Missing Value Estimation of Microarray Data Using Similarity Measurement

Soumen Kumar Pati[1] and Asit Kumar Das[2]

[1] Department of Computer Science/Information Technology,
St. Thomas' College of Engineering and Technology, 4, D.H. Road, Kolkata-23
[2] Department of Computer Science and Technology,
Bengal Engineering and Science University, Shibpur, Howrah-03
Soumen_pati@rediffmail.com, asitdas72@rediffmail.com

Abstract. DNA gene expression profiling plays an important role in a wide range of areas in biological science for handling cancer diseases. Data generated in microarray related experiments have many missing expression values which lose valuable information from the dataset. The proposed method first partitions the genes without missing values using clustering algorithm and then measures the similarity between a gene with missing values and the centroid of the clusters and finally, the missing values are estimated by the corresponding expression values of the centroid giving maximum similarity factor. The method explicitly depends on expression values to imputes missing values, completed the input dataset with low errors for data analysis and knowledge discovery. The method is compared with prominent approaches, such as zero-impute, row-average-impute and KNN-impute in terms of "Normalized Root Mean Square Error" to claim its novelty.

Keywords: DNA Microarray data, Gene expression value, Clustering algorithm, Similarity measurement, Missing value imputation.

1 Introduction

DNA microarray technology gives a global view of gene expression monitoring the mRNA levels of thousands of genes in particular cells or tissues. Microarray datasets [1] are usually in the form of large tables of expression levels of genes (rows) under different experimental samples (columns). The datasets frequently contain missing values due to insufficient resolution, spotting or scratches on the slide, image corruption, dust or hybridization failures and so on [2]. Unfortunately, most of algorithms for gene data analysis require a complete matrix as input. So the proper and more accurate prediction of Missing values remains an important preprocessing step to analyze microarray dataset. Several approaches [3-8] are proposed by the researchers to deal with missing values. The approach [3] repeats the original experiment to get microarray dataset without missing values, which is expensive and more time consuming. The approach [4] ignores genes containing missing values, that usually loses many useful information and may bias the results if the remaining genes

B.K. Panigrahi et al. (Eds.): SEMCCO 2012, LNCS 7677, pp. 602–610, 2012.

unable to represent the entire dataset. Some approaches [4, 5] estimate the missing values by a global constant such as zero (0), or by the average of the available sample values for that gene, which distort relationships among expression values for that gene. And others [7] consider the correlation structure among expression values for a gene. The estimating procedure consists of two steps: in the first step similar expression values related genes to the gene with missing value, are selected and in the second step the missing values are predicted using observed values of selected genes, for example the widely used weighted K-nearest neighbor (KNN) imputation, estimate the missing values using a weighted average of K most similar genes [6]. These methods have better performance than previous one, but the drawback is that their estimation ability depends on parameter K (number of K neighbor genes used to estimate missing value) for which no theoretical way exist to determine them appropriately and thus need to be specified by the user. Whereas, in [2, 8], cluster-based algorithms have been proposed to deal with missing values which don't need user to determine K parameters [7] but microarray dataset is very high dimensional and there exist large number of genes with large number of samples which may degrades the clustering performance. Also this method depends on number of clusters whose selection becomes very crucial. So this approach is also inefficient to deal with missing values.

In the article, a novel missing value estimation technique has been proposed on microarray dataset for imputing missing values that not only overcomes the constraints of the existing methods [2-8] but also gives significantly less Normalized Root Mean Square Error (NRMSE). The method of missing value estimation consists of the following steps:

