. Saha S, Dey KN, Dasgupta R, Ghose A, Mullick K. Missing value estimation in DNA microarrays using B-splines. *Journal of Medical and Bioengineering*. 2013;2(2).
13. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*. 2005;21(2):187-98.
14. Brown G, editor *A new perspective for information theoretic feature selection*. *Artificial intelligence and statistics*; 2009: PMLR.
15. Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of machine learning research*. 2003;3(Mar):1157-82.
16. Shannon CE. A mathematical theory of communication. *The Bell system technical journal*. 1948;27(3):379-423.
17. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*. 1994;5(4):537-50.

- 321 18. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency,
322 max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine
323 intelligence. 2005;27(8):1226-38.
- 324 19. Yang HH, Moody JE, editors. Data visualization and feature selection: new algorithms for
325 nongaussian data. NIPS; 1999: Citeseer.
- 326 20. Kwak N, Choi C-H. Input feature selection for classification problems. IEEE transactions on neural
327 networks. 2002;13(1):143-59.
- 328 21. Lin D, Tang X, editors. Conditional infomax learning: An integrated framework for feature extraction
329 and fusion. European conference on computer vision; 2006: Springer.
- 330 22. Fleuret F. Fast binary feature selection with conditional mutual information. Journal of Machine
331 learning research. 2004;5(9).
- 332 23. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.
- 333 24. Soltanian AR, Rabiei N, Bahreini F. Feature Selection in Microarray Data Using Entropy
334 Information. 2019.
- 335 25. Shannon CE. A mathematical theory of communication. Bell system technical journal.
336 1948;27(3):379-423.
- 337 26. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

Figures

Figure 1

illustrates the proposed method.

Figure 2

RMSE values for GSE510 dataset (a) 1%; (b) 3 %; (c) 5%; (d) 7% missing