Imputing Missing Values in Microarray Data with Ontology Information

Andy C. Yang, Hui-Huang Hsu*, Ming-Da Lu
Department of Computer Science & Information Engineering,
Tamkang University, Taipei, Taiwan, R.O.C.
*andyyung0215@gmail.com, *huihuanghsu@gmail.com*

Abstract

Microarray technology is a big step in bioinformatics. Hidden information within the large amounts of data provides scientists with molecular functions or essential biological meanings to study and analyze. However, these data often contain a certain portion of entities that are missing. Several methods to estimate these missing values are developed, but most of them are with disadvantages. In this paper, we propose a novel approach to deal with these missing values based on a practical similarity measurement between gene pairs. Our approach takes gene expression values and gene ontology (GO) information for genes into consideration. We implement our approach on a real microarray dataset and compare its imputation accuracy with other methods. Experimental results show that our approach can estimate missing values in microarray data effectively.

Key words: Microarray, missing value, gene ontology

1. Introduction

Microarray experiments are widely used for biological analysis or disease diagnosis in the passing decade. It offers an opportunity for biologists to monitor thousands of genes or proteins at the same time on single microarray chip. Numerous gene expression data can be generated simultaneously via this high throughput biological technology. Gene expression values on each microarray chip represent the reaction of each gene after the hybridization effect across time [1,2]. Each gene expression value represents different reaction degrees resulted from experiments. In the quantitative data matrix, each gene expression value is in the format of logarithm. This kind of data provides a possible means for the inference of transcriptional regulatory relationships among the genes on the microarray gene chips.

However, microarray gene expression data often contain multiple missing values. For real microarray datasets, a certain portion of gene expression values that do not exist in the raw data are called missing values. These empty values need to be correctly and effectively imputed before any further analysis is performed. The reason why these missing values occur can be due to human operations, experimental inaccuracy, or unobvious reaction at that time slot of certain genes [3]. Figure 1 is an example of missing value problem in microarray data. G_{33} in Figure 1 stands for a missing value of gene 3 at the third time point.

	E1	E2	E3	…	…	E5
Gene1	-0.3	0.5	0.1	0.4	-0.6	0.1
Gene2	0.4	$G_{2,2}$	-0.4	$G_{2,4}$	-1.1	0.9
Gene3	-0.2	0.3	$G_{3,3}$	0.5	-0.7	0.2
…	0.6	0.5	0.1	$G_{4,4}$	$G_{4,N-1}$	$G_{4,N}$
…	-0.5	$G_{N-1,2}$	0.3	0.4	-0.6	0.1
GeneN	0.7	0.1	$G_{N,3}$	-0.3	0.2	0.5

Figure 1.Missing values in microarray data

To solve this problem, several methods developed from different viewpoints are introduced in these years. Among all published literatures, existing methods for microarray missing value imputation mainly utilize k-nearest neighbor (KNN) or KNN-like approaches to estimate the missing values [4]. It is shown that KNN is a widely-used method to estimate an unknown target object with known information. When applying KNN to impute missing values in microarray time series-data, we have to choose a number of k similar genes without missing entries at the same time slot (experiment) as the target missing value. Besides, we still need to estimate how similar the two genes of interest are to identify whether the two genes have regulatory relations. For this purpose, most of the similarity

measurements take the statistical or mathematical correlations among genes into consideration. These principles may be unsuitable to the microarray time-series data because of the existence of outliers [5]. Outliers influence much on the correlation coefficient measurements, especially when there are two or more outliers occurring in the time-series data set. Also, when identifying similarity of two genes in microarray time-series data, comparing local similarity is usually more important than comparing all time slot points. This is because even genes with known regulations may have reaction delay or offsets among time axis in microarray experiment results [6].

Moreover, external information such as gene ontology for genes themselves is utilized as a hint to improve the imputation accuracy [7-9]. This kind of external information can provide extra messages for genes to improve the accuracy of imputation. Combining this external information requires heuristic works as well. Besides KNN or KNN-based imputation methods, there are also several other works proposed from different aspects. Oba et al. propose an estimation method for missing values based on Bayesian principal component analysis (BPCA) [10]. BPCA is shown to outperform others. However, it is not easy to decide the number of principal axes while applying BPCA for missing value imputation. Moreover, BPCA is not practical in the case when many missing values occur in one time slot.

In this paper, we propose a novel approach to impute missing values resulted from many reasons in microarray data. We first define a similarity measurement for gene pairs based on the combination of difference of gene expression values and gene ontology semantic closeness. This similarity measurement is then applied to work with the KNN method as the distance of each gene pair to estimate missing values in real microarray datasets.

