28th NATIONAL RADIO SCIENCE CONFERENCE
(NRSC 2011)
April 26-28, 2011, National Telecommunication Institute, Egypt

C2

The Impact of Missing Values Imputation Methods in cDNA Microarrays on Downstream Data Analysis

Vidan Fathi Ghoneim[1], Nahed H. Solouma[2], Yasser M. Kadah[3]
[1]Biomedical Engineering Department, Misr University for Science and Technology, Six of October City, Egypt
[2]Engineering Applications Department, NILES, Cairo University, Giza, Egypt
[3] Biomedical Engineering Department, Cairo University, Giza, Egypt

ABSTRACT

DNA microarray is a high throughput gene profiling technology employed in numerous biological and medical studies. These studies require complete and accurate gene expression values which are not always available in practice due to the so-called microarray missing value (MV) problem. Many attempts have been held to deal with this problem. MV imputation algorithms to estimate MV have been designed as the most reliable solution for this problem. Many of the schemes introduced to evaluate these algorithms are limited to measuring the similarity between the original and imputed data. While imputed expression values themselves are not interesting, rather whether their impact on downstream analysis is the major concern. In this work the success of three MV imputation methods is measured in terms of Normalized Root Mean Square Error as well as classification accuracy and detection of differentially expressed genes (biomarkers) for distinguishing different phenotypes. The classification accuracies computed on the original complete and imputed datasets gave a practical evaluation of the three imputation methods where it showed slight variations among them. Some of the identified biomarkers were found to be Gene Ontology annotated coding for proteins involved in cell adhesion/motility, lipid/fatty acid transport and metabolism, immune/defence response, and electron transport.

Keywords: microarrays, missing values, imputation algorithms, classification accuracy

I. INTRODUCTION

Microarray technology offers a powerful tool for modern biomedical research. Using microarrays, expression levels of thousands of genes can be measured simultaneously on a single chip what is called gene expression profiling. One microarray technology that is widely employed is cDNA microarrays, where microscope slides are spotted with thousands of cDNA fragments. The arrays are hybridized to fluorescent-labelled cDNAs generated by reverse transcription of RNA isolated from the cell sample or tissue under investigation. In standard terminology, the cDNAs spotted on to the arrays are called probes, and the cDNAs in the samples are called target genes. cDNA microarrays generate one or two channel data. In two channel use, the arrays are hybridized to a mixture of two samples (e.g. disease and normal), each labelled with a different dye (green and red) to allow for quantitative measurement of target genes abundance in the samples. In one channel use, each array is hybridized to a single sample, labelled with a single dye. The arrays are laser scanned at the wavelength(s) appropriate to the dye(s) used, and the images are processed to extract data for analysis. In one channel studies, the data usually consists of a measure for the spot intensity and its local background, for each spot on the array. In two channel studies, this is available for both dyes.

What we needed to measure from cDNA microarrays is these spots net intensities as it reflect the relative abundance of the corresponding target genes (gene activity) in two different samples (derived from competitive, two channel hybridizations). But these measurements may be biased by diverse effects such as efficiency of RNA extraction, reverse transcription, label incorporation, exposure, scanning, spot detection, etc. Furthermore, there are systematic effects due to characteristics of the array, such as effects of different probes, spotting effects, region effects, etc... Gene activity estimation has an impact on subsequent data analysis and interpretation. If the measured intensity of a gene is not due to the gene activity itself, subsequent analysis using this erroneous estimate will, of course, be misleading. Before any subsequent data analysis, the spots intensities are corrected for the background intensities. Most image analysis programs return 'local' background intensities as mentioned earlier. It is obtained from the mean or median of the pixel intensity values surrounding each spot. Local background is arguably an unbiased estimate of the local nonspecific signal, so subtracting it from the foreground intensity gives in principle an unbiased estimator of the true signal due to hybridization. Although well motivated, this traditional approach produces corrected intensities with undesirable statistical properties. It produces negative intensities whenever the background intensity is larger than the foreground intensity, leading to missing log ratios,

sometimes for a substantial proportion of probes on an array, what is called microarray missing value (MV) problem [1].

