A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data

Kohbalan Moorthy, Mohd Saberi Mohamad* and Safaai Deris

Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, University Technology Malaysia, 81310, Skudai, Johor, Malaysia

Abstract: Many bioinformatics analytical tools, especially for cancer classification and prediction, require complete sets of data matrix. Having missing values in gene expression studies significantly influences the interpretation of final data. However, to most analysts' dismay, this has become a common problem and thus, relevant missing value imputation algorithms have to be developed and/or refined to address this matter. This paper intends to present a review of preferred and available missing value imputation methods for the analysis and imputation of missing values in gene expression data. Focus is placed on the abilities of algorithms in performing local or global data correlation to estimate the missing values. Approaches of the algorithms mentioned have been categorized into global approach, local approach, hybrid approach, and knowledge assisted approach. The methods presented are accompanied with suitable performance evaluation. The aim of this review is to highlight possible improvements on existing research techniques, rather than recommending new algorithms with the same functional aim.

Keywords: Gene expression analysis, gene expression data, information recovery, microarray data, missing value estimation, missing value imputation.

INTRODUCTION

The rapid development of software tools for analysis and interpretation of vast amount of data has benefited many research endeavors. In regard to this, the microarray technology is one of the essential tools to monitor wide expression levels of genes in a given organism. This technology is a significant advancement in the genetic field as the small size chip is able to cater large amount of genes for gene expression analysis.

Microarray technology allows expansion of the sample's information to generate detailed expressions of the data for gene regulation and identification [1]. It has been used in studies related to cancer classification, identification of relevant genes for diagnosis or therapy, and investigation of drug effects on cancer prognosis [2]. Its advantages have been proven in areas such as virology, immunology, and microbiology [1]. Today, obtaining samples from different type of diseases or diagnostic classes has been made possible through the advancement of this microarray technology [3].

Nevertheless, microarray data analysis is still a challenge in the bioinformatics field as there are still thousands or more uncharacterized variables which require careful mining and interpretation [4]. Gene expression analyses often suffer from the problem of missing values, which is a common problem in statistical analysis [5]. Missing values rate of less than 1% is considered inconsequential, 1-5% is controllable, 5-15% requires refined methods to handle the imputation, and that more than 15% strictly influences the prediction or interpretation.

There are many reasons that cause such undesirable situation, and these include hybridization failures, low resolution, artifacts on the microarray itself, image noise, corruption, and problems related to the spotting process [6]. In a number of studies, it has been shown that missing values in the data can severely affect the interpretation and hinder downstream analysis such as supervised classification, unsupervised clustering of genes, differentially expressed genes detection, and construction of gene regulatory networks [7]. Besides that, it has been found that missing values have negative effect on algorithms such as singular value decomposition (SVD), support vector machines (SVM), and also principal component analysis (PCA). The reason is that these algorithms can become non functional when missing values in data are encountered [8].

With wrong missing value imputation, possibility of losing the informative genes in the process of variable selection can be significant. If this problem can be identified and solved through latest improvement in the algorithms of missing values estimation, the identification of important genes can be secured. This is crucial in leading discovery of target genes of a particular class type [3].

Missing values estimation is therefore a low cost and effective alternative to repetition of the entire microarray experiments to recover all missing data points. Previous researchers have demonstrated that missing values imputation can significantly improve the overall prediction or data analysis when information about the data is incorporated into the imputation. According to [9], many sophisticated statistical models have been developed and applied in medical research area to estimate missing values from the non-missing part of the data set. Nonetheless, there

*Address correspondence to this author at the Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, University Technology Malaysia, 81310, Skudai, Johor, Malaysia;
Tel: +60-7-553-3153; Fax: +60-7-556-6164;
E-mails: saberi@utm.my, mohd.saberi@gmail.com

2212-392X/14 $58.00+.00

are still cases where inaccurate values have been estimated for the missing value spot in the gene expression data, and thus affected the identification of disease related genes. This disrupts the overall ranking of the significant genes, leading to loss of informative genes [10].

