Method FCMSPimpute for Estimation of Missing Values in Microarray Data

Shekofeh Yaraghi ¹ and Mohammad Davarpanah Jazi ²

¹Department of Computer, Arak Branch, Islamic Azad University, Arak, Iran
shekofeh.yaraghi@yahoo.com

² Computer & Information Technology Department, Foulad Institute of Technology,
Fouladshahr, Esfahan, Iran
mdjazi@cc.iut.ac.ir

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 based on the fuzzy clustering and the shortest path algorithm for measuring the semantic similarity on GO to estimate missing value to microarray gene expression. In this proposed method, missing values are imputed with values generated from cluster centers. Genes similarity in clustering process resolute based on both the GO structure information and terms property. We have applied the proposed method on two datasets with different percentages of missing values. The experimental results indicate that proposed method provides a higher accuracy of missing value estimation because the semantic similarity obtained by sp algorithm better correlates with the expression similarity than other node-based methods.

KEYWORDS

Missing Value; Semantic Similarity; Fuzzy Clustering; Microarray.

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 $m \times n$ gene expression matrix with m genes and 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). There are two steps in KNN imputation. The first step is to select k genes with expression profiles similar to missing value. the second step is estimated by a weighted average of the known values in the j -th experiment of the k selected genes but their disadvantage is depend on K parameter.

2. RELATED WORK

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]. The gene ontology (GO) has become one of the most important and usefull resources in bioinformatics. Fuzzy C-means clustering algorithm and gene ontology (FCMGOimpute) have been proposed to avert the problem of those methods [7].

In this method two genes will be similar if they have the same annotations. This similar measure is not good enough. In [8] Fuzzy C-means clustering algorithm and semantic similarity (FCMSSimpute) have used semantic similarity for measure similarity between genes.

In this paper, we change semantic similarity method on GO to measure for increasing accuracy of imputation because in previously method similarity use only information content but in this paper we use structure and IC information to compute semantic similarity.

The remainder of this paper can be described as follows: Next section contains a description of FCMGOimpute method. In section 3 the proposed methodology and the semantic similarity methods are described and in section 4 discusses the results of method applied on yeast cell cycle dataset. The paper ends with conclusions and future work.

3. FCMGOimpute

Clustering analysis of data from DNA microarray hybridization studies is essential for identifying biologically relevant groups of genes .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[7].

Let $X = \{g_1, g_2, \dots, g_n\}$ be the set of given examples 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 [6]:

$$d_{ik}^2 = \sum_{j=1}^s (g_{kj} - c_{ij})^2 = \left(\frac{s}{e_k} \sum_{j=1}^s (g_{kj} - c_{ij})^2 e_{kj} \right) \left(1 - \frac{\sum_{t=1}^s U_{it}^m B_{kt}}{N} \right) \quad (4)$$

Where s is the feature space dimension, $e_k = \sum_{j=1}^s e_{kj}$ and B_{kt} is defined based on gene ontology annotation of gene k and gene t , as follows:

$$B_{kt} = \begin{cases} 1 & \text{if } g_k \text{ and } g_t \text{ have the same annotation} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

We calculate cluster canters and membership by Eq. (6) and (7) for minimize the objective function shown:

$$c_{ij} = \frac{\sum_{k=1}^n (U_{ik})^m e_{kj} g_{kj}}{\sum_{k=1}^n (U_{ik})^m e_{kj}} \quad (6)$$

$$U_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}^2}{d_{jk}^2} \right)^{\frac{1}{m-1}}} \quad (7)$$

To determine the fuzziness parameter (m) and the number of clusters (c), some methods were proposed in [7]. In FCMGOimpute two genes are similar, if they have same annotation, and they are dissimilar if their annotations are different. According to definition similarity value will be 0 or 1 but semantic similarity is not crisp and a real value between 0 or 1.

4. Semantic Similarity Methods

Semantic similarity measures can be used to calculate the similarity of two concepts in ontology. There are three types of approach to compare this term in a ontology such as Go: edge based, node-based, and hybrid of the both.

1- Edge based approaches [9] are based mainly on counting the number of edges in the graph path between two term.