i. The dataset is standardized to Z-score using Transitional State Discrimination method (TSD) [9] and the samples are characterized by N (here, N = 5) discrete sample values. As the samples are collected from both normal and cancerous patients, they are divided into two disjoint classes. For each gene, frequencies of sample values are computed in each class (i.e., normal and cancerous).
ii. Based on the frequencies of discrete sample values, the genes without missing value are partitioned into $3 \times N$ (here, 15) different groups, explained in the following section. N out of $3 \times N$ groups contains whole portion (i.e., normal and cancerous samples) of the genes while each N of remaining $2 \times N$ groups contains only one portion (i.e., either normal or cancerous samples) of the genes.
iii. Either a gene with missing values is associated to one of the N groups containing whole portion or each of its two portions (i.e., normal and cancerous) is associated to one of the $2 \times N$ respective groups containing only one portion.
iv. Now the determined group(s) is partitioned into optimal set of clusters using clustering algorithm [10] and similarity factors are measured between centroid of each partition and associated portion of the gene. The missing values of the associated portion of the gene are imputed by the respective values of the centroid with most similar partition. Thus, missing values of each gene are imputed by repeating step (iii) and step (iv).

The article is organized into four sections. Section 2 describes the proposed missing value estimation technique. The experimental results and performance of the proposed method for various benchmark gene expression datasets is evaluated in Section 3. Finally, conclusions are drawn in Section 4.

2 Missing Value Estimation Method

Microarray technology [1] is a very high throughput technology that evaluates the expression of immense number of genes simultaneously under different experimental conditions. These conditions may be a time series during a biological process or a collection of different tissue samples (e.g. normal versus cancerous samples). Usually data from microarray experiments contains missing values due to different reasons including dust or scratches on the slide, error in experiments, image corruption, insufficient resolution for which (5 – 50)% genes are affected. Therefore missing value estimation is an important data preprocessing step to impute proper expression values with less error.

2.1 Discretization of Gene Expression Values

Initially, experimental gene dataset (***U, C***) are discretized using Transitional State Discrimination method (*TSD*) [9], where *U*, the universe of discourse contains *n* genes and *C*, the condition attribute set contains *m* samples. In *TSD* [9], discretization factor f_{ij}, based on which the dataset is discretized, is computed for sample $C_j \in C$ of gene $g_i \in U$, using (1), for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

$$f_{ij} = \frac{M_i[C_j] - \mu_i}{\delta_i} \tag{1}$$

Where, μ_i and δ_i, the mean and standard deviation respectively of gene g_i and $M_i[C_j]$ is the value of sample C_j in gene g_i. Then mean (N_i) of negative values and mean (P_i) of positive values of each gene g_i are computed and discretized to one of *N* (here, *N* = 5) fuzzy linguistic terms using (2).

$$f_{ij} = \begin{cases} 'VL' & if\ f_{ij} \le N_i \\ 'L' & if\ N_i < f_{ij} < 0 \\ 'Z' & if\ f_{ij} = 0 \\ 'H' & if\ 0 < f_{ij} < P_i \\ 'VH' & if\ f_{ij} \ge P_i \end{cases} \tag{2}$$

After discretization, the dataset is divided into two sets, one set (*MISS*) contains genes with missing value and other set (*NOMISS*) contains genes without missing value.

2.2 Formation of Correlated Gene Subsets

Let, the samples of genes are collected from d_1 normal and d_2 cancerous patients; so each gene contains d_1 normal and d_2 cancerous samples. Let, each gene $g_i \in NOMISS$

is represented as $g_i = \{g^n_{i1}, g^n_{i2}, \ldots, g^n_{id_1}, g^c_{i1}, g^c_{i2}, \ldots, g^c_{id_2}\}$, where g^n_{ij} for $j = 1, 2, \ldots, d_1$ are normal samples and g^c_{ik} for $k = 1, 2, \ldots, d_2$ are cancerous samples. Frequencies of discrete expression values for samples $\{g^n_{i1}, g^n_{i2}, \ldots, g^n_{id_1}\}$ and $\{g^c_{i1}, g^c_{i2}, \ldots, g^c_{id_2}\}$ of gene g_i are computed as $\{f^{ni}_{VL}, f^{ni}_{L}, f^{ni}_{Z}, f^{ni}_{H}, f^{ni}_{VH}\}$ and $\{f^{ci}_{VL}, f^{ci}_{L}, f^{ci}_{Z}, f^{ci}_{H}, f^{ci}_{VH}\}$ respectively, where f^{ni}_{VL} is the frequency of expression value '*VL*' in normal samples of gene g_i, similar meaning of other terms. Let $f^{ni}_{max} = max\{f^{ni}_{VL}, f^{ni}_{L}, f^{ni}_{Z}, f^{ni}_{H}, f^{ni}_{VH}\}$ and $f^{ci}_{max} = max\{f^{ci}_{VL}, f^{ci}_{L}, f^{ci}_{Z}, f^{ci}_{H}, f^{ci}_{VH}\}$. The gene subsets are formed as follows:

If f^{ni}_{max} and f^{ci}_{max} are computed from
(i) Same discrete expression value, say '*VL*' then the gene $g_i = \{g^n_{i1}, g^n_{i2}, \ldots, g^n_{id_1}, g^c_{i1}, g^c_{i2}, \ldots, g^c_{id_2}\}$ is placed in subset $GENE\_WHOLE_{VL}$ (abbreviated as GW_{VL}, used synonymously in the paper), same situation for other discrete values. Thus, five subsets GW_{VL}, GW_L, GW_Z, GW_H and GW_{VH} are formed. Each of these five subsets contains genes of *NOMISS,* where maximum frequency of discrete value occurs for same discrete value in both normal and cancerous samples.
(ii) Different discrete expression value, say f^{ni}_{max} occurs for '*VL*' and f^{ci}_{max} occurs for '*VH*'. In this case, the normal part $\{g^n_{i1}, g^n_{i2}, \ldots, g^n_{id_1}\}$ of g_i is placed in subset $GENE\_NORMAL_{VL}$ (abbreviated as GN_{VL}), same situation for other discrete values. And cancerous part $\{g^c_{i1}, g^c_{i2}, \ldots, g^c_{id_2}\}$ of g_i is placed in subset $GENE\_CANCER_{VH}$ (abbreviated as GC_{VH}), same situation for other discrete values. Thus, GN_{VL}, GN_L, GN_Z, GN_H and GN_{VH} are formed, each of which contains normal samples of genes whose maximum frequency discrete value differs from that of cancerous samples. Similarly, gene subsets containing only cancerous samples are formed which are GC_{VL}, GC_L, GC_Z, GC_H and GC_{VH}.

Thus, genes without missing value (i.e., set *NOMISS*) are partitioned into fifteen subsets. These subsets are formed according to the gene expression values of the dataset and each subset contains similar type of genes.

2.3 Similarity Measurement

Fifteen gene subsets are formed from the set *NOMISS* of genes without missing values. Each set contains the genes of similar type. On the other hand, the set *MISS* contains genes with missing values which need to be estimated as data preprocessing step of knowledge discovery. Each gene $g_j \in MISS$ can also be thought of as $g_j = \{g^n_{j1}, g^n_{j2}, \ldots, g^n_{jd_1}, g^c_{j1}, g^c_{j2}, \ldots, g^c_{jd_2}\}$ with some g^n_{jk} and g^c_{jl} may be missed, for $k = 1, 2, \ldots, d_1$ and $l = 1, 2, \ldots, d_2$ which are estimated by the proposed method.

The same process is applied to compute the maximum frequency of discrete expression values in both normal and cancerous samples of gene $g_j \in MISS$. If maximum frequency occurs in both types of samples for same expression value, say *'VL'*, then g_j is associated with gene subset GW_{VL}. But if maximum frequency occurs for different expression values, say '*VL*' and '*VH*' for normal type and cancerous type respectively, then normal samples $\{g^n_{j1}, g^n_{j2}, \ldots, g^n_{jd_1}\}$ of g_j is associated with GN_{VL} and cancerous samples $\{g^c_{j1}, g^c_{j2}, \ldots, g^c_{jd_2}\}$ of gene g_j is associated with GC_{VH}. Thus

each gene $g_j \in MISS$ is either (*a*) associated with any one subset of {GW_{VL}, GW_L, GW_Z, GW_H, GW_{VH}} or (*b*) normal portion of it is associated with any one of {GN_{VL}, GN_L, GN_Z, GN_H, GN_{VH}} and cancerous portion of it is associated with any one of {GC_{VL}, GC_L, GC_Z, GC_H, GC_{VH}}. Similarity of gene gj in case of (*a*) is discussed below; whereas same logic is applied in case of (*b*), which is not described redundantly.