Many newly produced gene expression data from microarray experiments have less missing values or the values are mostly corrigible through repetitive microarray analysis. However, utilization of existing cancer patients' data that have been pre-processed with conventional missing value imputations, notably those based on k-nearest neighbor, is still widespread. The improvement of these data, used for the development of cancer prediction algorithms, is particularly important for the reproduction of hard or rare gene expression profiles that are used in cancer prediction studies such as brain cancer, colon cancer and etc. Early approaches in missing value imputation intend to remove the entire row that contains the missing values, then impute the missing values with row average values or median or even replace the missing values with zero [11]. However, filing missing values spots with zero or average values is not the best solution as it leads to biases because the correlation structure of the data is not counted.

MISSING VALUE IMPUTATION TECHNIQUES

In earlier imputation methods, the software used performs statistical analysis and biological information of the data is disregarded. Recent imputation methods, on the other hand, use the information mined from the microarray data to customize the analysis [7]. The methods for missing values imputation can be generally divided into three types, which are case or pairwise deletion, parameter estimation, and also imputation techniques [12]. [5] categorized the methods into four, namely case deletion (CD), mean imputation (MI), median imputation (MDI), and *K* nearest neighbor imputation (KNNI).

CD works by removing all cases or instances with missing values in at least one feature. The missing values are replaced by calculating the mean based on all known values of the attributes using the MI technique. Median is used particularly in MDI to assure reliability since the mean is affected by the existence of outliers. In the KNNI method, the similarity of two instances is determined using a distance function based on similar instances.

Besides that, [7] categorized the missing values imputation techniques into two main categories - generic statistical methods and application specific modifications. Mean imputation, hot deck imputation, model based imputation, multiple imputation, and cold deck imputation are all considered as generic statistical methods. In application specific modifications, quality issues or experimental designs are taken into account to impute the missing gene expression data.

In this review, the approaches of existing algorithms are categorized according to the type of information, which are global approach, local approach, hybrid approach, and knowledge assisted approach. The algorithms are grouped and as listed in Table **1**.

Table 1. List of Missing Value Imputation Algorithms Based on Category

Algorithms	Year	Category	References
SVDimpute	2001	Global	[11]
BPCA	2003	Global	[13]
MICE-CART	2010	Local	[14]
KNNimpute	2001	Local	[11]
GMCimpute	2004	Local	[15]
LLSimpute	2005	Local	[16]
SLLSimpute	2008	Local	[17]
CMVE	2005	Local	[18]
AMVI	2008	Local	[19]
ABBA	2010	Local	[20]
LinCmb	2005	Hybrid	[21]
POCSimpute	2006	Knowledge	[22]
GOimpute	2006	Knowledge	[23]
HAIimpute	2008	Knowledge	[24]

GLOBAL APPROACH

Using this approach, the algorithms perform missing values estimation based on global correlation information obtained from the entire data matrix. According to [8], if the algorithms assume that there exists a global covariance structure in all genes samples and that the genes exhibit dominant local similarity structures, then the imputation will become less accurate. Examples of such algorithms include the SVD imputation (SVDimpute) [11] and Bayesian principal component analysis (BPCA) [13].

LOCAL APPROACH

For local approach, the algorithms exploit only local similarity structure in the data sets to perform missing values imputation. The subsets of genes that show high correlation with the genes that contain the missing values are used to compute the missing values. The KNN imputation (KNNimpute) and local least square imputation (LLSimpute) are some of the earliest and well-known algorithms in this category [8].

The KNNimpute [11] uses pairwise information between the target gene with missing values and the *k* nearest reference genes to impute the missing values. Studies have shown that KNNimpute performs extremely well when strong correlation exists between genes in the data.

LLSimpute [16] uses a multiple regression model to impute missing values. This technique has been proven to be slightly more competitive than KNNimpute and is more complex than BPCA.

Sequential LLSimpute (SLLSimpute) [17] is an extension of LLSimpute algorithm which performs

imputation sequentially by starting from the gene with least missing rates. The imputed genes are then reused for imputation of other genes. It has been proven that SLLSimpute performs better than LLSimpute because the genes with missing values are reusable in this algorithm.