Many algorithms for microarrays data analysis require complete data such as hierarchical clustering, k-means clustering, and self organizing maps. Typically, 1-10% of the data on microarray can be missing, affecting up to 95% of the genes and so data analysts have limited options before carrying out analysis on the data. They can either discard the genes (or arrays) that contain missing data, repeat the experiment which is not only costly and time consuming, but also cannot come to identical gene expression profiling results, or estimate (impute) values of missing data entries [2]. The latter option is the most appropriate where imputation methods 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 [2], the Local Least Squares imputation (LLS) [3], Fixed Rank Approximation Algorithm (FRAA) [4], and Bayesian Principal Component Analysis (BPCA) [5].

Most of the imputation algorithms currently being used have been evaluated only in terms of the similarity between the original and imputed data points. For example 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. Based on this, the success of imputation methods should be evaluated in other terms besides NRMSE. The imputed expression values themselves are not interesting, while whether or not the imputed expression values can be trusted and used in downstream applications is the major concern. Evaluation can be 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 [6]. A recent 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 [7]. Another study considered the impact of imputation on 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 [8]. 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 [9]. 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 [6].

In this work in correspondence to [8] the effect of imputation methods in terms of downstream data analysis was rather chosen to be investigated besides using NRMSE to evaluate the imputation methods. Classification and gene selection processes were conducted to evaluate the success of three commonly used data imputation methods: KNN, SVD, and Zero replacement. For the classification process we employed two classifiers: Support Vector Machine (SVM) and Euclidean classifiers. The classifiers were trained using two cross validation methods: k-fold and hold out. For the gene selection process two samples t-test and Fisher Discriminate Analysis (FDA) were implemented within the cross validation cycles to provide an unbiased estimate of the accuracy rate of the classifiers. This process yielded significant genes that could be considered as biomarkers to distinguish between different disease phenotypes. In our case study we applied the experimental work on invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) phenotypes, which are the two major histological types of breast cancer [10].

II. METHODS

1. Imputation Methods

There are several alternative ways of dealing with MVs. This paper considers the common used imputation methods based on the review of MV imputation methods literature.

1.1 Zeroimpute

It fills the missing values with zero. Although it is very simple and efficient, obviously, Zeroimpute could artificially create erroneous relationships between genes since the integrity and usefulness of the non missing data in the expression matrix are not taken into account.

1.2 KNNimpute

It is a standard MV imputation method introduced in [2]. The KNN-based method takes advantage of the correlation structure in microarray data by selecting one or more genes with expression profiles similar to the gene of interest to impute MV. 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 other than *i*. 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 in gene *g* by either replacing the observed value of if one closest gene is selected or using the weighted average of values of the *k* closest genes in experiment *i.* In the weighted average, the contribution of each gene is weighted by the similarity of its expression to that of gene *g*. In our experimental work, this method is employed to take advantage of the correlation structure in microarray data but not along similar genes in one experiment but rather along similar arrays (experiments) for one gene [11].

1.3 SVDimpute

It 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 through the SVD of the expression matrix. SVD is studied and implemented in the context of microarray data by [2]. This study referred to these patterns, which in this case are identical to the principle components of the gene expression matrix as eigen genes. The most significant eigen genes are identified by sorting the eigen genes based on their corresponding eigen values. Once *k* most significant eigen genes 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, zeros are substituted in this study as an initial estimation for all MV in matrix A, obtaining A'. The first principal component is used in here termed "eigengene" corresponding to the highest eigen value. For more convenience we implemented the SVD on each class of data independently to avoid gene expressions of different phenotypic classes (samples) to influence the imputation [11].

