2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Missing Value Estimation in Microarray Data by Fuzzy Clustering and Gene Regulatory Information

Shekofeh Yaraghi [1+], Mohammad Davarpanah Jazi[2] and Vahid Rafeh[3]

[1]Department of Computer, Islamic Azad University, Arak Branch, Arak, Iran

[2] Computer& Information Technology Department, Foulad Institute of Technology, Fouladshahr, Esfahan, Iran

[3] Arak University, Arak, Iran

Abstract. Gene expression microarray experiments produce datasets with numerous missing expression value, which can significantly affect the performance of statistical and machine learning algorithms. In this paper, we proposed a novel method, namely the fuzzy clustering and histone acetylation (FCMHAimpute), to estimate missing value to microarray gene expression. In this proposed method, missing values are imputed with values created from cluster centers. External information such as gene regulatory information is used to determine the similar genes in clustering process. We have applied the proposed method on two datasets with different percentage of missing values. The experimental results demonstrate that the proposed method provides a higher accuracy of missing value estimation especially at higher missing percentages because histone acetylation information may be more highly correlated with gene expression than that of similarity.

Keywords: Missing Value, Semantic Similarity, Fuzzy Clustering, Microarray, Histone Acetylation.

1. Introduction

Gene expression microarray provides a popular technique to monitor the relative expression of thousands of genes under a variety of experimental conditions [1]. Gene expression microarray experiments can generate datasets with multiple missing expression values due to various reason, e.g. insufficient resolution, image corruption, dust or scratches on slides, or experimental error during the laboratory process. Datasets are an m*n gene expression matrix with m genes end n experiments. Unfortunately, many algorithms for gene expression analysis need a complete matrix of gene array values as input [2, 3]. Therefore, these missing values need to be fill because each value is important to determine the validity and accuracy of a special analysis.

Certainly, there are many strategies to deal with missing values such as: filling the missing values with zeros; using the row mean for imputation. These methods produce inaccurate estimating values. Then a number of complicated approaches have been proposed to predict missing value [4] such as K-Nearest Neighbor (KNN) that their disadvantage is depend on K parameter.

We use cluster-based algorithms for estimating missing value because they don't need user to determine parameters [5] and another limitation of the existing estimation methods for microarray estimation is that they use no external information and the estimation is based solely on the expression data [6]. In [7] a method based on the Fuzzy C-means clustering and Gene Ontology (FCMGOimpute) have been proposed that exploits the similarities in GeneOntology (GO) database to select the neighbor gene.

[+] Corresponding author.
E-mail address: Shekofeh.yaraghi@yahoo.com

In this paper, we propose a new missing value estimation method based on fuzzy C-means clustering algorithm and histone acetylation (FCMHAimpute) for detect similar genes to the gene with missing value.

The remainder of this paper can be described as follow: Next section contains a description FCMimpute method. In section 3 the proposed methodology and in section 4 shows the results of method applied on two datasets. The paper ends with conclusion.

2. FCMImpute

Clustering analysis of data from DNA microarray hybridization studies is essential for identifying biologically relevant groups of genes. Partitional clustering methods such as K-means or Self-Organizing Maps assign each gene to a single cluster. However, these methods do not provide information about the influence of a given gene for the overall shape of clusters. Here we apply a fuzzy partitioning method, Fuzzy C-Means (FCM), to attribute cluster membership values to the genes, where single genes may belong to several clusters. In fuzzy clustering, each point has a membership degree to cluster between 0 or 1.

FCM partitions a given data set, X= $\{g_{1,}g_{2,\ldots,}g_n\}$, and let C be the number of clusters. Then degree of belonging of data object x_k to cluster i is defined as U_{ik}, which explain the bellow constraints:

$$\sum_{k=1}^{n} U_{ik} > 0 \quad (1)$$

$$\sum_{i=1}^{c} U_{ik} = 1 \quad (2)$$

Fuzzy c-means clustering is based on minimization of the following objective function:

$$J(U,C) = \sum_{i=1}^{c}\sum_{k=1}^{n} U_{ik}^{m}\ d_{ik}^2 \quad (3)$$

Where m is the fuzziness parameter which is a real number greater than 1 and d_{ik}^2 is the Euclidean distance between data object x_k and cluster center i which is defined by:

$$d_{ik}^2 = \sum_{j=1}^{s}(g_{kj} - c_{ij})^2 \quad (4)$$

FCM clustering algorithm minimizes the objective function shown in Equation (3), by computing of the cluster center and membership degrees, iteratively by Equation (5) and (6).