Another notable local approach is the MICE-CART, which is consisted of multiple imputations by chained equations (MICE) and classification and regression trees (CART). It is a nonparametric approach done by [14] to perform multiple imputations through chained equations using sequential regression trees as the conditional models. It is implemented to reduce the usage of parameter and is able to perform tuning while capturing complex relations of the data.

Gaussian mixture clustering imputation or GMCimpute [15] is capable of using more global correlation information even though it is a local approach algorithm. In this algorithm, the data is clustered into S components of Gaussian mixtures using the EM algorithm, where S represents the estimates of missing values. A single S value is computed from each component and then averaged to obtain the final estimated missing values. GMCimpute uses the local correlation information in the data through the mixture of components.

Collateral missing value imputation (CMVE) technique utilizes the concept of multiple parallel estimations of missing values to improve the final estimation. CMVE [18] has been able to produce better accuracy in normalized RMS error (NRMSE) compared to BPCA, KNN and LSimpute on many datasets involving ovarian cancer and yeast sporulation time series data.

Ameliorative missing value imputation (AMVI) [19] is more advanced than CMVE in the sense that it uses Monte Carlo simulation to determine the optimal number of reference genes, K. The time series expression profiles used in AMVI have been reported to exhibit strong dependency between observations.

Adaptive Bicluster-Based Approach or ABBA is a missing value estimator for binary matrices. This algorithm has been made less complex for better understanding and usage, but the amount of parameters tweaking is also higher. [20] verified the performance of this algorithm and concluded that it is better than KNN, particularly when the rate of missing values are higher than normal.

HYBRID APPROACH

For heterogeneous data sets, local correlation between genes dominates and techniques such as KNNimpute or LLSimpute perform better compared to BPCA or SVDimpute. This shows that the correlation structure in the data affects the performance of the imputation techniques. In term of global correlation, approaches such as BPCA or SVDimpute are preferred.

Nevertheless, there are still some hybrid methods like LinCmb [21] that can capture both local and global correlation information in the data. Using this method, the missing values are estimated by convex combination of five different imputation methods, namely row average, KNNimpute, SVDimpute, BPCA, and GMCimpute. LinCmb generates fake missing entries at positions where the true values are known and uses the constituent methods to estimate the missing entries. This method is adaptive to the correlation structure of data matrix in the sense that when more missing entries are present, global methods will become the focus to determine the missing values.

KNOWLEDGE ASSISTED APPROACH

This approach integrates the domain knowledge or external information into the missing values imputation process. This approach is powerful as it significantly improves the imputation's accuracy using domain knowledge. Thus, it is better than data driven approach, especially for data sets with small number of samples which are noisy or have high missing rate.

An example of this approach is the Projection Onto Convex Set (POCS), which is a flexible set theoretic framework which exploits the biological occurrence of synchronization loss and correlation information between genes and arrays [22]. POCS executes local least square regression to capture gene-wise correlation, performs PCA imputation to capture array-wise correlation, and restricts the squared power of the expressions profiles to capture synchronization loss. Optimal solution can be achieved using POCS regardless of whether local or global correlation structure prevails in the data. This is due to the fact that the final solution is always dominated by the smallest yet most reliable constraint set to satisfy larger yet less reliable constraint sets.

According to [25], functionality of genes is likely to be expressed in a modular fashion with some showing higher degree of concerted reactions to certain stimuli. Gene ontology (GO) is a well-accepted standard for gene function categorization and has three independent ontologies that explain gene products in terms of associated biological processes (BP), cellular components, and molecular functions (MF) [23]. GO improves the imputation accuracy as proven in experimental results where the proportion of annotated genes was large at higher rates of missing values.

Histone Acetylation Information Aided Imputation (HAIimpute) combines histone acetylation information into KNNimpute and LLSimpute to improve the accuracy of missing value estimation [24]. The mean expression of genes from each of the clusters is used by HAIimpute to form the pattern expressions. The missing values are then obtained by fitting a linear regression model between the gene and pattern expressions.

The final estimates of the missing values are given by a convex combination of linear regression imputations and secondary imputation using both KNNimpute and LLSimpute. It has also been proven that HAIimpute has consistently improved the KNNimpute or LLSimpute, indicated by improved correlation between imputed genes and original complete genes.

