World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:9, No:1, 2015

Imputation Technique for Feature Selection in Microarray Data Set

Younies Mahmoud, Mai Mabrouk, Elsayed Sallam

Abstract—Analyzing DNA microarray data sets is a great challenge, which faces the bioinformaticians due to the complication of using statistical and machine learning techniques. The challenge will be doubled if the microarray data sets contain missing data, which happens regularly because these techniques cannot deal with missing data. One of the most important data analysis process on the microarray data set is feature selection. This process finds the most important genes that affect certain disease. In this paper, we introduce a technique for imputing the missing data in microarray data sets while performing feature selection.

Keywords—DNA microarray, feature selection, missing data, bioinformatics.

I. Introduction

DNA microarray technology is one of the most important tools in functional genomics research. DNA microarrays measure the expression levels of thousands of genes simultaneously. This technology is involved in variety of biological researches, such as classifying and detecting cancer and identifying genes relevant to a certain disease or phenomena.

One of the most critical issues in microarray data analysis is feature selection [2]. Instead of using all available gene expressions in the microarray, we can choose the most valuable subset of gene expressions for the purpose of data analysis. There are many advantages for performing feature selection such as reducing the computational cost, improving the classification precision and identifying and monitoring specific diseases or phenomena [7].

The two well known techniques in feature selection are filters and wrappers [16]. The filter based techniques select a relevant subset without interacting with the classifier. These techniques evaluate each feature independently on the basis of different characteristics including distances, information, dependences and consistency. So, these techniques are very fast and have lower computational complexity compared to wrapper techniques, which interact with the classifiers or predictive models to select the relevant features by trying all the possible subsets from the features, and select the ones with the highest accuracy. These techniques must use searching algorithms to find the relevant subset because it is impossible to examine all the subsets [5].

Performing feature selection in microarray data sets faces the problem of missing data because most of these techniques cannot work with missing data. Many studies showed that microarray data sets can contain up to 10% of missing data and in some cases up to 90% of genes have one or more missing data [6]. The reasons of this problem are hybridization failures, microarray scratching, imperfect resolution or image noise and impurity. One simple solution for solving missing data problem is to repeat the high cost microarray experiment more than one time, which is very expensive. Another solution is to impute the missing data from the completed ones, which avoids the high cost of repeating the experiments. We can classify the imputation techniques into simple computational approach and intensive computational approach.

There are several simple imputation techniques to handle missing data problem. One of these techniques is to remove all the gene expressions with missing data from the entire microarray data set. This technique is very simple and has no errors, but it wastes on average about 50% of the gene expressions data set [6]. Another technique is replacing the missing data with a default value, generally zero, or with the average value of the expression row [1]. Unfortunately, these techniques distort the gene expression reality and reduce the trustiness of gene expression data.

To overcome these downsides, many intensive techniques have been developed such as K-nearest neighbor (KNN) imputation[19], sequential K-nearest neighbor (SKNN) imputation [11], iterative K-nearest neighbor (IKNN) imputation[3], singular value decomposition based imputation (SVD)[19], local least squares imputation (LLSimpute) [10] and sequential local least squares imputation (SLLSimpute) [21]. But, the estimation error generated from these techniques still affects the performance of statistical and machine learning techniques including predicting or discovering classes and identifying genes [18]. Therefore, more reduction of the estimation error will be efficient for the microarray data analysis. So there are significant motivations for developing new techniques, which minimizes the estimation error.

In this paper, we present a technique that deals with missing data while performing feature/gene selection process. Also, we demonstrated the proposed techniques by hands-on experiments on a real microarray data set under variety of conditions. The rest of the paper is organized as follows: Section II presents a group of imputation techniques for missing data. Section III introduces the proposed imputation technique. Section IV shows the results of our experiment. Finally, Section V concludes the paper and gives the future work.

Y. Mahmoud and E. Sallam are with the Department of Computer & Control Engineering, Faculty of Engineering,Tanta University, Egypt, , 31111 Egypt (e-mail: younies.mahmoud@f-eng.tanta.edu.eg).