The remaining of this paper is organized as follows. In Section 2, details of our approach for missing value imputation are given by starting from the definition of our similarity measurement for gene pairs. Section 3 described the involved datasets and the estimation of imputation accuracy. Experimental results are then presented and discussed in Section 4. The concluding remarks are drawn in Section 5 along with future work.

2. Missing value imputation based on DTW, GO, and KNN

In order to impute missing values effectively, we propose a novel and effective approach. Our approach takes both gene expression values and external biological information for genes into account. It is carried out with the following four steps.

Step 1: **Calculation of DTW distance for each gene pair.**

To find the closeness of gene pairs within their expression values, we choose the dynamic time warping (DTW) distance measurement [11,12]. DTW algorithm is used in our approach as the substitution for commonly-used Euclidean distance while estimating similarity between gene pairs within their expression values. This is because the importance of finding whether there exist subsequences with highly similar relations is emphasized while analyzing whole microarray time series data [13]. If two series with time slot points are given as input, the DTW algorithm can discover the best possible alignment between them by calculating the minimum sum of whole matched points on the two time series. DTW is a recursive algorithm that starts with matching each point-to-point pair from the first element to the last element on the two input sequences. With DTW mapping method, local similarity can be found as the best mapping path within the two sequences to be aligned. Equations of DTW algorithm are as follows:

Distance of two time slot points:

The distance between the elements of the two time series is computed as:

$$dis(i,j) = |x_i - y_i| \tag{1}$$

Base Conditions:

$e(0,0) = 0$;
$e(1,1) = dis(x_1,y_1)*W_D$;
$e(i,0) = \infty$ for $1 \leqq i \leqq I$; (2)
$e(0,j) = \infty$ for $1 \leqq j \leqq J$;

where W_D is the weighted value for the paths in the diagonal direction.

Recursive Relation:

$$e(i,j) = \min \begin{cases} e(i,j-1) + dis(x_i,y_j)*W_V \\ e(i-1,j-1) + dis(x_i,y_j)*W_D \\ e(i-1,j) + dis(x_i,y_j)*W_H \end{cases} \tag{3}$$

where Wv, W_D, and W_H denote the weighted values for the paths in the vertical, diagonal, and horizontal directions respectively.

Output: DTW distance for two sequences X and Y:

$$\text{DTW}(X,Y) = \frac{1}{n+m} * e(i,j) \tag{4}$$

where lengths of X and Y are n and m respectively.

For the improvement of DTW algorithm, we also perform some adjustments on it in order to increase the effectiveness and accuracy of our approach for missing imputation. Due to space limitation, here we mention only the names of the two adjustments that generate the best results according to our evaluation: Fast DTW and Slope Weighting [14,15]. For more information, please refer to the references.

Step 2: Calculation of GO semantic similarity for each gene pair.

Analyzing only gene expression values in microarray data is not enough. External information of genes which provides biological functions or important meanings also needs to be concerned. For this purpose, we add gene ontology (GO) information for genes into our approach. GO is a biological definition and annotation for genes that describes the biological meanings of each gene. Most known genes have specific annotations in GO structure within three independent domains: molecular function (MF), biological process (BP), and cellular component (CC). Terms within three above domains record and represent various molecular or biological meanings for each annotated gene from different aspects respectively. [16]

With the application of GO annotations for genes, similarity measurement between gene pairs can be performed more accurately and effectively. Therefore, the quantitative representation of GO terms for each gene is required to estimate how similar each gene pair is. To our best knowledge, the first literature presenting methods that use gene ontology is in [17]. The authors utilize the concept of information content to identify the importance of each GO term by calculating the probability of occurrence of each GO term (called the p values) in the whole GO structure. Afterward, we can estimate the similarity of each gene pair within GO semantic aspect by calculating the mean p values of the shared GO terms used to annotate the two genes. Operations of the algorithm proposed by the authors can be briefly descried as follows:

- First find the sets of GO ids for each pair of genes being identified.
- Create a table recording the tracing path of all terms annotated for both genes.
- Calculate the probability of the occurrence of each term in the table.
- Estimate all the parent-children relations of each term in the path-tracing table to determine whether the two genes have common ancestors.
- For genes that have shared parent nodes in the GO tracing path, calculate the mean probability of occurrence of all their matched GO term combinations.
- The mean probability of occurrence is taken as the distance between gene pairs.