2- Node based approaches are based mainly on information content(IC). The IC value for a term t is defined as:

$$IC(t) = -\log p(t) \quad (8)$$

Where p(t) is the probability of occurrence of the term t in a certain corpus. All node based on methods are defined based on the IC value of the terms involved. The methods are shown to measure the semantic similarity for GO term:

Rensink's method [10] is shown in Eq(9):

$$sim_{rensink}(t_1, t_2) = \max IC(t) \quad (9)$$

Lin's [11] and Jaing's method [12] is shown in Eq(10, 11):

$$sim_{lin}(t_1, t_2) = \frac{2 \cdot IC}{IC(t_1) + IC(t_2)} \quad (10)$$

$$sim_{jaing}(t_1, t_2) = 1 - (IC(t_1) + IC(t_2) - 2 \cdot IC(t)) \quad (11)$$

Where t is the most informative common ancestor (MICA) of t_1 and t_2 . Edge based methods, the weights of the edges conflict with the property of GO, and for node-based methods, only IC value of the two terms and their MCIA are considered regardless of their position in GO.

3- Hybrid methods that consider both the substructure of GO and the property of terms involved, unlike the existing edge based(node based) methods that use only structure (IC) information.

5. Proposed Method

In FCMSSimpute have been used Lin's method to compute semantic similarity which is node-based and they rely on the property of the terms which is represented using the concept derived from information theory [8]. We used a new hybrid method namely shortest path algorithm for measuring the semantic similarity on GO (FCMSPimpute) that contains more information than the single IC values used in the node-based algorithm and the weights assigned to the substructure which are more consistent with common interpretation than the previous edge-based methods.

The shortest path algorithm can be described as follow [13]. Given two term t_i and t_j , the normalized semantic similarity between them is defined as:

$$Sim_{SP}(t_i, t_j) = 1 - \frac{\arctan \left[\frac{\sum_{t_1 \in path} \frac{1}{A_{IC}(t_i)} + \sum_{t_1 \in path} \frac{1}{B_{IC}(t_j)}}{\left(\frac{\pi}{2}\right)} \right]}{\left(\frac{\pi}{2}\right)} \quad (12)$$

The function $\arctan$ normalizes the distance obtained by summing the weight of the terms on the shortest path [0-1]. SP algorithm weights each term using the value of $\frac{1}{IC}$.

We will use gene semantic similarity as [14]:

$$s_{sp}(g_i, g_j) = \max \{ (sim_{sp}(t_i, t_j) | t_i \in annotation(g_i), t_j \in annotation(g_j)) \} \quad (13)$$

We modify the calculation of Euclidean distance in Eq. (14) as follow:

$$\sum_{j=1}^S (g_{kj} - c_{ij})^2 = \left(\frac{S}{e_k} \sum_{j=1}^S (g_{kj} - c_{ij})^2 e_{kj} \right) \left(\frac{\sum_{t=1}^n U_{it}^m s_{sp}(g_i, g_j)}{N} \right) \quad (14)$$

Calculation of cluster centers and membership degree is the Eq. (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 degrees of a gene to the cluster. The missing gene expression value is imputed by:

$$g_{ik} = \frac{\sum_{i=1}^c u_{ik}^m c_{ij}}{\sum_{i=1}^c u_{ik}^m} \quad (15)$$

6. Experimental Results

The time series dataset are used to evaluate propose method [15]. So we have conducted extensive experiments on dataset from the UCI machine learning repository [16]. We evaluate our method on the dataset abalone, Which contains 8 continuous attributes and 4177 instances in total. The genes annotation were retrieved from the Saccharomyces Genome Database (SGD) [17]. In our experiments, we used the annotations from BP ontology. We deleted rows with missing value to achieve a complete dataset as test dataset.

The performance of the missing value estimation is evaluated by the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{n}} \quad (16)$$

Where y_j and $\hat{y}_j$ are the real value the estimated value, respectively, and n is the number of missing value. We have applied FCMimpute, FCMGOimpute, FCMSSimpute and our proposed method (FCMSPimpute) on yeast cell cycle data with different percentage of missing values and compared the accuracy of them by means of RMSE. The result experiments are shown in the figure 1, 2. FCMimpute method does not use any useful external information and uses just microarray data for imputation process. As it can be seen from the results, FCMimpute has a lower performance, compared to other method. FCMGOimpute gene ontology annotation as an extra knowledge but similar concept is 0 or 1. FCMSSimpute and FCMSPimpute have lower RMSE. The proposed method (FCMSPimpute) we use gene ontology annotation as an external information, and semantic similarity on GO based on both the GO structure information and term's property. FCMSPimpute has better performance over FCMSSimpute because the semantic similarity obtained by sp algorithm better correlates with the expression similarity than other node-based methods.