2. Validation Methods

To evaluate the success of MV imputation methods we worked on two validation schemes. First, measuring Normalized Root Mean Square Error (NRMSE) which was more commonly adopted [2]-[9], [12]. The NRMSE measurement presumes that all the observed gene expression values, which are not considered as missing values, should accurately measure the hybridization intensities of the genes on the microarrays. This presumption, however, is not necessarily the case as discussed in the introduction [12]. The NRMSE is the root mean squared difference between the original **y** and imputed values **y'** of the missing entries, divided by the root mean squared original values in these entries as shown in (1):

$$NRMSE = \sqrt{\frac{mean((y - y')^2)}{mean(y^2)}} \quad (1)$$

where mean( ) stands for the arithmetic mean of the elements in its argument array.

Second, we relied on measuring MV imputation methods quality in terms of downstream microarray data analysis, which is the core interest of this work. Two-class classification was adopted together with finding gene biomarkers using gene selection methods. These biomarkers were referred to the Gene Ontology (GO) to identify their functions.

3. Machine Learning

3.1 Gene Selection

In our attempt to extract significant genes (biomarkers) also known as differentially expressed genes, feature selection is applied using Fisher Discriminate Analysis (FDA) and two samples t-test. 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 by (2):

$$J(gene) = (\mu_1 - \mu_2)^2 \Big/ \sigma_1 + \sigma_2 \quad (2)$$

where, $\mu 1$, $\sigma 1$ and $\mu 2$, $\sigma 2$ are the means and variances of the two classes; according to our study ductal and lobular sets respectively. It is clear that *j* has a higher value when the feature value differs greatly in the two classes and vice versa.

A standard statistical for detecting significant changes between the measurements of a variable in two groups is the t-test. We conducted a two samples t-test for each gene to identify significant changes in expression values between the IDC samples and ILC samples. Genes with p-values < 0.001 threshold were selected for the classification process. Using the cross validation approach as will be discussed in the coming sub-section, the employed dataset was subjected to feature selection in both phases of testing and training.

3.2 Classification and Cross Validation

Discrimination and clustering have been described as **class prediction** and **class discovery**. In the machine learning literature they are known as supervised and unsupervised learning. The learning in question being of the combinations of measurements, here gene expression values, which assign units to classes . In the statistical literature they are known as discrimination and clustering. The distinction is important. Clustering or unsupervised methods are likely to be appropriate if classes do not exist in advance. If the classes are preexisting, then discriminate analysis or supervised learning methods are more appropriate and more efficient than clustering methods. Cluster methods tend to be over used in microarray data analysis relative to discrimination methods . A common practice for example is to suppress existing class assignments, use an unsupervised learning technique to define new classes and assign the units to these classes, and then see how well the existing class assignments are reflected in the new classes. A more direct and efficient approach would be to use a supervised method to discriminate the classes in conjunction with a method such as cross validation to evaluate the repeatability of the results on new data. The efficiency of direct discrimination over clustering becomes increasingly important as the prediction problem becomes more challenging. Discrimination methods include linear discriminate analysis in various forms, nearest neighbor classifiers, classification trees, aggregating classifiers, neural networks and support vector machines.

In this work we employed the latter approach using Euclidean and SVM classifiers. Given a complete gene expression matrix with all samples being labeled with their class memberships, we first employed the k-fold cross validation to avoid the possible data over fitting problem. For doing the k-fold cross validation, the complete dataset is randomly partitioned into k equal parts. Each part of the k equal parts is used as the testing dataset at one time by removing its sample labels, while the rest (1- k) parts are used as the training dataset. This process is repeated for each of the k parts. Based on the classifier built on the training dataset, the sample labels of the testing dataset are predicted and compared with the original true sample labels. The percentage of the correctly predicted samples is the classification accuracy of the classifier. We carried out in this study 5-fold cross validation, where each time 5 or 6 samples were used for testing. The random partition process was repeated 10 times to cover almost the whole 57 data samples. We also employed hold out cross validation which returns logical index vectors for cross validation of N observations (57 samples) by randomly selecting P*N (approximately) observations to hold out for the evaluation set. P must be a scalar between 0 and 1. P was assigned to 0.1 corresponding to holding out 10% (approximately 5 samples) for testing at one time. And so the process was repeated 10 times as the case with k-fold cross validation method.