However, the authors in the work only find the minimum p value of shared ancestors of GO terms for two genes. This operation is insufficient because in GO structure two GO terms that are used to annotate different genes may have several relations. Theoretically, two genes with GO terms in common tend to be more relative than two genes having GO terms that only have shared ancestors. As a result, our approach gives different weighted values while calculating p values for the three term-term relations: the same terms, parent-children relation, and ancestor-sharing relation. The three relations are marked as case1, case2, and case3 in order as shown in Figure 2.

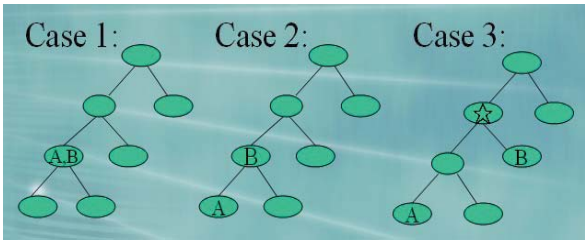

Figure 2.The three relations between GO term pairs

The star symbol in Figure 2 illustrates the closest shared ancestors or two GO terms A and B. To find the best weighted values for these three relations, we implement several parameters and the results will be discussed in Section 4. Finally, the mean p value of all GO term pair combinations for two genes is used to the further semantic similarity measurement of these two genes.

Moreover, some GO terms in the whole GO structure tend to be more informative than the others. These representative GO terms are commonly used to annotate differentially expressed genes between an object sample and a control sample in one microarray experiment. In other words, if we can find out these informative GO terms and give corresponding weighted values to them as we calculate the GO semantic similarity of gene pairs, the similarity measurement will be more accurate because these terms are found to significantly annotate a set of genes with similar reactions to the experiment. There is one method proposed to find these data-specific GO terms [18]. According to the authors, the steps to find these GO terms are as follows:

(I) Preparation of a pairwise comparison matrix for each GO term

The algorithm starts by preparing a pairwise comparison matrix for each GO term. To identify whether one GO term significantly annotates certain of differentially expressed genes, a corresponding matrix for the comparison of expression values of genes annotated by this GO term is created. The X-axis and the Y-axis in the matrix are expression values of the genes annotated by the GO term at time slot I, II, III…etc. For example, if we are going to identify a GO term A and see if it is informative or not, we will first create a matrix belonging to GO term A with expression values at each time slot for the genes annotated by GO term A listed in the matrix.

(II) Identification of differentially expressed genes between two microarray data

After the comparison matrixes of each GO term are created, we are now going to fill in each cell in the matrixes. A table recording all genes annotated by the GO term and the ratio of their gene expression values at certain time slots is build as the content of each cell in the matrix. For example, the cell (I, II) in the matrix indicates that gene expression values of all genes annotated by the GO term at time slot 1 divided by gene expression values at time slot 2 is performed. After repeating this operation, all cells except the diagonal line in the matrix are processed with different ratios standing for how differentially the genes are expressed between the two time slots. Next, we need to define a threshold for ratios of gene expression values so that genes with ratios of expression value for two time slots larger than the threshold are taken as differentially expressed genes. Cells with more than one differentially expressed gene in them are then marked gray in the comparison matrix. To define the threshold for the ratio of gene expression values, a statistical method is used. The statistical testing method is based on the following equation:

$$P-value=\sum_{j}^{\min(n,M)}\frac{C_{j}^{M}*C_{n-j}^{N-M}}{C_{n}^{N}} \qquad (5)$$

where N is the number of genes examined by the microarray experiment that we refer to as "population gene set", M is the number of genes annotated to the matrix-linked GO term in the population gene set, n is the number of differentially expressed genes between microarray data, and j is the number of genes assigned to the matrix-linked GO term in the differentially expressed genes. Figure 3 is an example of this operation.

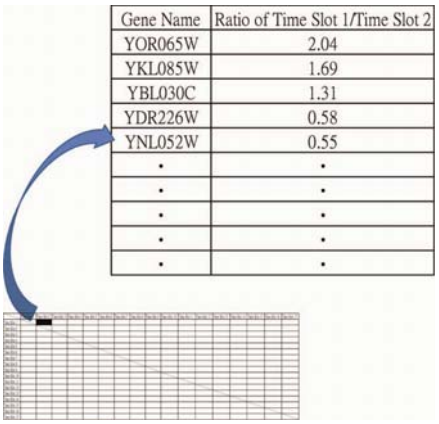

Figure 3.Example of each cell in the comparison matrix

(III) Identification of data-specific GO terms

The last step in the method is to determine whether gray cells are concentrated in any rows compared to whole cells. The theorem is very intuitive that if one GO term annotates genes which are significantly differentially expressed, then this GO term tends to be more informative than the other terms. After comparison matrixes are built for all involved GO terms, we now need to determine how enriched the gray cells are in each matrix. Similar to previous step, it requires a statistical operation that cannot be explained in detail due to space limitation. Please refer to [18] if interested.

