Applied Soft Computing 55 (2017) 276–288

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Gene selection for designing optimal fuzzy rule base classifier by estimating missing value

Amit Paul [a,*], Jaya Sil [b], Chitrangada Das Mukhopadhyay [c]

[a] Computer Science and Engineering of St. Thomas' College of Engineering and Technology, Kidderpore, Kolkata 700023, India
[b] Computer Science and Technology of Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, India
[c] Centre for Healthcare Science and Technology of Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, India

ARTICLE INFO

Article history:
Received 4 May 2016
Received in revised form 1 December 2016
Accepted 25 January 2017
Available online 9 February 2017

Keywords:
DNA microarray
Fuzzy rule base classifier
Impute missing value
Gene selection
Fuzzy similarity

ABSTRACT

DNA microarray technology, a high throughput technology evaluates the expression of thousands of genes simultaneously under different experimental conditions. Analysis of the gene expression data reveals that not all but few important genes are responsible for the diseases. However, the DNA microarray data set usually contain multiple missing value and therefore, selection of important genes using the incomplete data set may be erroneous, resulting misclassification in disease prediction. In the paper we propose an integrated framework, which first imputes the missing value and then in order to achieve maximum accuracy in classifying the patients a classifier has been designed to select the genes using the complete microarray data set.

Here functionally similar genes are employed to estimate the missing value unlike the existing gene expression value based distance similarity measure. However, the functionally similar genes may differ in their protein production capacity and so the degree of similarity between the genes varies from gene to gene. The problem has been dealt by proposing a novel method to impute the missing value using the concept of fuzzy similarity. After imputing the missing value, the continuous gene expression matrix is discretized using fuzzy sets to distinguish the activation levels of different genes. The proposed fuzzy importance factor (FIf) of each gene represents its activation level or protein production capacity both in the disease and normal class. The importance of each gene is evaluated while optimizing the number of rules in the fuzzy classifier depending on the FIf. The methodology we propose has been demonstrated using nine different cancer data sets and compared with the state of the art methods. Analysis of experimental results reveals that the proposed framework able to classify the diseased and normal patients with improved accuracy.

1. Introduction

A DNA microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions. For example conditions may be time series during a biological process or a collection of different tissue samples (e.g., normal versus cancerous tissues). DNA microarray data set is represented by a real-valued matrix, where the rows form the expression patterns of the genes, columns represent the expression profiles of the samples and each cell is the measured expression level of the gene under certain condition. Though DNA microarray technology simultaneously measures the expression levels of thousands of genes, only a few underlying genes may account for significant data variation, making the difference between the normal human being and the patient. Redundant and irrelevant genes lead to inaccurate classification and add extra burden in finding potentially useful knowledge [1,2] by analyzing the data set. Therefore, gene selection is one of the main sub-fields in bioinformatics research [3–5], which ultimately establishes certain biomarkers.

The aim of gene selection research is to find an optimal gene subset from the set consisting of thousands of genes, in order to improve performance of the classifier in terms of accuracy, precision and sensitivity. However, DNA microarray experiments often generate multiple missing value due to different reasons including dust or scratches on the slide, error in experiments, image corruption and insufficient resolution. Typically 1–50% of the data are missing [6], affecting up to 95% of the genes and becomes a predominant problem in analyzing the DNA microarray data [7,8]. Therefore, estimation of missing value for imputing the data is an essential preprocessing step for knowledge extraction from the

* Corresponding author.
E-mail addresses: amitpaul83@gmail.com (A. Paul), js@cs.iiests.ac.in (J. Sil), chitrangadadas@yahoo.com (C.D. Mukhopadhyay).

http://dx.doi.org/10.1016/j.asoc.2017.01.046
1568-4946/

complete DNA microarray data set. Existing missing value prediction algorithms require a full DNA microarray data set [8] for statistical analysis because the underlying statistical methodology is based on balanced data sets.

In our earlier work, Pattern Similarity Matching (PSM) [9] algorithm was applied for imputing missing value and experimented on the data set "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization". Performance of the PSM algorithm was evaluated using normalize root mean square error (NRMSE). In [10] we presented Optimized Fuzzy Rule Generation (OFRG) algorithm for classification and tested on "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer". Performance of the OFRG algorithm was evaluated based on the classification accuracy.