III. RESULTS

Given a complete microarray gene expression dataset (which also can be regarded as a dataset with missing ratio 0%), based on the uniform distribution, we randomly simulated 2 datasets for each of the missing ratios *r* (1%, 2%, 3%, 4%) making up a total of 4 datasets for each of the 2 simulations of the random generator algorithm. On each simulated dataset of the 8 datasets, all the three missing value imputation methods, KNNimpute, SVDimpute, and Zeroimpute, were run separately to estimate the missing values. Then, on both the original complete dataset and the imputed complete dataset, each of the two gene selection methods, t-test, and FDA was

applied on the randomly picked samples according to the cross validation methods used. Where two cross validation methods were used: k-fold cross validation and hold out cross validation. Applying k-fold cross validation with k=5 corresponds to choosing 1/5 out of the 57 samples about 5 samples for testing and leaving the other 4/5 about 52 samples for training. For the hold out cross validation 0.1 out of the samples were adjusted for testing in each cycle corresponding to approximately 10%, about 5 samples were hold out for testing. The random partition process was repeated 10 times to cover almost the whole 57 data samples using both cross validation methods. The gene selection method used was applied in each of the cycles of cross validation to provide an unbiased measure of the accuracy rate of the classifier. For the phenotypic information of the test samples if used in training the classifier, the estimate may be biased because of some re-substitution effects as discussed in [8]. The Euclidean classifier and the SVM classifier were then built based on the selected genes to predict the class of the testing samples.

1. Data Set Description

The data set used in this study was downloaded from Stanford Microarray Database (SMD): http://genome-www.stanford.edu/breast_cancer/lobular/. The dataset includes fifty seven samples represent thirty six invasive ductal carcinoma (IDC) and twenty one invasive lobular carcinoma samples (ILC). The data set is acquired by Genepix from cDNA microarrays spotted by a total of approximately 42000 clones [10]. The output Genepix results are stored in excel sheets giving a total of fifty seven raw data files for all samples.

1.1 Data Preprocessing

The raw data containing approximately 42000 has many redundant genes; many spots hybridized by the same gene. Also some genes were not present in all raw data files. A common set of genes for all raw data files was obtained compromising 15798 genes. And all expression values of redundant genes were averaged to give a single value for each gene. This step lowered a lot the number of MV in the data set as having gene replicates is one of the solutions for MV as implied by [2].The net data size we applied upon all the experiments in this work is then 15798genesx57samples. All gene expressions were log base two transformed. This transformation sufficiently reduces the effect of outliers on gene similarity determination [2]. Furthermore we applied Loess normalization to promote uniformity within arrays (samples). It is a technique for fitting smooth non-linear functions of a set of predictor variables to a continuous response variable also known as local regression. We figured an improvement in the classification results after imputing MV with normalized data rather than non normalized data [1]. Considering the uncertainty in relying on the NRMSE for evaluating imputation methods as discussed in the introduction. When the expression values of an input microarray dataset are all of high confidence, that is, they do accurately measure the actual spots intensities, NRMSE could be a better imputation quality measurement, considering both its effectiveness and its computational complexity [12]. All preprocessing steps were applied using Limma package in R - Bioconductor project.

1.2 Gene Filtering

Gene profiling experiments have genes that exhibit little variation in the profile and are generally not of interest in the experiment. These genes are commonly removed from the data. The variance and entropy for each gene expression were calculated and genes with variances and entropies less than the 10^{th} percentile were discarded. The obtained filtered set has reduced the total number of genes from 15798 to 8086. Since in this work, we focus on the idea of using gene selection based classification to evaluate missing imputation methods, rather than examining how the dataset quality affect the classification accuracy, the dataset size is not too much concerned here. Large dataset may extremely consume runtime with insignificant effect on the classification accuracy more over it might deteriorate the classification accuracy as non informative genes are introduced. Using the gene filtering process can improve the classification accuracy to some extent as long as proper filtering threshold is chosen. By experiment, the 10^{th} percentile as a threshold was preferred rather than the 15^{th} and 20^{th}. This step was implemented using the Bioinformatics toolbox-Matlab.

