J. Bio-Sci. 27: 31-41, 2019 ISSN 1023-8654
http://www.banglajol.info/index.php/JBS/index
DOI: https://doi.org/10.3329/jbs.v27i0.44669

IMPROVED *K*-NEAREST NEIGHBORS APPROACH FOR INCOMPLETE AND CONTAMINATED GENE EXPRESSION DATASETS

MI Asifuzzaman[1], H Akter[1], MM Rashid[1], MNH Mollah[2], SMS Islam[3] and M Shahjaman[1]*

[1]*Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh*
[2]*Bioinformatics Lab., Department of Statistics, University of Rajshahi, Bangladesh*
[3]*Instititute of Biological Sciences, University of Rajshahi, Bangladesh*

Abstract

With the rapid development of high-throughput DNA microarray technologies, researchers can measure expression profiles of thousands of genes simultaneously with low costs. These massive amounts of gene expression (GE) data often contain missing values or outliers due to various reasons of data generating process. Most of the statistical methods were developed based on complete dataset. As a result, for subsequent analysis using incomplete dataset, these methods strongly suffer and we cannot find our target. A numerous methods have been developed to impute missing values and they are available in the literature. Albeit, missing values imputation and outliers handling both are equally important for analyzing GE, most of the methods perform these tasks separately and produce misleading results. Therefore, in this paper, an attempt is made to develop a new hybrid approach which is robust against outliers and missing values, simultaneously. We demonstrate the performance of the proposed method in a comparison of popular missing value imputation method K-NN while performing feature selection using both simulated and real GE datasets. The Results obtain from simulated as well as real data studies show that the proposed method outperforms K-NN in presence of different percentages of missing values and outliers. On the other hand, in absence of outliers with missing values, the proposed method keeps equal performance with the other methods.

Key words: Gene expression data, IQR, Missing values, Outliers, Robustness

Introduction

Microarray technology allows researchers to measure the expression profiles for tens of thousands of features/genes in parallel by a single experiment and produce huge amounts of datasets (De Risi et al. 1997, Lockhart et al. 2000, Alam et al. 2017). It has been widely used in different biological disciplines such as cancer classification, drug discovery, stress response, regulation of cell cycle, clustering to discover the co-regulated gene groups, cancer prognosis, and identification of important features that are relevant to a certain disease etc. (Wang et al. 2006, Colombo et al. 2011). Microarray gene expressions (GE) datasets are high-dimensional with small sample sizes, usually $n<p$. Thus statistical methods that are used to analyze these datasets often suffer from computational complexities. One of the most important tasks of microarray GE data analysis is to select the most important features/genes from a large number of features (Li et al. 2004). Feature selection (FS) can enhance the performance of the methods for downstream analyses (Shahjaman et al. 2017a). Despite the wide spread use of microarray technology, GE data often contain missing values as well as outliers. Missing values and outliers are usually common in the high-dimensional OMICS datasets with dozens of variables/features and hundreds of samples/individuals. A variety of reasons

*Author for correspondence: shahjaman_brur@yahoo.com involve for missing values in GE data such as corruption of image, scratches on the slides, poor hybridization, inadequate resolution, fabrication errors and so on (Schuchhardt et al. 2000, Tuikkala et al. 2006). Microarray GE datasets typically contain 1-10% missing values that could affect up to 90% of genes (Chiu et al. 2013). On the other hand, outliers may also occur in GE datasets due to different steps of data generating process from hybridization to image analysis for various reasons (Shahjaman et al. 2017b). Outliers can deteriorate the performance of the feature selection methods. Therefore, outlier detection is very important for microarray GE data analysis (Nadon et al. 2002, Alam et al. 2016). Furthermore, for subsequent analysis, most of the methods were formulated based on complete datasets only. The first and simplest way to overcome these problems is to remove the genes corresponding to the missing values or outliers. However, in this procedure, we might be lost important information. The second method is the replace the missing value by zero (Alizadeh et al. 2000). In this case, researchers may puzzle between missing values and the values of real data that are close to zero. Therefore, the methods which replace the missing values by their estimated values have been developed. The first and most classical method to impute these values is the *K*-nearest neighbor (KNN) (Troyanskaya et al. 2001). Then the update version of KNN were developed which includes sequential *K*-nearest neighbor (SKNN) and iterative *K*-nearest neighbor (IKNN) (Kim et al. 2004) etc. These are known as local procedures. There are also many global missing value imputation procedures such as Bayesian principal component analysis (BPCA) (Fix et al. 1951), singular value decomposition (SVD) (Troyanskaya et al. 2001), partial least squares (PLS) and so on. The non-parametric random forest (RF) imputation (Stekhoven et al. 2012) and parametric expectation-maximization (EM) imputation (Dempster et al. 1977) also have been widely used in GE data analysis. Most of the traditional missing value imputation approaches cannot deal with outliers. Hence, they produce misleading results. Therefore, in this paper, an attempt is made to improve the popular K-NN approach by incorporating an IQR rule to detect and modify the outliers, which can deal with both missing values and outliers, simultaneously while performing feature selection.

