Evaluation of Missing Values Imputation Methods in cDNA Microarrays Based on Classification Accuracy

Vidan Fathi Ghoneim
Biomedical Engineering Dept.
Misr University for Science and Technology
Six of October City, Egypt
vida_eng@yahoo.com

Nahed H. Solouma
Engineering Applications Dept.
NILES, Cairo University
Giza, Egypt

Yasser M. Kadah
Biomedical Engineering Dept.
Faculty of Engineering, Cairo University
Giza, Egypt

Abstract— **Many attempts have been carried out to deal with missing values (MV) in microarrays data representing gene expressions. This is a problematic issue as many data analysis techniques are not robust to missing data. Most of the MV imputation methods currently being used have been evaluated only in terms of the similarity between the original and imputed data. While imputed expression values themselves are not interesting, rather whether or not the imputed expression values are reliable to use in subsequent analysis is the major concern. This paper focuses on studying the impact of different MV imputation methods on the classification accuracy. The experimental work was first subjected to implementing three popular imputation methods, namely Singular Value Decomposition (SVD), weighted K-nearest neighbors (KNNimpute), and Zero replacement. The robustness of the three methods to the amount of missing data was then studied. The experiments were repeated for datasets with different missing rates (MR) over the range of 0-20% MR. In applying supervised two class classification we adopted a twofold approach, introducing all genes expressions to the classifiers as well as a subset of selected genes. The feature selection method used for gene selection is Fisher Discriminate Analysis (FDA), which improved noticeably the performance of the classifiers. The retained classifiers accuracies using imputed data after applying the three proposed imputation methods show slight variations over the specified range of MR. Thus, assessing that the three imputation methods in concern are robust.**

Keywords-microarrays; imputation; evaluation; classification

I. INTRODUCTION

The data from microarray experiments is usually in the form of large matrices of gene expressions (rows) under different experimental conditions or different subject samples (columns) and frequently with some values missing. Missing values (MV) occur from diverse reasons, including insufficient resolution, image corruption, or simply due to dust or scratches on the slide. Missing data may also occur due to systematic effects concerning the arrays [1]. Although microarrays technology has developed much during the past years, it is still underlying uncertainties, resulting in datasets with compromised accuracy because of the existence of these MVs. Since many algorithms for microarray analysis require a complete data matrix as the input, MVs must be imputed before the subsequent analyses. If a complete dataset is required, as is the case for most clustering tools, data analysts typically have three options before carrying out analysis on the data: they can either discard the genes (or arrays) that contain missing data, replace missing data values with some constant (zero), or estimate (impute) values of missing data entries [1] . Most commonly applied statistical techniques for dealing with missing data are model-based approaches. The influence of specific modeling assumptions in the employed methods is minimized as much as possible in this work in agreement with [1]. Although Normalized Root Mean Square Error (NRMSE) can be calculated to measure the imputation accuracy, since the original values are known. This method is problematic for two reasons. First, most of the time the selection of artificial missing entries is random and thus is independent of the data quality whereas imputing data spots with low quality is the main scenario in real world. Secondly, in the calculation of the NRMSE, the imputed value is compared against the original, but the original is actually a noised version of the true signal value, and not the true value itself. Although this randomized MV generating scheme is widely used, it ignores the underlying data quality. Therefore, a real dataset is employed in this study and its MVs are imputed. Many imputation methods are available that utilize the information present in the non-missing part of the dataset. Such methods include, for example, the weighted K-Nearest Neighbors (weighted KNN) and Singular Value Decomposition (SVD) approach [1], the Local Least Squares imputation (LLS) [2], and Bayesian Principal Component Analysis (BPCA) [3]. While most of the imputation algorithms currently being used have been evaluated only in terms of the similarity between the original and imputed data points, the success of imputation methods should be evaluated also in other terms, for example, based on clustering methods to identify groups of co-regulated genes, disease classification and their biological interpretation, that are of more practical importance for the biologist [4]. A recent