M. Mabrouk are with Biomedical Engineering Department, Misr University for Science & Technology.

II. Missing Value Iputation for Microarray Data

The challenge of missing data in microarray data is to find estimated values to replace these missing data with the accurate estimated values. These techniques are significant for analyzing microarray data because most of the data analysis techniques cannot work in the presence of missing data. One possible solution is to repeat the high cost microarray experiment more than one time to ensure the absence of the missing data, which is very costly in terms of time and experiment cost.

Henceforth, the researchers developed different techniques to work with the challenge of missing data, which can be classified into simple computational approach and intensive computational approach, which are briefly discussed in the following subsections.

A. Simple Computational Approach

The techniques under this class uses primitive method to the challenge of missing data such as:

1) Gene Expressions: This technique depends on removing all of the gene expressions with missing data even if it has only one missing value. This technique wastes, approximately, fifty percent of the microarray data because the average missed data in the microarray data sets is fifty percent [12]. So, this technique is the worst feasible one.

2) Replacing With Default Value: This technique replaces all the missing data with a single value chosen randomly (commonly zeros). Although this technique does not waste any gene expression, it gives fake estimated values in the gene expression, which makes the result of the data analysis untrustable.

3) Replacing With Average: In this technique, the missed data for each feature will be replaced by a single value, which is the average of the data in this feature. This technique reduces the feature priority for the predictive models because it decreases the mean distance between the feature classes and increases the standard deviation for each class.

B. The Intensive Computational Approach

There are around seventeen intensive computational techniques [4] used for imputing missing data in microarray data sets. All of these techniques are trying to find the coherence between gene expressions without missing data to impute the missing data in other gene expressions. Brief discussion on some of the these techniques is presented in the following.

1) k-nearest neighbors (k-NN) and weighted k-NN impute techniques: KNN and weighted KNN impute is looking for k nearest gene expressions with most similar to the target gene expression (with missing data). Then, the missing data is estimated by the average or weighted average of related data for these k closest gene expressions [19],[9].

2) Bayesian principal component analysis: This technique is very efficient in dealing with missing data [15]. So, Zhou et al. [6] have used the Bayesian gene selection for estimating missing data with linear and non-linear regression. But, the k-NN technique is still the most popular one [15].

III. Proposed Imputation Technique

Most of the imputation techniques are developed to impute all the missing data in the microarray data sets, but in feature selection; only a small subset of the entire gene expressions is needed. So, the proposed computational technique is developed to impute the missing data only in the significant genes. To implement this idea, the technique is divided into several phases. In the following subsections, we will present the phases of the technique and the algorithm used for implementing this technique.

A. Phases of The Proposed Techniques

The proposed technique consists of five phases. Each phase is responsible for performing a specific operation in which the result of each operation leads to the next one as following.

1) Dividing The Gene Expressions into Groups: In this phase, the entire gene expressions are divided into groups. Each group contains the gene expressions with the same number of missing data. This result in a number of groups each one is different from the other in its number of missing data for their gene expressions.

2) Replacing With Average: This phase implements the average computation technique on each group because the average values reduce the gene expression rank and since each group has the same amount of missing data in each gene expressions, so the reducing rank for gene expression in a specific group will be equal.

3) Performing Feature Selection: This phase applies feature selection for the target disease or phenomena on each group independently to select the high ranked subsets of the gene expressions in each group. The resulted subsets from all the groups are logically connected with each other, because all of them are related to a specific disease or phenomena.

4) Predicting The Missing Data: As the selected subsets are interconnected, a predictive model can be built based on the subset of the completed gene expressions to estimate the missing data in the other selected gene expressions with missed data.

5) Selecting The Final Subset: After predicting the missing data, all the subsets are grouped into a one subset to select the final subset of the gene expressions.