1.3 Generating MV

Missing values were randomly simulated in the original complete gene expression matrix (8086 genes) with certain overall missing ratios (MR) r (r = 1%, 2%, 3%, 4%). In more details, if a complete expression matrix contains m genes, n samples and c classes; we randomly pick m x n x r entries from it and erase them to form a dataset containing missing values. Although the MVs on the original microarray chip may occur not completely at random, we simulated the MVs at random as previous studies assumed [2], [8], [12]. Moreover, we emphasis what we are trying to focus on is the impact of using gene selection based classification in evaluating MV imputation methods. So as long as all the imputation methods are applied on datasets with the same MVs distribution, the results for downstream analyses are comparable. The distribution of genes with at least one MV at all values of MR is illustrated in table 1.

Table 1.Distribution of genes with different MR

MR%	Total number of MV	No. of genes with at least one MV*	Percentage of affected genes*
1%	4609	3514	43.45%
2%	9218	5540	68%
3%	13827	6638	82%
4%	18436	7290	90%

*The values are approximate average of two randomly simulated datasets

2. Validation Results

To summarize, by regarding the original complete dataset as a dataset of 0% MR, we have 4 missing ratios (1%, 2%, 3%, 4%), each associated with 2 simulated datasets (except for 0% MR), three missing value imputation methods (except for 0% MR), 2 gene selection methods, and 2 classifiers, using 2 cross validation schemes, which is repeated for 10 times. The validation results for the three imputation methods in terms of classification accuracy and NRMSE are illustrated in tables 2(a), 2(b), 2(c), and 2(d).

Table 2(a). Validation Results on 1%MR

Imputation Method	Validation*					NRMSE
	Cross Validation	Classification				
		FDA-SVM	FDA-Euclidean	ttest-SVM	ttest-Euclidean	
KNN	K-fold	0.82455	0.8246	0.85	0.83	0.7155
	Hold out	0.83	0.85	0.82	0.84	
SVD	K-fold	0.8421	0.82455	0.877	0.8246	0.6332
	Hold out	0.83	0.82	0.89	0.86	
Zero	k-fold	0.8158	0.82455	0.868	0.807	1
	Hold out	0.8	0.81	0.86	0.81	

*All values are approximate average of two randomly simulated datasets

Table 2(b). Validation Results on 2%MR

Imputation Method	Validation*					NRMSE
	Cross Validation	Classification				
		FDA-SVM	FDA-Euclidean	ttest-SVM	ttest-Euclidean	
KNN	K-fold	0.82455	0.8158	0.859	0.80705	0.688
	Hold out	0.81	0.75	0.84	0.72	
SVD	K-fold	0.859	0.8246	0.868	0.82455	0.62
	Hold out	0.84	0.85	0.86	0.91	
Zero	k-fold	0.833	0.807	0.85	0.82455	1
	Hold out	0.81	0.8	0.81	0.84	

*All values are approximate average of two randomly simulated datasets

Table 2(c). Validation Results on 3%MR

Imputation Method	Validation*					NRMSE
	Cross Validation	Classification				
		FDA-SVM	FDA-Euclidean	ttest-SVM	ttest-Euclidean	
KNN	K-fold	0.868	0.8158	0.872	0.83335	0.6992
	Hold out	0.85	0.84	0.92	0.78	
SVD	K-fold	0.84	0.8158	0.85	0.80705	0.628
	Hold out	0.86	0.88	0.87	0.77	
Zero	k-fold	0.815	0.807	0.85	0.83335	1
	Hold out	0.86	0.81	0.79	0.85	