Materials and Methods

Improved *K*-Nearest Neighbors (*K*-NN) Approach (proposed)

There are mainly two types of statistical approaches for data analysis when the data are contaminated by outliers: one is the application of robust method using original datasets and the other is the application of classical method using modified datasets. Modified dataset preserve all the information to select the important features. Therefore, in these findings, we use interquartile range (IQR) for outlier modification. If Q_1 and Q_3 are the first and third quartiles respectively, then IQR is defined by IQR = Q_3-Q_1. An observation is said to be outlier if it does not belongs to the interval [Q_1- β × IQR, β × IQR+Q3], where β = 1.5. *K*-nearest neighbors approach (Troyanskaya et al 2001) dependent on parameter tuning and it works through three stages: (i) distance measure, (ii) choice of *K* and (iii) adaption method.

(i) Distance measure

The distance measure sometimes called as dissimilarity measure. Suppose we have two instances x_i and x_j, the smaller distance between them represent the higher similarity. The widely used distance measures are Euclidean distance and Manhattan distance. The Euclidean and Manhattan distance *between* $x_{i\ and}$ x_j is defined as

$$d_{euclidean}\left(x_{i,}x_j\right) = \sqrt{\sum_{p=1}^{G} w_p\left(x_{i,p} - x_{j,p}\right)^2} \qquad (1)$$

$$d_{\mathrm{manhattan}}\left(x_{i,}x_j\right) = \sum_{p=1}^{G} w_p\left|x_{i,p} - x_{j,p}\right| \qquad (2)$$

where *G* denotes the total number of features, and w_p is the normalized weight of *p* th feature.

(ii) Selection of the neighborhood, *K*

An important parameter of *K*-NN approach is *K*. In this approach the *K*-should be given in advance and it is also dependent on dataset used for *K*-NN imputation. Many researchers only consider *K*=1 (Walkerden and Jeffery 1999), some others consider *K* between 1 to 3 (Mendes et al. 2003). The best results are found when K=1 to 5 (Li et al. 2009). *K* can also be found by the square root of the number of instances (Kocaguneli et al. 2012). However, the optimal value of *K* can also be determined using a cross-validation approach.

(iii) Adaption method

In this stage we obtain estimates for the missing values. There are few common ways of adaptations to estimate the missing values: mean, median, inverse distance weighted mean (IDWM), inverse rank weighted mean (IRWM) and so on.