$$c_i = \frac{\sum_{k=1}^{n}(U_{ik})^m x_k}{\sum_{k=1}^{n}(U_{ik})^m} \quad (5)$$

$$U_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{m-1}}} \quad (6)$$

The FCM algorithm do not consider the missing values of data genes in the clustering procedure. Consequently for an incomplete dataset the Euclidean distance between g_k and cluster center c_i is calculated by [8]:

$$d_{ik}^2 = \sum_{j=1}^{s}(g_{kj} - c_{ij})^2 = \frac{s}{e_k}\sum_{j=1}^{s}\left(g_{kj} - c_{ij}\right)^2 e_{kj} \quad (7)$$

Where $e_k = \sum_{j=1}^{s} e_{kj}$

We calculate cluster canters and membership degree by equation (8) and (9) for minimize the objective function shown:

$$c_{ij} = \frac{\sum_{k=1}^{n}(U_{ik})^m\ e_{kj} g_{kj}}{\sum_{k=1}^{n}(U_{ik})^m\ e_{kj}} \quad (8)$$

$$U_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{m-1}}} \quad (9)$$

To determine the fuzziness parameter (m) and the number of clusters(c), some methods were proposed in [5, 7].

3. Combine FCM and Histone Acetylation

In this paper we have utilized the incorporation of fuzzy c-means and histone acetylation for detect similar genes to the gene with missing value.

The sites of acetylation include at least four highly conserved lysines in histone H4 (K5, K8, K12 and K16), five in histone H3 (K9, K14, K18, K23 and K27), as well as less conserved sites in histone H2A and H2B [9]. The acetylation sites can be integrated to create unique acetylation patterns that involve differently acetylated sites. Here we used the clustering results of [10] to formulate clusters of genes with similar

acetylation patterns across the 11 residues, which are marked as $A_{O_1},\ldots A_{O_n}$ and $A_{I_1},\ldots.A_{I_n}$ for intergenic regions (IGRs) and open reading frames (ORFs), relatively, where O_n and I_n are the maximal number of clusters for ORFs and IGRs. Each cluster consists of the genes which share common acetylation pattern and M is denote to be the total number of the acetylation patterns on IGRs and ORFs, M= $A_{O_n}+A_{I_n}$ [11].

We modify the calculation of Euclidean distance in (10) as follow:

$$d_{ik}^2=\sum_{j=1}^{s}\left(g_{kj}-c_{ij}\right)^2= \left(\frac{s}{e_k}\sum_{j=1}^{s}\left(g_{kj}-c_{ij}\right)^2 e_{kj}\right)\left(1-\frac{\sum_{t=1}^{m}U_{it}^{m}H_{kt}}{M}\right) \quad (10)$$

In the above formula, the first term is Euclidean distance of gene g_k and cluster center i for incomplete data, based on their expression level. The second term use gene regulatory information.

In the formula, is defined based on histone acetylation of gene k and gene t, as follows:

$$H_{kt}=\begin{cases}1 & if\ g_k\ and\ g_t\ have\ the\ same\ histone\ acetylation\\ 0 & otherwise\end{cases}, 1\le t\le M \quad (11)$$

Therefore the histone information of g_k is compared with histone information of all genes belonging to cluster i, more genes have the same histone information, more the distance shrink. So the genes which belong to cluster i with higher membership degree, have more influence. Calculation of cluster centers and membership degree is the (6) and (7).

We impute missing values by making use of the weighted mean of the values of the corresponding attribute over all clusters. The weighting factors are the membership degree of a gene to the cluster. The missing gene expression value is imputed by:

$$g_{ik}=\frac{\sum_{i=1}^{c}u_{ik}^{m}c_{ij}}{\sum_{i=1}^{c}u_{ik}^{m}} \quad (12)$$

4. Experimental Results

We compared our proposed method (FCMHAimpute) with FCM, FCMGOimpute and KNNimpute. In order to evaluate the effectiveness of our method, We used yeast cell cycle data [12, 13]. The histone acetylation data used in this work were from [10]. We deleted rows with missing value to achieve a complete dataset as test dataset.