B. Proposed Algorithm

The algorithm implements the proposed technique with considering that the predictive model is built based on the highest possible gene expressions. So, the algorithm performs the imputation in iterations. Each iteration implements all the phases of the proposed technique. As shown in algorithm 1, the input is an array of the $n+1$ subsets, where n is the maximum number allowed for missing data in a gene expression. Each subset contains gene expressions with the same amount of missing data in each one of them. The subsets in the array is sorted in an ascending order corresponding to the maximum amount of the missing data allowed for the gene expression in each subset, where the *zero*_based subset contains the gene expressions with *zero* missing data, the

Data: we have array *missed_sets[n+1]* that has n + 1 sets of features with 0-missing data , 1-missing data and so on
Result: producing the *final_set* from the *missed_sets* array
predictor_set = feature_selection(missed_sets[0])
i = 1 , final_set = [] ;
final_et = predictor_set;
while $i <= n$ **do**
  missed_set = feature_selection(missed_set[i]);
  foreach *feature in the missed_set* **do**
    missed_feature = get_feature(feature.id());
    completed_missed_features.add(predict_value(predictor_set, missed_feature));
  end
  final_set = feature_selection(predictor_set + completed_missed_set)
end

Algorithm 1: The proposed algorithm

*one*_based subset contains gene expressions with *one* missing data and so on.

The role of the *predictor_set* is to contain the subset of the gene expressions, which are responsible for building the predictive model to estimate the missing data in the other missed subsets. The *predictor_set* in the first contains the result of performing feature selection on the *zero*_based subset which contains the completed gene expressions. The *final_set* contains the output of the feature selection process and imputing process, which in the beginning equal to the *predictor_set*. After that the algorithm iterate through the *missed_subsets* from *one_missed_subset* to *nth_missed_subset* to perform the iterations.

The Iterations of The Algorithm:

1) Perform the *average_imputing* on the *ith_missed_set*.
2) Select the high gene expressions by applying *feature_selection* to get the *missed_gene_expressions*, which will be imputed using the *predictor_set*.
3) For each feature in the *missed_gene_expressions*, predictive model will be built to estimate the missed data in this gene expression.
4) The *final_set* will be produced after applying *feature_selection* on the *missed_gene_expressions* after the imputing and *predictor_set*.
5) The *predictor_set* will be equal to the *final_set*
6) Increase *i* by one to go to the next *missed_set*
7) Go to step no. *1*

After these iterations the output result, which is the *final_set*, contains final selected gene expressions with their estimated data.

IV. Experimental Results

To conduct the experiment, the pre-experiment phase is essential for choosing the feature selection technique, its classifier and the classification test model, besides choosing the predictive model for estimating the missing data. We compared

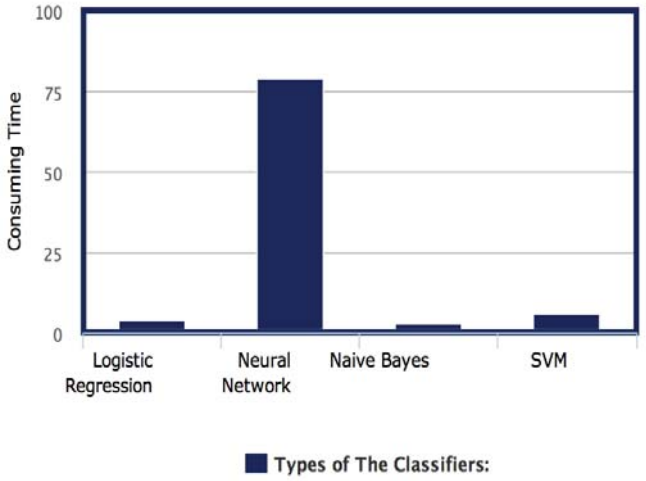

Fig. 1 Time consuming by each classifier

the proposed imputation technique with k-NN imputation technique because it is still widely used and implemented in the last edition of the bioinformatics toolbox in MATLAB [17].

Both of the pre-examine phase and the experiment phase are conducted on a real microarray data with gene expressions profiling of hepatocellular carcinoma (HCC) consists of 105 samples with 34 of them are HCC specimens (17 hepatitis B virus [HBV]-related and 17 hepatitis C virus [HCV]-related) and 71 non-tumor liver specimens (36 chronic hepatitis B [CH-B] and 35 chronic hepatitis C [CH-C])[20].

The concern in the experiment is with classifying the data based on the hepatocellular carcinoma (HCC).