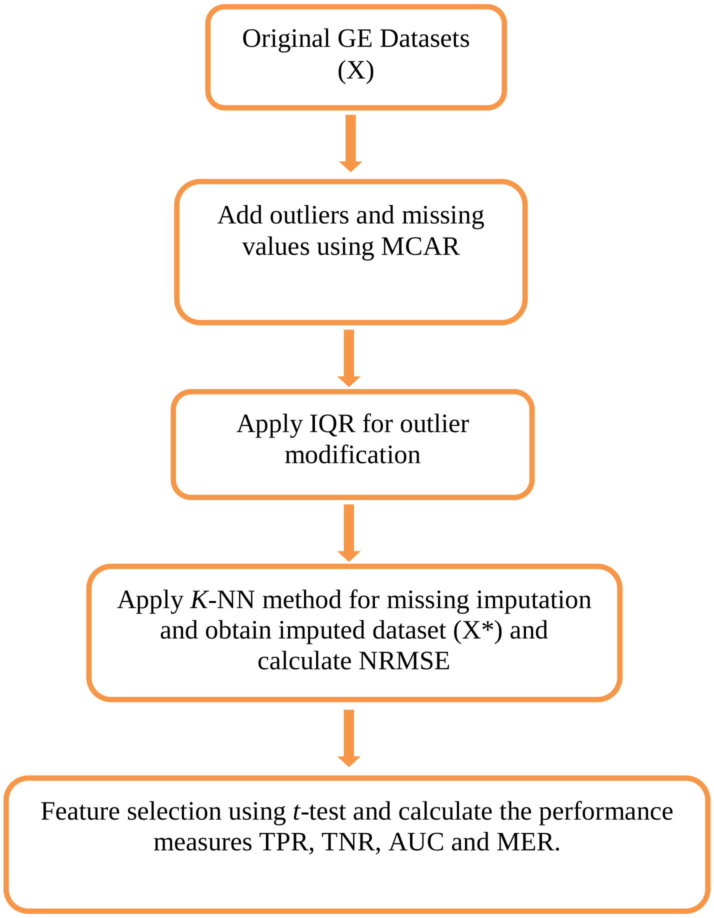

Original GE Datasets (X)
Add outliers and missing values using MCAR
Apply IQR for outlier modification
Apply *K*-NN method for missing imputation and obtain imputed dataset (X*) and calculate NRMSE
Feature selection using *t*-test and calculate the performance measures TPR, TNR, AUC and MER.

Fig. 1. Work flow of the proposed procedure.

Performance evaluation

The performance of K-NN and improved K-NN imputation methods were evaluated by most commonly used measure normalized root mean squared error (NRMSE). NRMSE is the difference between imputed and true values defined as follows:

$$NRMSE = \sqrt{\frac{\text{mean}\,(X^{\text{true}} - X^{\text{imputed}})^2}{\text{var}\,(X^{true})}}$$

Where, X^{true} is the true data and X^{imputed} is the imputed data.

We also used *t*-test as a feature selection method (FS) to evaluate the performance of the proposed missing imputation approach with traditional K-NN. Accordingly we used the following performance measures:

True positive rate (TPR) = TP / TP + FN, False positive rate (FPR) = FP / (FP + TN), Accuracy (ACC) and area under the receiving operating characteristics (ROC) curve (AUC). Where, TP, TN, FP and FN are the number of true positives, number of true negatives, number of false positives and number of false negatives, respectively. The flowchart of the proposed procedure has been shown in Fig. 1.

Simulated dataset

We applied K-NN and our proposed improved K-NN method in the simulated dataset. The simulated dataset was generated using the following one-way ANOVA model developed by Kerr et al. (2000):

$$x_{jk} = \mu_j + \epsilon_{jk};\, (j = 1,2;\, k = 1,2, \dots, n_j) \qquad (3)$$

where, x_{jk} is the *k*th observed expression of a gene in the *j*th condition, μ_j is the mean of all expressions of a gene in the *j*th condition and ϵ_{jk} is the random error term that follows $N(0, \sigma^2)$. The outlying datasets were generated by multiplying a constant (say, 5) with the mean of equation (3). We introduced varying percentages of missing values (1%, 5% and 10%) under the missing completely at random (MCAR) assumption.

Results and Discussion

To investigate the performance of the proposed missing value imputation method in a comparison of classical K-NN imputation approach, we generated 100 datasets from one-way ANOVA model using equation (3) with ($n_1 = n_2 = 10$) samples. The gene expression profiles of 1000 genes were generated with ($n_1 + n_2$) = 20 samples for each of the dataset. The number of DE gene is set to 200 and the rest of the 800 genes are considered as the non-DE genes. The mean and common variance for each group were set as $(\mu_1, \mu_2) \in c(3, 5)$ and σ^2= 0.1. Here we have considered 1%, 5% and 10% missing values under the MCAR assumption for each of the dataset. We also added different percentages of outliers (1%, 5% and 10%) in the datasets with missing values. We first calculated the average values of NRMSE between original and imputed data matrices by K-NN and proposed methods with different rates of missing values in absence and presence of outliers. The values of NRMSE against different rates of missing values in absence and in presence of outliers have been plotted in Fig. 2. From Fig. 2a we can observe that in absence of outliers with different rates of missing values (1%, 5% and 10%) both K-NN and proposed method produces almost similar values of NRMSE. However, the