Figure 1. Comparison of the accuracy of FCMimpute, FCMGOimpute, FCMSSimpute, FCMSPimpute methods by RMSE for time series dataset over 1 to 20% missing data.

Figure 2. Comparison of the accuracy of FCMimpute, FCMGOimpute, FCMSSimpute, FCMSPimpute methods by RMSE on dataset abalone, over 1 to 20% missing data.

7. CONCLUSIONS

In this paper, an efficient method for estimating missing values in microarray data, namely FCMSPimpute, is proposed. We use benefit of the correlation structure of the data to estimate missing expression values by clustering, as well as using gene semantic similarity which improves the imputation accuracy.

We have explained the performance of our method on two datasets and compared the accuracy with FCMimpute, FCMGOimpute, FCMSSimpute and FCMSPimpute. The result experiments are shown, the proposed method (FCMSPimpute) out performs other methods in terms of accuracy. In FCMSPimpute, we have changed semantic similar method and We have considered two genes similar if they have same biological process annotations instead Molecular Function.

REFERENCES

- [1] Brazma, A. Quackenbush, J. Causton, H. (2003). Microarray Gene Expression Data Analysis: A Beginner's Guide, ", Wiley-Blackwell.
- [2] Troyanskaya, O. Cantor, M. Sherlock, G. Brown, P. Hastie, T. Tibshirani, R. Botstein ,D and Russ B. Altman,(2001): Missing value estimation methods for DNA microarrays, *Bioinformatic*, pp. 520-525.
- [3] A.Kaur, S.Bir, R.kamel, (2010) "Approximation of Missing Values in DNA Microarray Gene Expression Data", *Interational Journal of Computer Application*, No 3, pp.20-28.
- [4] Xiang, Q. Dai, X (2008): Improving Missing Value Imputation in Microarray Data by Using Gene Regulatory Information, *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008*, pp.326-329.

- [5] Luo, J. Yang, T and Wang, Y .(2005) , Missing Value Estimation For Microarray Data Based On On Fuzzy C-means Clustering,” in *Proceedings of the Eighth International Conference on High-Performance Computing in Asia-Pacific Region.*”
- [6] Tuikkala, J. Elo, L. Nevalainen, O. S. and Aittokallio, T. (2006), Improving missing value estimation in microarray data with geneontology, *Bioinformatics*, vol. 22, pp. 566-572
- [7] A . Mohammadi, M . Saraee, (2008)” Dealing with Missing Values in Microarray Data,” *2008 International conference on Emerging Technologies*, pp.258 – 263.
- [8] Pourhashem, M . Kelarestaghi, M . Pedram, M . (2010), Missing Value Estimation In Microarray Data Using Fuzzy Clustering And Semantic Similarity, “*Global Journal of Computer science and technology*”, pp 18-22.
- [9] Pekar, V. and Staab, S. (2002),Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision, “In: *Proceedings of COLING-2002*. Taipei”. pp. 786-792.
- [10] Resnik, (1999), Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language,” *Journal of Artificial Intelligence Research*”, vo.11, pp. 95-130.
- [11] Lin, D. (1998), An information-theoretic definition of similarity, “in *Proceedings of the 15th International Conference on Machine Learning*.” Morgan Kaufmann, San Francisco, CA, pp. 296–304.
- [12] Jiang, J. and Conrath, D. (1997), Semantic similarity based on corpus statistics and lexical taxonomy, “in *Proceedings of International Conference on Research in Computational Linguistics*.” ROCLING X.
- [13] Shen, Y. Shaohong Zhang, Wong, H. (2011), A New Method for Measuring the Semantic Similarity on Gene Ontology,” *2010 IEEE International Conference on Bioinformatics and Biomedicine*”, pp.533-538.
- [14] Chen, Z.Tang, J. (2010), Using Gene Ontology to Enhance Effectiveness of Similarity Measures for Microarray Data, “*IEEE International Conference on Bioinformatics and Biomedicine*,” pp.66-71.
- [15] <http://www-genome.stanford.edu/mec1/data.shtml>, 2009.
- [16] Blake, C. land Merz, C. J (1998). UCI Repository of machine learning databases.
- [17] SGD project, <http://www.yeastgenome.org/>