Step 3: Definition of the similarity measurement of gene pairs.

Our approach aims to provide an accurate similarity measurement that takes both gene expression values and external information for genes into account so that the similarity measurement can be combined with the KNN method to impute missing values. We modify the equation proposed in [18] according to our evaluation as follows:

$$Sim(g_x,g_y)=D^{GO\_NEW}(g_x,g_y)^{\alpha}*D^{DTW}(g_x,g_y) \qquad (6)$$

where D^{GO_NEW} is our estimation of p values of all GO term pairs used to annotate gx and gy, α is the positive weighted parameter, and D^{DTW} is the DTW distance of (gx, gy). In equation (6), we replace Euclidean distance with DTW distance, and replace original p value estimation with our approach. This is because we consider that DTW is more suitable than Euclidean distance while calculating distance between gene expression values. Equally, we use our new estimation for semantic distance between gene pairs to retrieve higher accuracy.

Step 4: Combination of the similarity measurement with the KNN method.

Our imputation approach is combined with the KNN method but with some modifications. Here we use our similarity measurement as the estimation to determine the closeness of gene pairs. The steps of our approach for missing value imputation are as follows:

1. In order to impute the missing value $G_{I,J}$ for gene I at time slot J, the KNN-impute algorithm chooses k genes that are most similar to the gene I and with the values in position J not missing.
2. The missing value is estimated as the weighted average of the corresponding entries in the selected k expression vectors:

$$G_{I,J} = \sum_{i=1}^{k} W_i \times e_{iJ} \quad (7)$$

where e_{iJ} are expression values of the k selected genes in the similar gene set.

3. The weighted value

$$W_i = \frac{1}{Sim(g^*, g_i) \times \Delta} \quad (8)$$

$$\Delta = \sum_{i=1}^{k} [1/(Sim(g^*, g_i)] \quad (9)$$

where and g* denotes the set of k genes closest to g_i, $Sim(g^*, g_i)$ is our similarity measurement as shown in equation (6). Missing values for the target gene are hence imputed with our approach.

3. Dataset and assessment of imputation accuracy

We use the microarray dataset obtained for genes of Yeast Saccharomyces cerevisiae cells with four synchronization methods: alpha-factor, cdc15, cdc28, and elutriation by Spellman et al [2]. Spellman's dataset is widely used as the real dataset in microarray research. These four subsets of the dataset contain totally 6178 gene ORF profiles with their expression values across various amounts of time slots. In the dataset, the alpha sub-dataset contains 18 time points with seven minutes as the time interval, while the cdc28 sub-dataset contains 17 time points with ten minutes as the time interval.

For the assessment of imputation accuracy, genes with missing values in microarray gene expression data are first filtered to generate a complete matrix. Missing values with different missing rates ranging from 1%, 5%, 10%, 15% and 20% in the complete matrix are deleted at random to create testing datasets. We then calculate the Normalized Root Mean Square (NRMS) error. Equation for NRMS error is as follows:

$$\sqrt{mean[(y_{predict} - y_{known})^2]} \Big/ std[y_{known}] \quad (10)$$

where $y_{predict}$ and y_{known} are estimated values and known values in the complete matrix respectively, and $std[y_{known}]$ is the standard deviation of known values. An imputation method is said to outperform others if the NRMS error of it is less than that of other imputation methods.

4. Experimental results and discussion

For our evaluation, we use only parameters that generate the best imputation results (the lowest NRMS values). We perform missing value imputation on alpha and cdc28 sub-datasets with our approach. We also implement existing methods such as the KNN method and BPCA for comparison. Experimental results are shown in Figure 4 and Figure 5 for alpha and cdc28 sub-datasets respectively. According to the results, our approach is the most effective method modified based on FastDTW with slope weighting and proper parameters for GO similarity measurement. Sequences of effectiveness of these imputation methods may change a little in certain percentage of missed data. This may result from the randomness while deciding which values to be removed in the complete matrix. To sum up, using our approach with suitable parameters can retrieve the best imputation results. He we only show the best imputation results of our approach with the most proper parameters due to space limitation.

5. Conclusion

In this paper, we present a novel approach for missing value imputation in microarray data. Our approach takes both expression values and GO semantic information for genes into account. Compared with other methods, our approach is more effective in terms of NRMS errors. We are now working on the survey of the influence of different parameters used in our approach on the imputation results. We hope to further improve our approach by adjusting proper parameters.