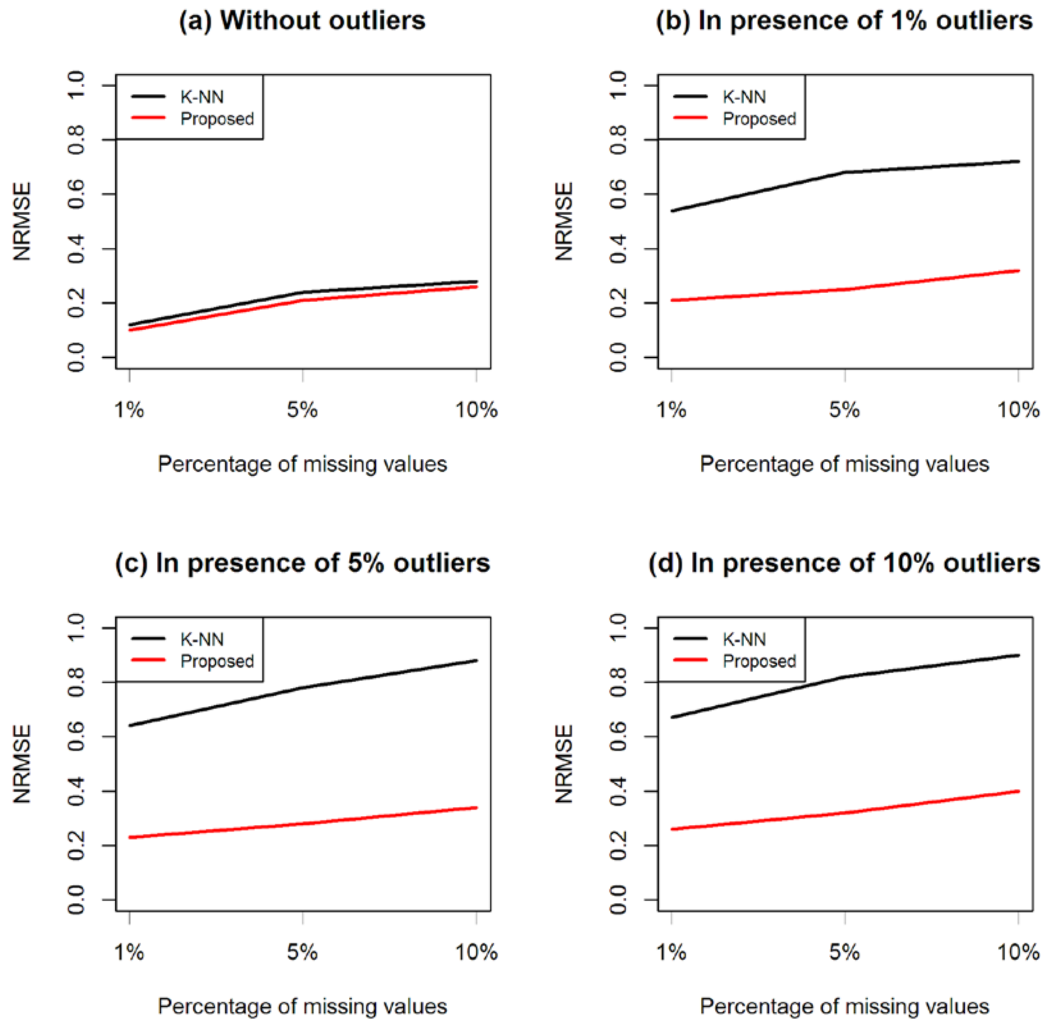
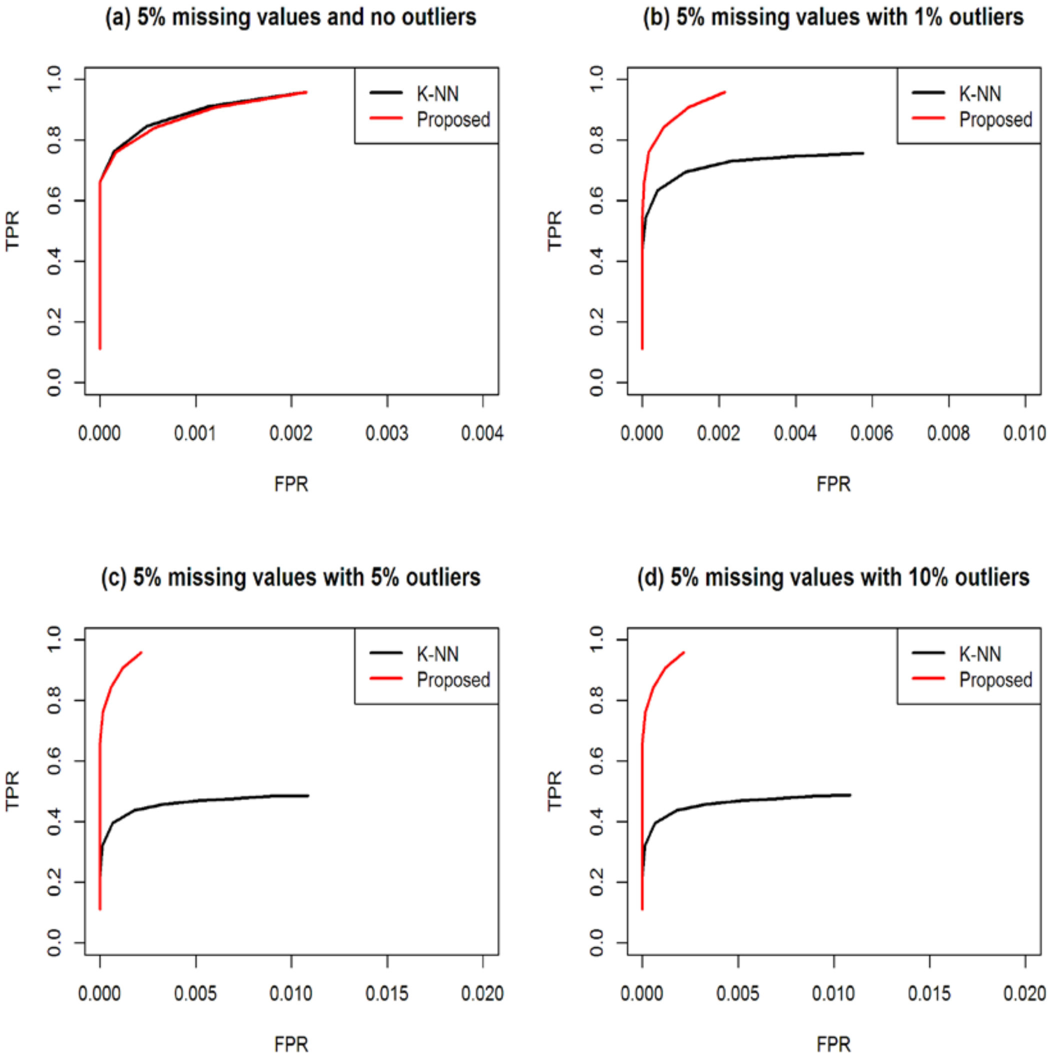