978-1-4244-7000-6/11/$26.00 ©2011 IEEE.

study investigated the influence of imputation on the detection of differentially expressed genes from cDNA microarray data. They proposed a method for imputation named (LinImp), fitting a simple linear model for each channel separately, and compare it with the widely used KNN method [5]. Another study considered impact of imputation on the related downstream analysis, disease classification; they discovered that while the ZERO imputation resulted in poor classification accuracy, the KNN, LLS and BPCA imputation methods only varied slightly in terms of classification performance [6]. Two other studies investigated the effect of MV and their imputation on the preservation of clustering solutions. One study concentrated on hierarchical clustering and the KNN imputation method; their main findings were that even a small amount of MV may dramatically decrease the stability of hierarchical clustering algorithms and that the KNN imputation rarely improves this stability [7]. The second one aimed to investigate the effect of MV on the partitioned clustering algorithms, such as k-means, and to find out whether more advanced imputation methods, such as LLS, Support Vector Regression (SVR) and BPCA, can provide better clustering solutions than the traditional KNN approach [4]. In this work in correspondence to [6] the effect of imputation methods in terms of classification accuracy was rather chosen to be investigated. The impact of three commonly used data imputation methods: weighted KNN, SVD, and Zero replacement are studied on the performances of three classifiers: Support Vector Machine (SVM), Euclidean distance and Backpropagation Neural Network (BpNN).

II. METHODS

A. Dataset Description

The data set used in this study was downloaded from: http://www.ncbi.nlm.nih.gov/geo. The dataset includes twenty samples presenting thyroid papillary cancer tissues versus patient matched adjacent non-tumor thyroid tissues (GEO accession: GSE3950). The data set is acquired by Genepix from cDNA microarrays spotted by a total of 16200 genes. All gene expressions are log base two transformed. This transformation sufficiently reduces the effect of outliers on gene similarity determination [1].

B. Experimental Methods

First, the original data set (16200 genes x 20 samples) is compromised to five data subsets each at a different missing rate (MR): 0%, 5%, 10%, 15%, and 20%. Using threshold 5% means that out of the 16200 genes some genes are selected, those genes with missing expression data points <=5% of the 20 data points (20 samples). This corresponds to 10008 genes either with complete measurements (representing data points of 20 samples) and genes with only one missing measurement (1MV representing 1 data point of 1 sample). The distributions of the genes with different MR are shown in Table I. Secondly, three well known MV imputation methods are implemented using Matlab. The imputation methods used in this work are weighted KNN, SVD, and Zero replacement. The three techniques employed the above described subset data sets to examine imputation methods robustness at different MR. Three classifiers are used in evaluating these imputation methods: SVM, Euclidean distance, and BpNN.

TABLE I. DISTRIBUTION OF GENES WITH DIFFERENT MR

MR	0%	5%	10%	15%	20%
Number of Retained Genes	9228	10008	10488	12196	13594
Number of Genes with MV	0	780	1260	2968	4366

C. Weighted KNN Algorithm

Weighted KNN is a standard MV imputation method introduced in [1]. The KNN-based method takes advantage of the correlation structure in microarray data by selecting genes with expression profiles similar to the gene of interest to impute missing values. Accordingly, the imputation process is typically divided into two steps. In the first step, a set of genes nearest to the gene with a missing value is selected. To explain the way that this step works, consider gene g in experiment *i* so, let's say $V_{g,i}$ is missing value, thus, this method would find k other genes, with a known value for experiment *i*, and with the expression profile most similar to g considering all the experiments. The authors examined a number of metrics for gene similarity (Pearson correlation, Euclidean distance, variance minimization). Euclidean distance was found to be a sufficiently an accurate norm in spite of its sensitivity to outliers which could be present in microarray data. The reason behind this finding lies in using the log-transform to normalize the data, which in turn reduces the effect of outliers on gene similarity determination. The second step involves the prediction of the MV using the observed values of the selected genes. At this stage, a weighted average of values in experiment *i* from the k closest genes is then used as an estimate for the MV in gene g. In the weighted average, the contribution of each gene is weighted by the similarity of its expression to that of gene g. In this work this method is employed to take advantage of the correlation structure in microarray data but not along similar genes for one experiment but rather along similar arrays (samples) for one gene. The number of neighboring samples k is varied to monitor the manner of this method at different conditions.