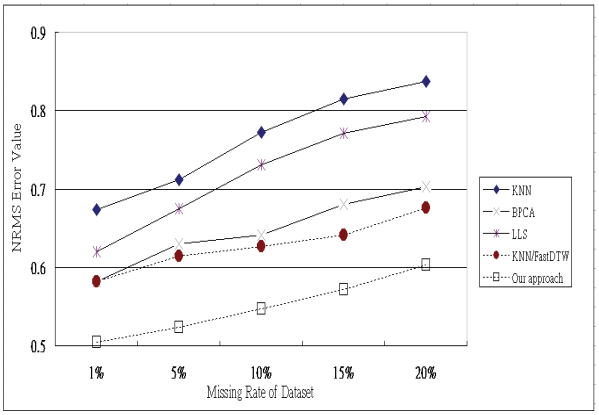

Figure 4.Imputation results for alpha sub-dataset

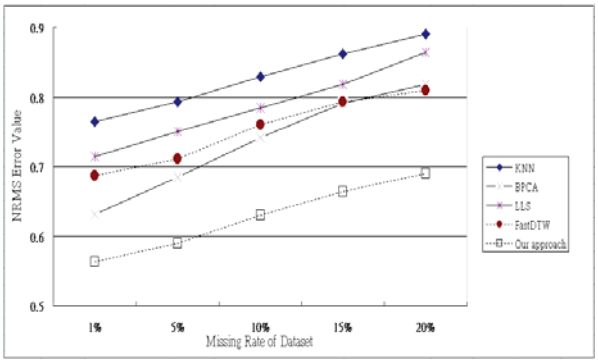

Figure 5.Imputation results for cdc28 sub-dataset

Acknowledgment

This work was supported in part by National Science Council (NSC) with project number NSC 99-2221-E-032-063.

References

[1] J. DeRisi, R. Iyer, and Brown P, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol.278, pp.680-686, 1997.

[2] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell*, vol.9, pp.3273-3297, 1998.

[3] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy," *Classification, Clustering and Data Mining Applications*, pp.639-648, 2004.

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

[5] A.C. Yang, H.H. Hsu, and M.D. Lu, "Outlier filtering for identification of gene regulations in microarray time-series data," in: *Proc. of the 3rd Intl. Conf. on Complex, Intelligent and Software Intensive Syst.*, pp.854-859, 2009.

[6] V.S. Tseng, L.C. Chen, and J.J. Chen, "Gene relation discovery by mining similar subsequences in time-series microarray data," in: *Proc. of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biol.*, pp.106-112, 2007.

[7] A. Mohammadi and M.H. Saraee, "Estimating missing value in microarray data using fuzzy clustering and gene ontology," in: *Proc. of the IEEE Intl. Conf. on Bioinformatics and Biomedicine*, pp.382-385, 2008.

[8] Q. Xiang and X. Dai, "Improving missing value imputation in microarray data by using gene regulatory information," in: *Proc. of the 2nd Intl. Conf. on Bioinformatics and Biomedical Eng.*, pp.326-329, 2008.

[9] Y. Yamada, K.I. Hirotani, K. Satou, and K.I. Muramoto, " An identification method of data-specific GO terms from a microarray data set," *IEICE Trans. on Inf. and Syst.*, vol.E92-D, pp.1093-1102, 2009.

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

[11] C. Furlanello, S. Merler, and G. Jurman, "Combining feature selection and DTW for time-varying functional genomics," *IEEE Trans. on Sig. Processing*, vol.54, pp.2436-2443, 2006.

[12] H.M. Yu, W.H. Tsai, and H.M. Wang, "Query-by-singing system for retrieving karaoke music," *IEEE Trans. on Multimedia*, vol.10, pp.1626-1637, 2008.

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

[14] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol.11, pp.561-580, 2007.

[15] J.B. Kruskall and M. Liberman, "The symmetric time warping algorithm: from continuous to discrete," in: *Time Warps, String Edits, and Macromolecules: The theory and Practice of String Comparison*, 1983.

[16] Gene ontology website, URL: http://www.geneontology.org/, last accessed on October 5th, 2010.

[17] J. Tuikkala, L. Elo, O.S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in micorarray data with gene ontology," *Bioinformatics*, Vol. 22, pp.566-572, 2006.

[18] Y.Yamada, K. Hirotani, K. Satou, and K. Muramoto, "An identification method of data-specific GO terms from a microarray data Set," IEICE Transactions on Information & Systems, Vol. E92-D, No.5, May 2009