PERFORMANCE EVALUATION

Evaluation of the algorithm imputation results is a crucial step to demonstrate reliability and accuracy. Validation methods can be broadly classified into two categories –

internal and external validation. In internal validation, the performance indices are computed between the imputed and known original values. This validation is also known for using information gained from the dataset itself. As for external validation, the operation is done through subsequent biological analysis to assess the imputation effects. External validation usually uses the knowledge gathered externally rather than internal data information. A list of comparative validation methods is listed in Table **2**.

Table 2. Performance Evaluation Method for Missing Value Imputation Algorithms

Validation Type	Algorithms Involved
Internal Validation	
NRMSE or Variants of it	[11,13,15-17,21,22,24]
Pearson Correlation	[24]
Preservation of differentially expressed genes	[21]
Preservation of prediction/classification problem	[19]
External Validation	
GO enrichment	[23]
Presence of biologically relevant genes	[19]

INTERNAL VALIDATION

Analyzing the missing value imputation algorithms are done by computing normalized root mean square error (NRMSE). If NRMSE is low, it means that the imputation algorithm is more accurate. The NRMSE is defined as:

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik} - \tilde{g}_{ik})^2}{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik})^2}} \quad (1)$$

where g_{ik} denotes the kth experiment for gene g_i, and g and $\tilde{g}$ denote the true value and imputed value respectively.

EXTERNAL VALIDATION

Contribution of external information such as pathway information and functional annotations helps to determine the validity of the imputation algorithms in external validation. Usually, biologists look up for the GO term that has been significantly enriched among the genes. This GO term is then applied to characterize the functional roles of the genes. Based on [26], the enrichment P-value for each GO term, t, and for each cluster is calculated using this formula:

$$p = \sum_{i=k}^{\min(b,T)} \frac{\binom{T}{i}\binom{B-T}{b-i}}{\binom{B}{b}} \quad (2)$$

where b is the number of genes in the cluster, K is the number of genes in the cluster represented with the GO term, t, B is the number of genes in the dataset, and the total number of genes represented with a GO term is T.

LIMITATIONS

Each algorithm has its own advantages and disadvantages, so does the datasets being used for each missing values technique. Several studies have shown that the performance of missing values imputation algorithms is significantly affected by factors such as correlation structure in the data, the missing data mechanism, the distribution of missing entries in the data, and the percentage of missing values in the data [2].

Selecting the right algorithm may significantly boost the accuracy of the imputation results since there is no single imputation algorithm that performs the best in every situation. Global methods such as SVDimpute and BPCA perform better on data sets with low entropy, whereas local methods such as LLSimpute and KNNimpute perform better with high entropy data sets.

Another limitation is related to assumptions made on the distribution of missing values. Missing values in microarray data sets are normally assumed to be missing at random in most studies on missing values imputation, even though this is not a realistic assumption as missing values tend to arise in a systematic manner in practice. This pattern can significantly affect the imputation's performance and thus need to be properly considered in data analysis or algorithm design [12]. It also depends on the experimental conditions across the columns as variations in conditions may result in a non-random distribution of missing values in the data matrix. This is because each column in a microarray data matrix comes from different experiments.

FUTURE WORKS

Even though there are many algorithms for missing values imputation, more reliable algorithms are needed to accommodate the special characteristics of individual data sets. According to [2], adaptive methods that can capture both global and local correlation information in data set are useful in many situations. Thus, combining multiple estimates based on the variance in each estimate is a possible strategy. As more experimental data from different kinds of domains become available, new imputation algorithms that can handle mixed domain data sets with missing continuous and categorical data are required.

CONCLUSION

In large-scale experiment analyses and studies, missing values have become a frequent problem in microarray gene expression and this affects the result or conclusion being made. High throughput gene expression profiling techniques such as cDNA microarray technology also suffer from the problem of missing values due to various experimental reasons. Since many analyses require complete data set, missing values imputation is an essential pre-processing step in microarray data analysis.