D. SVD Algorithm

SVD is used to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the data set. SVD is studied and implemented in the context of microarray data by [1]. This study referred to these patterns, which in this case are identical to the principle components of the gene expression matrix as eigen genes and introduced the following formula in (1).

$$A_{mxn} = U_{mxm} \sum_{mxn} V^{T}_{nxn}. \quad (1)$$

where matrix V^T contains eigen genes, whose contribution to the expression in the eigen space is quantified by corresponding eigen values on the diagonal of matrix Σ. The most significant eigen genes are identified by sorting the eigen genes based on their corresponding eigen value. Once k most significant eigen genes from V^T are selected, a MV *j* in gene *i* is estimated by first regressing this gene against the k eigen genes and then use the coefficients of the regression to

reconstruct j from a linear combination of the k eigen genes. The j^{th} MV value of gene i and the corresponding j values of the k eigen genes are not used in determining these regression coefficients. As SVD can only be performed on complete matrices; therefore, 1s are substituted in this study as an initial estimation for all MV in matrix A, obtaining A'. The first principal component is used, here termed eigen genes corresponding to the highest eigen value each time we used a data matrix at different MR. Thus, the model and imputation are expressed in (2) and (3) respectively:

$$g_i(C) = \sum_{k=1}^{k} \beta_k \, v_k(C). \quad (2)$$

$$\hat{g}_i(\approx C) = \sum_{k=1}^{k} \widehat{\beta_k} \, v_k(\approx C). \quad (3)$$

E. Machine Learning Techniques

This study focuses on two-class classification using hard-margin SVM with a first-degree dot product kernel function, BpNN multiple-layer network with nonlinear differentiable transfer functions, and Euclidean classifiers. A known problem in classification specifically, and machine learning in general, is to find ways to reduce the dimensionality of the feature space to overcome the risk of "over fitting". Data over fitting arises when the number of features is large (16200 genes) and the number of training patterns is comparatively small (20 samples). In our attempt to extract differentially expressed genes, feature selection is applied using Fisher Discriminate Analysis (FDA). FDA is a simple algorithm applied mainly to reduce the dimensionality of the data thus outputting the most discriminate features (genes expressions), according to the value of the Fisher factor j given in (4):

$$j(gene) = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + \sigma_2)}. \quad (4)$$

where, μ1, σ1 and μ2, σ2 are the means and variances of the two classes; normal and tumor sets respectively. It is clear from (4) that, j has a higher value when the feature value differs greatly in the two classes and vice versa. Using the cross validation approach, the available dataset is used for both phases testing and training.

III. Results

A. Imputation Methods

The three imputation methods studied weighted KNN, SVD, and Zero replacement show robustness at different MR. There are very slight differences in the classification accuracies each time a different imputation method is used as illustrated in Fig. 1. From the results obtained the following is observed, weighted KNN performance slightly decreases at higher MR only when used with the FDA based Euclidean classifier. It is very stable when used with FDA based SVM and Euclidean distance classifiers without applying any gene selection method. All the illustrated results are obtained at tuning $k=5$. For SVD, the results show that it is a robust imputation method. It gives slight changes in terms of classification accuracy at different MR when used with FDA based Euclidean and FDA based SVM classifiers. Compared to weighted KNN and SVD, Zero replacement performed well with all used classification attempts as shown in figure1 and at different MR. Although the three imputation methods perform better for the FDA based SVM classifier at lower MR (5%, 10%), it is observed that the weighted KNN, SVD and Zero imputation algorithms still make little difference in affecting classification performances.