*All values are approximate average of two randomly simulated datasets

Table 2(d). Validation Results on 4%MR

Imputation Method	Validation*					
	Cross Validation	Classification				NRMSE
		FDA-SVM	FDA-Euclidean	ttest-SVM	ttest-Euclidean	
KNN	K-fold	0.833	0.807	0.859	0.85965	0.7
	Hold out	0.82	0.85	0.84	0.8	
SVD	K-fold	0.807	0.8246	0.868	0.833	0.63
	Hold out	0.85	0.78	0.85	0.88	
Zero	k-fold	0.807	0.8158	0.859	0.8158	1
	Hold out	0.81	0.86	0.84	0.74	

*All values are approximate average of two randomly simulated datasets

3. Agreement with GO Terms

The selected genes identified by the specified gene selection methods, FDA and t-test statistics were checked with the SMD supplementary file associated with the data files. This supplement lists the genes annotations according to the gene ontology database (GO) for some hybridized genes in the beforehand data set. Each time an imputed data set was introduced to a classifier some of the selected genes, using either of the proposed feature selection methods, were found to be annotated in the GO. They are observed to code for proteins involved in cell adhesion/motility, lipid/fatty acid transport and metabolism, immune/defense response, and electron transport. A sample of some genes annotations for a common set of genes obtained when using the imputed datasets and the two feature selection methods are shown in table 3.

Table 3. Some Significant Genes Identified by FDA and ttest Feature Selection Methods

Unigene	Name	Symbol	GO Annotations
Hs.180878	lipoprotein lipase	LPL	heparin binding activity| lipoprotein lipase activity| lipid transporter activity| lipid transport| fatty acid metabolism| circulation| extracellular| lipid catabolism| hydrolase activity
Hs.20447	p21(CDKN1A)-activated kinase 4	PAK4	protein kinase activity| cell shape and cell size control| cell motility| signal transduction| Golgi apparatus
Hs.386793	glutathione peroxidase 3 (plasma)	GPX3	selenium binding activity| glutathione peroxidase activity| electron transporter activity| response to lipid hydroperoxide| soluble fraction| extracellular| peroxidase reaction| oxidoreductase activity
Hs.74034	caveolin 1, caveolae protein, 22kDa	CAV1	structural molecule activity| tumor suppressor| caveola| integral to plasma membrane
Hs.76392	aldehyde dehydrogenase 1 family, member A1	ALDH1A1	aldehyde dehydrogenase (NAD+) activity| androgen binding activity| electron transporter activity| aldehyde metabolism| cytosol| oxidoreductase activity
Hs.198241	amine oxidase, copper containing 3 (vascular adhesion protein 1)	AOC3	amine oxidase (copper-containing) activity| electron transporter activity| amine metabolism| cell adhesion| inflammatory response| integral to membrane| plasma membrane

IV. DISCUSSION

In this work we adopted the scheme of generating MV in a uniform distribution completely at random. Using a complete matrix that is extracted from the original MV contained gene expression matrix. Then, entries of the complete matrix are randomly removed to generate the artificial MVs as detailed in former sections. Finally, MV imputation is applied. In a former study [11] natural MV not artificially generated ones were employed. Then we adopted another scheme in determining the amounts of MV in agreement with [8]. Using the MV rate threshold (MVthld) throughout the study rather than using the true MV rate. Where, for a given MVthld (MVthld = 5n%, where n = 0, 1, 2, 3, 4), the genes with MV rate less than MVthld were retained to design the classifiers. As a result, the true MV rate of the remaining genes was not equal to MVthld and, in fact, could have been much less than MVthld. Hence, the parameter MVthld might not be a good indicator. Moreover, plotting the classification accuracies against a number of values for MVthld, as MVthld increased, the number of genes retained to design the classifier became larger and larger, so that the increase or decrease in the classification accuracy may be largely due to the additional included genes (especially if the genes are biomarkers) and may only weakly depend on MVthld.