Instead of introducing new algorithms, variations based on existing imputation approaches are worthy contributions to systematic evaluation of existing algorithms. With so

many different algorithms, evaluation of suitable parameters and operating platform has to be performed for optimal execution of the algorithm. This is to avoid overwhelming development of new missing value imputation algorithms before the full functional capability of previous existing ones has been fully explored.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank Malaysian Ministry of Higher Education for supporting this research by an Exploratory Research Grant Scheme (Grant number: R.J130000.7807. 4L096) and a Fundamental Research Grant Scheme (Grant number: R.J130000.7807.4F190). This research is also funded by an e-science research grant (Grant number: 06-01-06-SF1029) from Malaysian Ministry of Science, Technology and Innovation.

REFERENCES

[1] Pham TD, Wells C, Crane DI. Analysis of microarray gene expression data. Curr Bioinform 2006; 1: 37-153.

[2] Liew WC, Law NF, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief Bioinform 2010; 12: 498-513.

[3] Asyali MH, Colak D, Demirkaya O, Inan MS. Gene expression profile classification: A review. Curr Bioinform 2006; 1: 55-73.

[4] Sethi P, Alagiriswamy S. Association rule based similarity measures for the clustering of gene expression data. The Open Med Inform J 2010; 4: 63-73.

[5] Acuña E, Rodriguez C. The Treatment of Missing Values and its Effect on Classifier Accuracy; Classification, Clustering, and Data Mining Applications, Springer Berlin Heidelberg 2004; pp. 639-647.

[6] Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 2002; 11: 108-136.

[7] Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief Bioinform 2010; 11: 253-264.

[8] Liew AW, Law NF, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief Bioinform 2011; 12: 498-513.

[9] Donders ART, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59: 1087-1091.

[10] Zhang W, Li J, Su R, Jianghong W. Microarray data analysis to find diagnostic approach and identify families of disease-altered genes based on rank-reverse of gene expression. Curr Bioinform 2009; 4: 242-248.

[11] Troyanskaya O, Cantor M, Sherlock G, *et al.* Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17: 520-525.

[12] Little RJ, Rubin DB. Statistical Analysis with Missing Data. Second Edition. John Wiley and Sons: New York 2002.

[13] Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 2003; 19: 2088-2096.

[14] Burgette LF, Reiter JP. Multiple imputation for missing data *via* sequential regression trees. Am J Epidemiol 2010; 172: 1070-76.

[15] Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. Bioinformatics 2004; 20: 917-923.

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

[17] Zhang X, Song X, Wang H, Zhang H. Sequential local least squares imputation estimating missing value of microarray data. Comput Biol Med 2008; 38: 1112-1120.

[18] Sehgal MSB, Gondal I, Dooley LS. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. Bioinformatics 2005; 21: 2417-2423.

[19] Sehgal MSB, Gondal I, Dooley LS, Coppel R. Ameliorative missing value imputation for robust biological knowledge inference. J Biomed Inform 2008; 41: 499-514.

[20] Colantonio A, Pietro RD, Ocello A, Verde NV. ABBA: adaptive bicluster-based approach to impute missing values in binary matrices. Proceedings of the 2010 ACM Symposium on Applied Computing; Sierre, Switzerland. 1774304: ACM; 2010. p. 1026-33.

[21] Jörnsten R, Wang HY, Welsh WJ, Ouyang M. DNA microarray data imputation and significance analysis of differential expression. Bioinformatics 2005; 21: 4155-4161.

[22] Gan X, Liew AW, Yan H. Microarray missing data imputation based on a set theoretic framework and biological knowledge. Nucl Acids Res 2006; 34: 1608-1619.

[23] Tuikkala J, Elo L, Nevalainen OS, Aittokallio T. Improving missing value estimation in microarray data with gene ontology. Bioinformatics 2006; 22: 566-572.

[24] Xiang Q, Dai X, Deng Y, He C, Wang J, Feng J, Dai Z. Missing value imputation for microarray gene expression data using histone acetylation information. BMC Bioinformatics 2008; 9: 252-269.

[25] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999; 402: C47-C52.

[26] Tuikkala J, Elo LL, Nevalainen OS, Aittokallio T. Missing value imputation improves clustering and interpretation of gene expression microarray data. BMC Bioinformatics 2008; 9: 202-215.

Received: July 19, 2012 Revised: October 31, 2012 Accepted: November 10, 2012