B. Calssification

The sensitivity of the alternative classification algorithms to various data imputation methods is addressed in this work, which can be problematic for some uncertainties in interpreting and comparing the disease classification results based on cDNA microarray data. This is clearly sensed in the obtained results using BpNN. When applying this classifier its results deteriorate greatly with the different MV imputation methods, different parameters (k values) and different MR. For example without FDA for gene selection it yields at 10% MR: 75% overall accuracy (70% sensitivity, 80% specificity) using weighted KNN, 65% overall accuracy (30% sensitivity, 100% specificity) using SVD and 60% overall accuracy (70% sensitivity, 50% specificity) using Zero replacement. Moreover, its results fluctuate greatly through successive iterations. In the same example mentioned above it yields 60% sensitivity and fluctuates in the range 60-90% for specificity using weighted KNN, 30-50% sensitivity 100% specificity using SVD and 70-80% sensitivity 50-80% specificity using Zero replacement. When applied with FDA there is progress in the results as it yields higher sensitivity and specificity % and consequently yields higher over all accuracy. However, the fluctuation problem in the results obtained is not resolved by applying FDA gene selection. This reveals that BpNN cannot be applied as a reliable classifier for the data set used in this study. The Euclidean distance classifier is simple, intuitive, and stable. It performs remarkably well as shown in the results plotted in figure1. All its results are slightly changed with the different imputation methods and MR. Moreover, among the three proposed classifiers it yields the most stable and accurate performance based on all genes measurements, see Fig. 1(a). Its robustness to different imputation methods and over different MR is observed. It yields 75% accuracy based on all genes under all conditions and yields slightly different accuracies over the range 80-85% based on selected genes using FDA. For the SVM classifier it shows robustness to different imputation methods and MR based on FDA. SVM reaches its best accuracy 90% (80% sensitivity, 100% specificity) based on FDA when using all MV imputation methods at 5% and 10% MV rates. On the other hand, it does not yield satisfactory results when applying it on all genes. This fact directed the study to use a gene selection method.

IV. Discussion

As the purpose of this study is to assess the MV imputation methods, rather than search for the best classifier. The results of only the reliable classifiers are presented: Euclidean distance and FDA based SVM classifiers as introduced in figure1. While conducting the experiments for weighted KNN, it is observed that having k column neighbors C_j for j=1,2,…,k used in imputing a MV in a column sample S, then C and S, both belong to the same class. This observation validates the robustness of applying weighted KNN across samples

(columns). By exploring the results in figure1, we show that the commonly used weighted KNN is a robust imputation method and is comparable with other complicated methods such as SVD. This finding agrees with [6]. The three imputation methods weighted KNN, SVD, and Zero replacement show robustness at different MR. There are very slight differences in the classification accuracies at different MR using the three imputation methods. In all experiments, the imputed datasets at different MR thresholds give classification accuracies very close to the accuracy of the complete dataset filtered of all missing data compromising 9228 genes out of 16200 genes corresponding to almost 57% of the total number of genes and 0% MR. This gives us a clue that there are still more informative genes beyond the filtered 9228 genes but having MV. Good imputation of these genes missing points contribute in maintaining classification accuracy if not improving it as being observed in the results of almost all classifiers as shown in Fig. 1. Having preexisting classes, discriminate analysis or supervised learning methods are more appropriate and more efficient than clustering methods. And thus in this work this direction is headed. In this concern Euclidean distance classifier shows robustness to the different MV imputation methods and MR either when using all measured genes or feature genes selected using FDA. This is not the case with SVM and BpNN as it is a necessary step to apply gene selection based classification. Using the linear discriminate FDA, it performs remarkably well, where it enhances the Euclidean distance and SVM classification accuracies. Raising the Euclidean distance classifier from 75% over all accuracy being based on all genes to 80-85% accuracy based on FDA. Among the three kinds of classifiers evaluated in this study, based on either selected genes or all genes, Euclidean distance classifier is robust to varied MV imputation methods, while the BpNN classifier is the unstable one. The SVM classifier is robust to varied MV imputation methods and MR, only when built on gene selection method FDA. This is may be due to the high dimensional and noisy data set employed.