A. Pre-experiment Phase

1) For performing feature selection, we chose the wrapper method because of its accuracy, although it is computationally intensive [5].
2) For choosing the appropriate classifier to preform the feature selection process, we used Weka, a reliable open source tool in machine learning [8], to conduct an experiment on microarray data with several classifiers. The microarray data is imputed using average imputation technique. This experiment compared each feature selection with a classifier upon the amount of time for feature selection and the accuracy using the *2_fold cross validation* [13]. As shown in Fig. 1 and in Fig. 2, the logistic regression consumed low amount of time for performing feature selection and gave the highest classification accuracy. So, we will use the logistic regression in our experiment.
3) For predicting the data, First Order Linear regression is appropriate because it is simple in the implementation, gives a reasonable accuracy and it is the counterpart of the logistic regression [14].
4) For choosing the maximum number allowed for missing data in a gene expression *n*, we found that the number of gene expressions in a specific subset decreases by

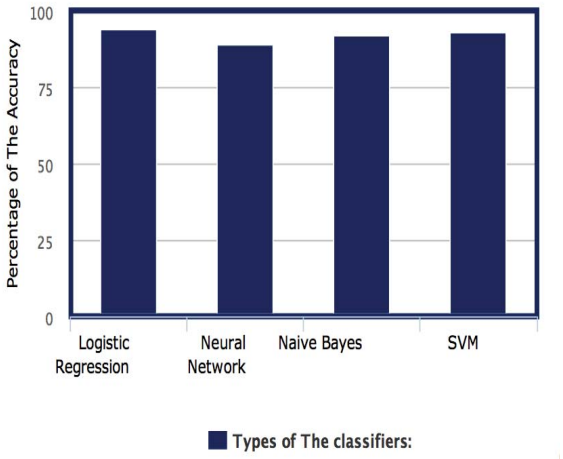

Fig. 2 Accuracy for Each Classifier

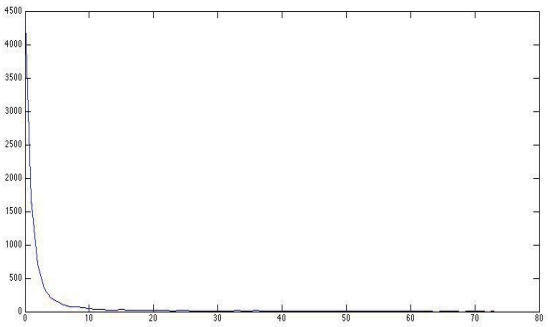

Fig. 3 The distribution for the amount of gene expressions in each level of missing data from zero to seventy five

the number of the missed data in the group gene expressions. As shown in Fig. 3 and Fig. 4, we found that the remaining data after the *5_missed_set* is too small compared to total gene expressions. So, we chose the *n* equal to *5* which wastes only 10% from the entire microarray data.

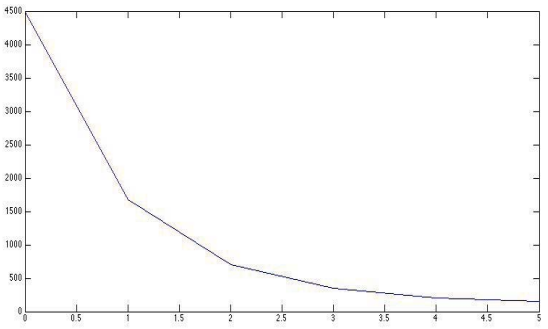

Fig. 4 The distribution for the amount of gene expressions in each level of missing data from zero to five

TABLE I: RESULTS OF THE PROPOSED ALGORITHM IN FIVE ITERATIONS

Iter#	Number of features in the final_set	Added Features	Accuracy
0	8	8	93.3333%
1	8	0	93.3333%
2	8	1	95.2333%
3	8	0	95.2333%
4	8	0	95.2333%
5	8	0	95.2333%

B. Experiment Phase

This phase contains two parts. The first part is to perform the proposed imputation technique using the proposed algorithm with the considerations in the pre-experiment phase. The second part is comparing the results of the proposed imputation technique with the k-NN imputation technique with time and accuracy perspective.