(*a*) Now, associated gene subset with real values is partitioned using clustering algorithm [10] which provides optimal set of K-clusters. Centroids of all K-clusters are computed and discretized using (2). Thus, K-centroids of (d_1+d_2)-tuples, one for each cluster is obtained. Let, the centroids of cluster T is $CENTRE_T = \{C^n_{t1}, C^n_{t2}, \ldots, C^n_{td_1}, C^c_{t1}, C^c_{t2}, \ldots, C^c_{td_2}\}$, for $t = 1, 2, \ldots, k$, where, C^n_{tj} is the mean (centroid) of j-th normal samples in cluster T, for $j = 1,2,\ldots,d_1$ and C^c_{tj} is the mean (centroid) of j-th cancerous samples of cluster T, for $j = 1, 2, \ldots, d_2$. Now the similarity S_{jT} of gene $g_j \in MISS$ with cluster T is the number of samples having discrete value equals to that of centroid of T, define by following function:

```
Function: Similarity (gene g_j, cluster T {
  /* gene g_j = {g^n_j1, g^n_j2, ..., g^n_jd1, g^c_j1, g^c_j2, ..., g^c_jd2} and centroid of
     Cluster T is CENTRE_T = {C^n_t1, C^n_t2, ...., C^n_td1, C^c_t1, C^c_t2, ...., C^c_td2} */
  S_jT =0; //similarity between gene g_j and cluster T
  For i = 1 to d_1
     If (g^n_ji = C^n_ti )
        S_jT = S_jT + 1;
  For i = 1 to d_2
     If (g^c_ji = C^c_ti )
        S_jT = S_jT +1;
  Return (S_jT);
}
```

Thus, similarity of g_j with all K clusters are obtained and if S_{jP} is maximum for $1 \le P \le K$ and the missing g^n_{jq} will be estimated by C^n_{pq}, $1 \le q \le d_1$ and missing g^c_{jr} will be estimated by C^c_{pr}, $1 \le r \le d_2$. Thus, all gene g_j with missing values are estimated. The overall algorithm of missing value is described below:

```
Algorithm. MISSING_VALUE_IMPUTATION (U, C)
Input:  U is the gene dataset containing n genes, C is
        the sample set containing m samples.
Output: Gene dataset with estimated missing values.

Step1: Discretized dataset U with N number of discrete
       values, using (1) and (2).
Step2: Create gene set MISS with missing values and
       NOMISS without missing values.
```

Step3: Find maximum frequency f_1 and f_2 of discrete values of a gene of *NOMISS* for normal and cancerous samples.
Step4: If (f_1 and f_2 occurs for same discrete value) then
Put whole gene into one of N gene subsets associated with respective discrete value.
Else, Put normal and cancerous part of gene separately into two subsets of 2×N gene subsets
Step5: Repeat Step3 and step4 for all genes of *NOMISS*.
Step6: Take a gene from *MISS* and select its associated set among 3×N gene subsets based on samples behavior.
Step7: Perform clustering algorithm [10] on selected gene subset and find optimal number of clusters.
Step8: Select cluster to which considering gene has maximum similarity.
Step9: Impute missing value of the sample of the gene by the corresponding value in centroid of the selected cluster.
Step10: Repeat Step6 to Step9 for all genes of *MISS*.
Step11: Stop.