V. CONCLUSION

Weighted KNN, SVD, and Zero imputation methods are robust to different MR. In this work weighted KNN is implemented in a different manner other than that introduced in other studies. The obtained results are satisfactory and greatly assess the algorithm robustness. Taking advantage of weighted KNN method that provides accurate estimation for MV in genes that belong to the same small tight expression cluster, it is applied here along samples that belong to the same class instead. It is thought that it would be more logical to find relations between similar samples of the same class for a certain gene rather than to relate different genes along the same sample. Euclidean distance classifier performs well with such high dimensional data either based on all genes measurements or on selected genes. Using FDA for feature gene selection enhances the performance of both Euclidean distance and SVM classifiers. Where both classifiers show robustness to the different imputation methods and different MV rates, meaning that their classification performances are slightly affected by making either or both changes.

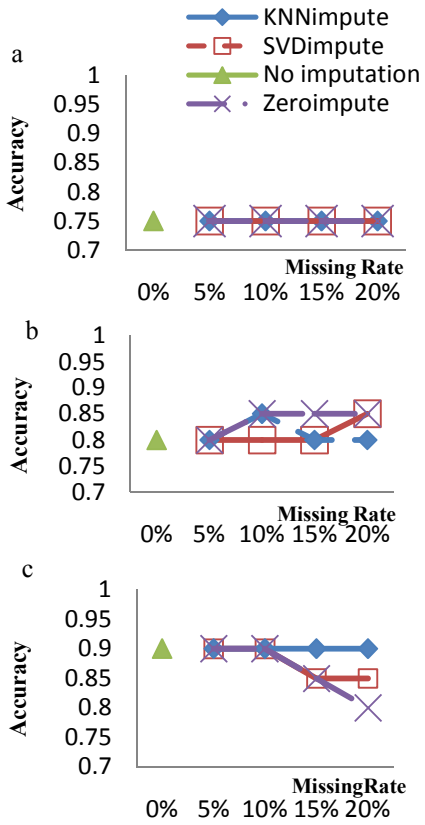

Figure 1. Classification accuracy of (a) Euclidean distance,(b) FDA-Euclidean distance ,(c) FDA-SVM

REFERENCES

[1] O. Troyanskaya, M.Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman, "Missing value estimation methods for DNA microarrays." Bioinformatics,vol. 17, pp. 520-525, 2001.

[2] H. Kim, G.H. Golub and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation." Bioinformatics, vol. 21,pp. 187-198, 2005.

[3] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data". Bioinformatics, vol. 19, pp. 2088–2096, 2003.

[4] J. Tuikkala, L.L. Elo, O. S. Nevalainen and T. Aittokallio, "Missing value imputation improves clustering and interpretation of gene expression microarray data." BMC Bioinformatics,vol. 9, pp. 202, 2008.

[5] I. Scheel, M. Aldrin, I. K. Glad, R. Sørum, H. Lyng and A. Frigessi, "The influence of missing value imputation on detection of differentially expressed genes from microarray data." Bioinformatics, vol. 21, pp. 4272–4279, 2005.

[6] D. Wang, Y. Lv, Z. Guo, X. Li, Y. Li, J. Zhu, D. Yang, J. Xu, C. Wang, S. Rao and B. Yang, "Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules." Bioinformatics, vol. 22, pp. 2883–2889, 2006.

[7] A. G. de Brevern, S. Hazout and A. Malpertuy, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering." BMC Bioinformatics, vol. 5, pp.114, 2004.