Fig. 2. Plot of NRMSE against different rates of missing values. (a) without outliers, (b) in presence of 1% outliers, (c) in presence of 5% outliers, and (d) in presence of 10% outliers.

proposed method outperforms K-NN method in presence of outliers (1%, 5% and 10%) with different rates of missing values (Fig. 2b-d). Then we employ *t*-test for identification of DE genes from each of 100 imputed datasets and estimated different performance measures TPR, FPR, AUC, MER and ACC. The ROC curve has been presented in Fig. 3 using *t*-test based on 100 imputed datasets using K-NN and proposed method in absence and presence of outliers with 5% missing values. This figure also supports the results of Fig. 2. The average values of accuracies (ACC) for detection of 200 DE genes using t-test based on 100 imputed datasets by K-NN and proposed method have been summarized in Table 1. In this table the columns represent the different conditions of outliers and rows represent the different conditions of missing values. From this table we clearly notice that *t*-test with proposed method has produced larger accuracies (ACC) than the t-test with K-NN in presence of outliers (see bold text in Table 1). Nevertheless, in absence of outliers, both methods are performed alike. The boxplot of AUC values estimated by t-test based on 100 imputed datasets by K-NN and proposed methods has been shown in Fig. 4. From Fig. 4a we observe that in absence of outliers both K-NN and proposed method produces almost similar values of AUC at different percentages of missing values. Whereas, in presence of 1%, 5% and 10% outliers (Fig. 4b-d) the proposed method outperformed K-NN method. Therefore, from this simulation study we may conclude that the proposed method outperforms K-NN method in presence of outliers and in absence of outliers it keeps equal performance with K-NN while performing feature selection using t-test.