3 Experimental Results and Performance Evaluation

Experimental studies presented here provide an evidence of effectiveness of the proposed missing values imputation method on experimental microarray dataset. Experiments were carried out on large number of different kinds of microarray data, few of which are summarized below:

(i) Leukemia dataset: Training dataset consists 27 ALL and 11 AML samples, over 7129 human genes. The raw data is available at http://www-genome.wi.mit.edu/cgi bin/cancer/datasets.cgi.
(ii) Diffuse Large B-cell Lymphoma (DLBCL) dataset: The dataset contains 58 DLBCL and 19 Follicular Lymphoma (FL) samples, over 7129 genes. Raw data are available at http://www-genome.wi.mit.edu/cgibin/cancer/datasets.cgi.
(iii) Lung Cancer dataset: Training dataset contains 16 samples labeled as "MPM" and 16 samples labeled as "ADCA" with around 12533 genes. The raw data are available at http://www. chestsurg.org/microarray.htm.
(iv) Prostate Cancer dataset: Training dataset consists 52 "relapse" and 50 "non-relapse" samples, over 12600 genes. The raw data are available at http://www-genome. wi. mit.edu/mpr/prostate.

The microarray gene expression dataset is divided into two subsets where one contains without missing value related genes and other contains randomly created

missing values related genes with randomly created missing positions where predicted values are imputed by the proposed method. The performance of the proposed method with compare to some traditional missing value estimation methods (i.e., Zero Imputation, Row Average and KNN) are measured by Normalize Root Mean Square Error (*NRMSE*). The *NRMSE* is computed for different methods using (3).

$$NRMSE = \frac{1}{std\_dev(X_{known})}\sqrt{\frac{\sum_{i=1}^{n}\left(X_{predict} - X_{known}\right)^2}{n}} \quad (3)$$

Where, X_{known} is the original gene expression value and $X_{predict}$ is the estimated value of the proposed algorithm, $std\_dev(X_{known})$ is the standard deviation of original expression values and n is the total number of missing values. The number n is computed randomly according to 5%, 10%, 15%, 20%, 25% and 30% of missing values and *NRMSE* are computed for all methods. The result shows that *NRMSE* produced by the proposed algorithm are significantly less than the other methods for different dataset, which confirms the potentiality and superiority of the proposed method. The KNN technique is applied for different values of *K* and taking the best results among them. The outstanding estimation ability of proposed missing value imputation method is important due to the use of correlation structure of gene expression values, novel clustering algorithm and similarity factor measurement. The other methods depends how far sample values of number of genes are closed with the missing value ignoring the characteristic of expression values of genes, which might be different. But the proposed method depends on the characteristics of gene expression values. The Fig. 1 to Fig. 4 shows the visual proof of several dataset by computing NRMSE for several methods.