By exploring the NRMSE results shown in tables 2(a), 2(b), 2(c), and 2(d) we can observe very slight deterioration of the NRMSE for all imputation methods by the increase in the MR. Moreover when comparing the three imputation methods at each MR one can observe that best NRMSE results are arranged in descending order for the three imputation methods as follows: SVDimpute, KNNimpute, and Zeroimpute. But one important finding is observed from the resulting values of the NRMSE. The importance of normalizing the data prior to imputation so that the data is of confidence to represent the actual spots intensities and thus the NRMSE becomes a reliable measure for evaluating the imputation methods.

On the other hand evaluating the three proposed imputation methods in terms of classification accuracies shows very slight variations when using the three imputation methods even at different MR. This consistency in the results of the classification accuracies are shown in tables 2(a), 2(b), 2(c), and 2(d).

Observing the retained classification accuracies for all imputed datasets at different MR, it is clear that they are very close to the classification accuracy of the complete dataset (0% MR) and in some experiments the accuracy for imputed datasets surpass that for the complete one. This implies that the imputed data are more accurate than the observed gene expressions that are considered as non missing data and they might be just a noisy version of the true signals. The overall classification accuracy of the complete dataset with FDA-SVM equals to 0.8421 using k-fold and 0.8 using holdout. For FDA-Euclidean it retains 0.8 using k-fold and 0.86 using holdout. For t-test-SVM with k-fold and holdout cross validations the classification accuracies equal 0.8596 and 0.88 respectively. And For t-test-Euclidean the classification accuracy equals 0.8 using k-fold and 0.88 with holdout cross validation.

The selected genes identified by both specified gene selection methods, FDA and t-test through all experiments were compared each time to the SMD supplement listing the gene ontology annotations (GO). In correspondence to [10] some of the selected genes retained during these feature selection processes are found to code for proteins involved in cell adhesion/motility, lipid/fatty acid transport and metabolism, immune/defense response, and electron transport as illustrated in table 3.

V. CONCLUSIONS

This work emphasizes the reliability of employing downstream data analysis in evaluating MV imputation methods as well as employing NRMSE. Data preprocessing was conducted in this work prior to imputation to enhance the data quality, the issue that increased the reliability of using NRMSE in this consensus.

As the gene expression values of the left out test samples may influence the imputation result, it might be proper to consider as a future work performing the very time consuming MV imputation for samples in each of the cycles of cross validation the same way the gene selection was implemented.

REFERENCES

[1] V. F. Ghoneim, N. H. Solouma, and Y. M. Kadah, "The Influence of Pre-processing and Gene Rank Aggregation on Microarray Data Analysis", in Proc. 27th National Radio Science Conference, 2010.

[2] 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, 17, 520-525, 2001.

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

[4] S. Friedland, A. Niknejad and L. Chihara, "A Simultaneous Reconstruction of Missing Data in DNA Microarrays", Linear Algebra Appl., to appear, Institute for Mathematics and its Applications, Preprint Series, No. 1948.
[5] 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, 19, 2088–2096, 2003.
[6] 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, 9:202, 2008.
[7] 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 no. 23, pages 4272–4279, 2005.
[8] 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 no. 23, pages 2883–2889, 2006.
[9] 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, 5:114, 2004.
[10] H. Zhao, A. Langerod, Y. Ji, K. W. Nowels, J. M. Nesland, R. Tibshirani, I. K. Bukholm, R. Karesen, D. Botstein, A. Borresen, and S. S. Jeffry, "Different Gene Expression Patterns in Invasive Lobular and Ductal Carcinomas of the Breast", Molecular Biology of the Cell, Vol. 15, 2523–2536, June 2004.
[11] V. F. Ghoneim, N. H. Solouma, and Y. M. Kadah, "Evaluation of missing values imputation methods in cDNA microarrays base on classification accuracy", in press.
[12] Y. Shi, Gene Expression Microarray Missing Value Imputation and Its Effects in Downstream Data Analyses, University of Alberta, 2007.