Table 1. Performance evaluation of K-NN and proposed method based on average ACC for different rates of missing values in absence and presence of outliers

Methods with different conditions		Without outliers	1% outliers	5% outliers	10% outliers
1% missing values	K-NN + t	0.991	0.8819	0.7322	0.7327
	Proposed + t	0.991	0.9826	0.9780	0.9776
5% missing values	K-NN + t	0.982	0.8717	0.7271	0.7183
	Proposed + t	0.982	0.9813	0.9780	0.9780
10% missing values	K-NN + t	0.978	0.8710	0.7230	0.7051
	Proposed + t	0.978	0.978	0.978	0.978



(a) 5% missing values and no outliers

(b) 5% missing values with 1% outliers

(c) 5% missing values with 5% outliers

(d) 5% missing values with 10% outliers

Fig. 3. Performance evaluation between K-NN and proposed method using ROC curve with 5% missing values and different conditions of outliers. (a) 5% missing values and no outliers, (b) 5% missing values with 1% outliers, (c) 5% missing values with 5% outliers, and (d) 5% missing values with 10% outliers.

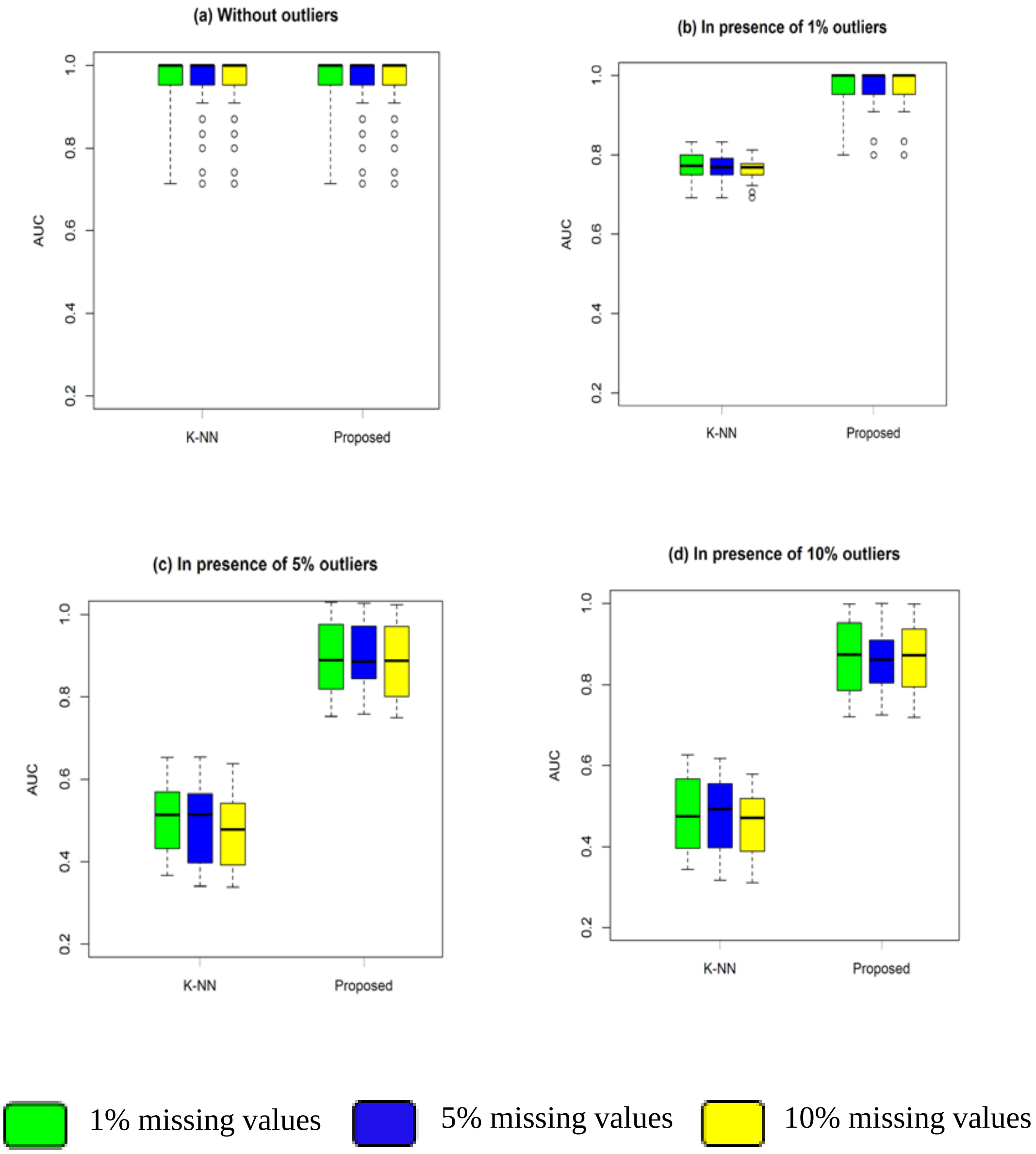


Fig. 4. Performance evaluation between K-NN and proposed method using boxplot of AUC values associated with varying percentages of missing values and outliers. (a) without outliers, (b) in presence of 1% outliers, (c) in presence of 5% outliers, and (d) in presence of 10% outliers. The green, blue and yellow boxes indicate 1%, 5% and 10% missing values, respectively under MCAR assumption.