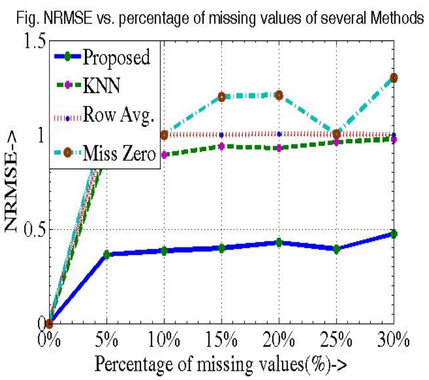

Fig. 1. Comparison of NRMSE value with different methods for Leukemia dataset

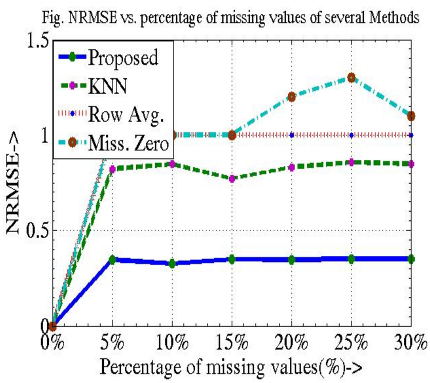

Fig. 2. Comparison of NRMSE value with different methods for DLBCL dataset

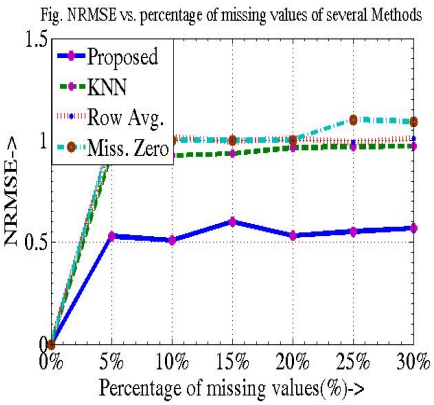

Fig. 3. Comparison of NRMSE value with different methods for Lung cancer dataset

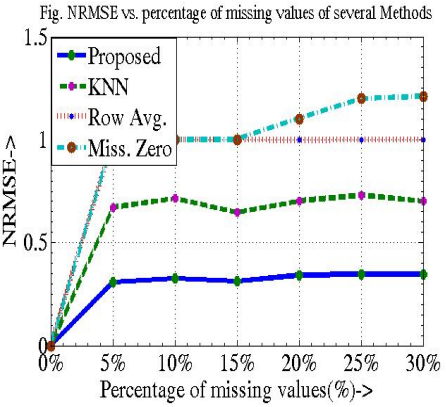

Fig. 4. Comparison of NRMSE value with different methods for Prostate cancer dataset

All the algorithms are implemented using Mat lab 7.8.1 version. Also all comparison figures are drawn using Mat lab 7.8.1 version. The comparisons are performed on PC (Intel(R) Core(TM) 2 Duo T5750 2.0 GHz, 2.0 GHz with 2.0 GB of Ram).

4 Discussions and Conclusion

Systematic Missing data can bring lots of difficulties in microarray data analysis simply because most existing methods were not designed for them and without imputing these values properly, the result will be erroneous. So this is most important preprocessing step to deal with missing values in the context of the integration of post-genomic experimental dataset. The existing statistical techniques incorporate with the context estimate missing values without measuring the correlation between normal and cancerous samples, which may give some valuable information about the nature of the gene. To this circumstance, the proposed method is conceptual and computational challenge totally depends on expression values and independent on number of genes. To measure the correlation between the normal and cancerous samples, the dataset is split into small subsets which help to estimate the missing values effectively. The performance of proposed method is analyzed on four common publicly available microarray dataset and compared the accuracy with Zero-impute, Row-average and KNN in terms of the NRMSE which shows the goodness of proposed method.

References

1. DeRisi, J., et al.: Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. 14(4), 457–460 (1996)

2. Luo, J., Yang, T., Wang, Y.: Missing Value Estimation for Microarray Data Based On Fuzzy C-means Clustering. In: Proceedings of the Eighth International Conference on High-Performance Computing in Asia-Pacific Region (2005)
3. Butte, A.J., Ye, J.: Determining Significant Fold Differences in Gene Expression Analysis. In: Pac. Symp. Biocomput., vol. 6, pp. 6–17 (2001)
4. Alizadeh, A.A., et al.: Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. Nature 403, 503–511 (2000)
5. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods 7, 144–177 (2002)
6. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525 (2001)
7. Huynen, M., Snel, B., Lathe, W., Bork, P.: Genome Res. 10, 1204–1210 (2000)
8. Zhang, S., Zhang, J., Zhu, X., Qin, Y., Zhang, C.: Missing Value Imputation Based on Data Clustering. Transactions on Computational Science (TCOS) 1, 128–138 (2008)
9. Velarde Cristina, C., Escudero, R., Zaliz, R.R.: Boolean Networks: A Study on Microarray Data Discretization. In: ESTYLF 2008, Cuencas Mineras, Mieres, Langreo, pp. 17–19 (2008)
10. Pati, S.K., Das, A.K.: Cluster Analysis of Microarray Data Based on Similarity Measurement. International Journal of Bioinformatics Research 3(2), 207–213 (2011) ISSN: 0975-3087