1) Performing The Proposed Imputation Technique: The proposed imputation technique performed in five iterations because we choose the *n* equal to *5*. As it shown in the table I, in the initial, the *final_set* contains eight gene expressions and the accuracy is *93.333%*. After the first iteration, the number of the gene expressions in the *final_set* is still the same because the new added feature is equal to *zero*, so, the accuracy is still the same. But, After the second iteration, the accuracy increased to *95.2333%* because there is a new gene expression added to the *fina_set* and the number of gene expressions in the *final_set* is still the same because Also there is a gene expression removed from the old *final_set*. In the remaining iterations, there is no added gene expressions, so, the number of gene expressions in *final_set* is remain the same and the accuracy too.

2) Comparing the result with k-NN impute: After performing the k-NN impute in the microarray data, we make comparison between it and the proposed imputation technique in the perspective of the time it takes for the imputation, the classification accuracy and the number of gene expressions in the imputation process. As it shown in table _refcompared_results, The effect of the reduction of the imputed gene expression in the proposed techniques is so obvious. The time for imputation is so small compared to the k-NN imputation technique. Beside the accuracy of the classification is greater than compared accuracy for the k-NN.

V. Conclusion & Future Work

This paper proposed imputation technique for microarray fata to perform feature selection. It is only used for feature selection and depends on reducing the imputed feature and

TABLE II: COMPARISON BETWEEN THE PROPOSED ALGORITM AND THE KNN-IMPUTATION TECHNIQUE

Iter#	Time consumption	Accuracy	Imputed features
Proposed Technique	1 second	95.2333 %	38
KNN Technique	45 seconds	94.3333%	3168

using predictive model for estimating the missing data rather than using imputation algorithms. The results demonstrate the efficiency of the proposed technique in both the consuming time and the resulted accuracy. In the future, finding a new way for reducing the imputed feature and choosing another predictive model is a great challenge to improve the imputation of microarray data.

REFERENCES

[1] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] V Bolón-Canedo, N Sánchez-Maroño, A Alonso-Betanzos, JM Benítez, and F Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014.

[3] Lígia P Brás and José C Menezes. Improving cluster-based missing value estimation of dna microarray data. *Biomolecular engineering*, 24(2):273–282, 2007.

[4] Magalie Celton, Alain Malpertuy, Gaëlle Lelandais, and Alexandre G De Brevern. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC genomics*, 11(1):15, 2010.

[5] Kyriacos Chrysostomou, M Lee, SY Chen, and X Liu. Wrapper feature selection., 2009.

[6] Alexandre G De Brevern, Serge Hazout, and Alain Malpertuy. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC bioinformatics*, 5(1):114, 2004.

[7] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[8] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[9] Rebecka Jörnsten, Hui-Yu Wang, William J Welsh, and Ming Ouyang. Dna microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22):4155–4161, 2005.

[10] Hyunsoo Kim, Gene H Golub, and Haesun Park. Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005.

[11] Ki-Yeol Kim, Byoung-Jin Kim, and Gwan-Su Yi. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 5(1):160, 2004.

[12] Alan Wee-Chung Liew, Ngai-Fong Law, and Hong Yan. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics*, 12(5):498–513, 2011.

[13] Rosa J Meijer and Jelle J Goeman. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155, 2013.

[14] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[15] Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[16] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[17] Henning Schmidt and Mats Jirstrand. Systems biology toolbox for matlab: a computational platform for research in systems biology. *Bioinformatics*, 22(4):514–515, 2006.

[18] Muhammad Shoaib B Sehgal, Iqbal Gondal, and Laurence Dooley. Statistical neural networks and support vector machine for the classification of genetic mutations in ovarian cancer. In *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on*, pages 140–146. IEEE, 2004.

[19] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[20] Teruyuki Ueda, Masao Honda, Katsuhisa Horimoto, Sachiyo Aburatani, Shigeru Saito, Taro Yamashita, Yoshio Sakai, Mikiko Nakamura, Hajime Takatori, Hajime Sunagozaka, et al. Gene expression profiling of hepatitis b-and hepatitis c-related hepatocellular carcinoma using graphical gaussian modeling. *Genomics*, 101(4):238–248, 2013.

[21] Xiaobai Zhang, Xiaofeng Song, Huinan Wang, and Huanping Zhang. Sequential local least squares imputation estimating missing value of microarray data. *Computers in biology and medicine*, 38(10):1112–1120, 2008.