Breast cancer real dataset

In this paper we have used 70 signature datasets from 78 patients. This dataset was taken from Buyse et al. (2006). Yan et al. (2015) also used this dataset to investigate the performance of their proposed method with other traditional methods. We first apply t-test in the original breast cancer dataset to identify DE genes. Using this test at 5% level of significance we identified 31 DE genes. We call these DE genes as true DE gene set. The heat map in Fig. 5 shows the expressions pattern of these genes. We added different percentages of missing values under MCAR assumption and corresponding to each percentage of missing values we consider various conditions of outliers (0%, 1%, 5% and 10%). We then employ the K-NN and proposed method in this incomplete and contaminated datasets to obtain imputed datasets. Then we investigated the true DE genes (31) identification performance of t-test based on imputed datasets by K-NN and proposed and calculated AUC values. We summarized these results in Table 2. From this table we revealed that in absence of outliers at different rates of missing values the K-NN + t and proposed + t performed almost equal. Whereas, in presence of outliers for every percentages of missing values (1%, 5% and 10%) the proposed + t out performed K-NN + t.

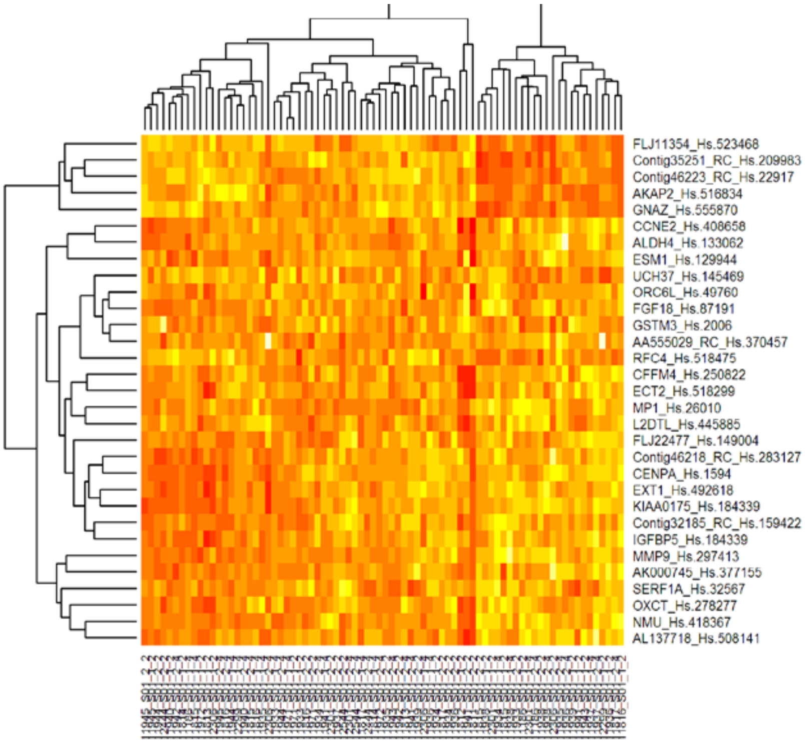


Fig. 5. Heatmap of 31 DE genes detected by t-test using original breast cancer dataset.