We compared accuracy of different imputation methods by the Normalize Root Mean Squared Error (NRMSE):

$$NRMSE=\frac{\sqrt{mean\left[\left(y_{predict}-y_{known}\right)^2\right]}}{std[y_{known}]} \quad (13)$$

Where $y_{predict}$ and y_{known} are vectors whose elements are the predict values and the known values, respectively, and std[y_{known}] is the standard deviation of the known values. We have applied KNNimpute, FCMimpute, FCMGOimpute and our proposed method (FCMHAimpute) on two datasets with different percentage of missing values and compared the accuracy of them by means of NRMSE. The result experiments are shown in the figure 1, 2. KNNimpute has a lower performance compare to other methods because KNNimpute depend on K parameter (number of gene neighbor) to estimate missing values. FCMimpute has better performance than KNNimpute but FCMimpute method doesnot use any useful external information and it uses just microarray data for imputation process. FCMGOimpute and FCMHAimpute have lower RMSE because both used external information. The proposed method (FCMHAimpute) has better result in term of accuracy because histone information may be more highly correlated with gene expression than that of similarity on GO.

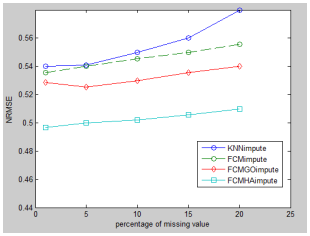

Fig. 1: comparison of the accuracy of KNNimpute, FCMimpute, FCMGOimpute and FCMHAimpute methods by RMSE for non-time series dataset of Gasch over 1 to 20% missing data.

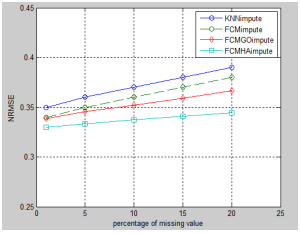

Fig. 2: comparison of the accuracy of KNNimpute, FCMimpute, FCMGOimpute and FCMHAimpute methods by RMSE for time series of yeast over 1 to 20% missing data.

5. Conclusion

In this paper, an efficient method for estimating missing values in microarray data, namely FCMHAimpute, is proposed.

We have analyzed the performance of our method on two datasets and compared the accuracy with KNNimpute, FCMimpute, FCMGOimpute, FCMHAinpute. We used from advantage of the acetylation information to promote the imputation. The theoretical basis is that acetylation state in chromatin provides a straight forward mechanism to coordinate the regulation of co-expressed genes. The proposed method considers gene expression and a acetylation information in a combined way for missing value imputation. Experimental results show that FCMHA method outperforms other methods in term accuracy, especially at higher missing percentage. For future work we make a decision to use from the other external information.

6. References

[1] A. Brazma, J. Quackenbush, H. Causton, "Microarray Gene Expression Data Analysis: A Beginner's Guide", *Wiley-Blackwell,* 2003.

[2] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T.Hastie, R. Tibshirani, D.Botstein and Russ B. Altman , " Missing value estimation methods for DNA microarrays," *Bioinformatics*, 2001, p.p 520–525.

[3] A.Kaur, S.Bir, R.kamel,"Approximation of Missing Values in DNA Microarray Gene Expression Data", *Interational Journal of Computer Application*, 2010, No 3, pp.20-28.

[4] I. Scheel, M. Aldrin, I. K. Glad, R. Sorum, H. Lyng, and A. Frigessi. , "The influence of missing value imputation on detection of differentially expressed genes from microarray data". *Bioinformatics*, 2005 , pp.4272-4279.

[5] J. Luo, T. Yang, and Y. Wang, " 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.

[6] J. Tuikkala, L. Elo, O. S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in microarray data with geneontology", *Bioinformatics*, 2006, vol. 22, pp. 566-572.

[7] A . Mohammadi, M . Saraee," Dealing with Missing Values in Microarray Data," *2008 International conference on Emerging Technologies*, 2008, pp.258 – 263.

[8] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-Means clustering of incomplete data," *IEEE, Transactions On Systems, Man, Cybernetics,* 2001, vol.31, pp.735-744.

[9] O. J. Rando, "Global patterns of histone modifications," *Current Opinion in Genetics and Development*, 2007, vol -17, pp.94-99.

[10] S. K. Kurdistani, S. Tavazoie and M. Grunstein, “Mapping global histone acetylation patterns to gene expression,” *Cell*, 2004, vol 117, pp.721-733.

[11] Q. Xiang, X. Dai, “Improving Missing Value Imputation in Microarray Data by Using Gene Regulatory Information,” *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008,* 2008 , pp.326-329.

[12] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, “Genomic expression programs in the response of yeast cells to environmental changes ,” *mol Biol Cell*, 2000, vol. 11, pp. 4241-57.

[13] http://genome-www.stanford.edu/Mec1/data.shtml.