Table 2. Performance evaluation of K-NN and proposed method based on AUC for different rates of missing values in absence and presence of outliers for breast cancer dataset

Methods with different conditions		Without outliers	1% outliers	5% outliers	10% outliers
1% missing values	K-NN + t	0.982	0.783	0.728	0.693
	Proposed + t	0.982	0.971	0.970	0.966
5% missing values	K-NN + t	0.985	0.769	0.716	0.644
	Proposed + t	0.985	0.953	0.955	0.952
10% missing values	K-NN + t	0.987	0.730	0.684	0.637
	Proposed + t	0.988	0.948	0.933	0.921

Conclusion

Microarray GE data often contain missing values or outliers due to several steps of data generating process. Missing values and outliers can adversely affect the downstream analysis such as feature selection. There are several missing value imputation and outliers handling approaches in the literature. Unfortunately they conduct their task without regard to each other. Among the various missing value imputation techniques, K-NN is the oldest and popular one. However, it cannot deal with outliers. As a result, using K-NN imputed dataset for further analysis produces misleading results. Therefore, in the present findings, we have introduced an IQR rule for outlier detection and modification. Then we tag this IQR rule with popular K-NN approach. We investigated the performance of our proposed method with traditional K-NN approach through feature selection. Both simulation and real data analysis results confirmed that the proposed method outperforms the K-NN in presence of outliers with any percentages of missing values (1%, 5% and 10%). On the other hand, in absence of outliers, it keeps equal performance with K-NN.

References

Alam MA, Komori O, Calhoun V and Wang YP (2016). Influence function of multiple kernel canonical analysis to identify outliers in imaging genetics data. *In*: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016, pp. 210-219.

Alam MA, Lin HY, Calhoun V and Wang YP (2017). Kernel method for detecting higher order interactions in multi-view data: an application to imaging, genetics, and epigenetics. arXiv preprint arXiv:1707.04368.

Alizadeh AA, Eisen MB, Davis RE, Ma C and Lossos IS (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769): 503-511.

Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, Assignies MS, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F and Piccart MJ (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Nat Cancer Inst., 98: 1183-92.

Chiu CC, Chan SY, Wang CC and Wu WS (2013). Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Systems Biology, 7(6): 12.

Colombo PE, Milanezi F, Weigelt B and Reis-Filho JS (2011). Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. Breast Cancer Research, 13: 212.

De Risi JL, Iyer VR and Brown PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278: 680-686.

Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society: Series B (Methodological), 39: 1-38.

Fix E and Hodges J (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine Randolph Field.

Kim KY, Kim BJ and Yi GS (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics, 5(1): 160.

Kocaguneli E, Member S and Menzies T (2012). Exploiting the essential assumptions of analogy-based effort estimation. IEEE Trans. Softw. Eng., 38: 425-439.

Li T, Zhang C and Ogihara MA (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20: 2429-37.

Li YF, Xie M and Goh TN (2009). A study of project selection and feature weighting for analogy based software cost estimation. J. Syst. Softw., 82: 241-252.

Lockhart DJ and Winzeler EA (2000). Genomics, gene expression and DNA arrays. Nature, 405: 827-836.

Mendes E, Watson I, Triggs C, Mosley N and Counsell S (2003). A comparative study of cost estimation models for web hypermedia applications. Empir. Softw. Eng., 8: 163-196.

Nadon R and Shoemaker J (2002). Statistical issues with microarrays: Processing and analysis. Trends in Genetics, 18(5): 265-271.

Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H and Herzel H (2000). Normalization strategies for cDNA microarrays. Nucleic Acids Research, 28(10): E47.

Shahjaman M, Kumar M, Ahmed MS, Begum AA, Islam SMS and Mollah MNH (2017a). Robust feature selection approach for patient classification using gene expression data. Bioinformation, 13(10): 327-332.

Shahjaman M, Kumar M, Mollah MMH, Ahmed MS, Begum AA, Islam SMS and Mollah MNH (2017). Robust significance analysis of microarrays by minimum β-divergence method. BioMed Research International, pp. 1-18.

Stekhoven DJ and Bühlmann P (2012). Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28: 112-118.

Troyanskaya O, Cantor M, Sherlock G and Brown P (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6): 520-525.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D and Altman RB (2001). Missing value estimation methods for DNA microarrays. Bioinformatics (Oxford, England), 17(6): 520-525.

Tuikkala J, Elo L and Nevalainen O (2006). Improving missing value estimation in microarray data with gene ontology. Bioinformatics, 22(5): 566-572.

Walkerden F and Jeffery R (1999). Empirical study of analogy-based software effort es- timation. Empir. Softw. Eng., 4: 135-158.

Wang S and Cheng Q (2006). Microarray analysis in drug discovery and clinical applications. Methods Molecular Biology, 316: 49-65.

Yan L, Tian L and Liu S (2015). Combining large number of weak biomarkers based on AUC. Stat. Med., 34: 3811-3830.

(Manuscript received on February 02, 2018 and revised on July 30, 2018)