The aim of the paper is to impute the missing value in the DNA microarray experimental data, then observing the effect of imputed value in gene selection and finally the role of the selected genes in classifying the human as disease or normal. A novel fuzzy similarity metric has been devised to estimate the missing value by considering biological functionality of both the positive and negative correlated genes in the DNA microarray experimental data. Similar biological functional genes are identified based on their expression patterns occur either in the same direction (both increasing or decreasing), known as positively correlated or in the opposite direction (one increasing, another decreasing or vice versa), negatively correlated. In the existing methods [11] missing value is estimated based on the distance between the expression values of the genes, ignoring their biological characteristics. In the paper, functional similarity between the genes has been determined by assigning soft class membership value using the concept of fuzzy similarity measure.

The complete gene expression matrix without missing value is normalized to obtain zero mean distribution of expression values. The normalized values are discretized and represented by five fuzzy sets like; high, very high, low, very low and moderate, depending on the variance of each sample. The fuzzy values are relevant for discriminating alternative phenotypes and represent activation level of the gene. It has been observed that the range of continuous expression value in experimental data while discretized follow the statistical property. The discretized data is represented by five fuzzy values (*very high*, *high*, *moderate*, *low*, and *very low*) sufficient for identifying the variation of original gene expression values between the normal and disease human being. In addition, discretized attributes generate finite rule set unlike the rule set with continuous values, which are unstable even for slight change of training data set.

The comprehensibility of the fuzzy rule-based systems is related to interoperability of each fuzzy set in the rule, separation with the neighboring fuzzy sets and the number of fuzzy levels assigned for each linguistic variable. Fuzzy if-then rule base classifier has been employed in the research for classifying the microarray gene expression data through important gene selection. The rule set in the classifier inducts more number of important genes and we obtain optimum number of gene subset provides maximum classification accuracy. In the proposed method gene selection and classification have been performed in an integrated framework by designing a robust fuzzy rule base classifier. The classifier consisting of fuzzy rules in order to identify the genes having significant importance compare to others in the microarray data set. The importance of each gene is defined by the number of unique fuzzy value, appearing maximum number of times in the disease class of a given data set. However, the same fuzzy value of a particular gene may also be present in the normal class, contributing toward misclassification of the data. Therefore, we calculate degree of importance of each gene considering its influence both in the disease and normal class and defined as the fuzzy importance factor (FIf). In the paper we propose an algorithm to build a fuzzy rule base classifier with optimum number of important genes selected by analyzing the data set. The genes are ranked by FIf and initially, fuzzy rules are framed with higher ranking genes. Gradually next higher ranking genes are included and thus optimum number of fuzzy rules are obtained consisting of optimum number of genes with an objective to achieve maximum classification accuracy.

We compare the proposed Pattern Similarity Matching (PSM) algorithm for missing value estimation with the K-nearest neighbor imputation (KNNimpute) algorithm [11] and SVD imputation algorithm by evaluating Normalize Root Mean Square Error (NRMSE). The results show superiority of the PSM algorithm while imputing missing values in the DNA microarray data set. The proposed fuzzy rule base classifier is compared with the existing classifiers like Naive Bayes', SVM, and C4.5 after employing well known feature (gene) selection algorithms (noFS, CFS, FCBF, INT, IG, ReliefF, SVM-RFE, mRMR). The Optimized Fuzzy Rule Generation (OFRG) algorithm shows large true positive and true negative, least false positive and false negative, therefore effective in disease diagnosis. The results show improvement in accuracy, sensitivity and specificity compare to the existing classifiers while classifying cancerous and normal people.

The rest of the paper is organized as follows. Literature review has been presented in Section 2. The proposed method is described in Sections 3 and 4. Experimental results are summarized in Section 5. Concluding remarks are given in Section 6.

2. Review work

Earlier imputation methods are based on statistical analysis where biological information of the DNA microarray data is ignored. Recent imputation methods on the other hand use information which is mined from the DNA microarray data to customize the analysis [12]. The methods for imputing missing values are generally divided into three categories:

- Pairwise deletion
- Parameter estimation
- Imputation techniques [13].

Acuña and Rodriguez categorized the methods namely, case deletion (CD) [14], mean imputation (MI), median imputation (MDI) and KNN impute [11]. CD method removes the entire row/column having missing values in at least one feature (gene). The missing values are replaced by the mean of the available gene expression values in MI method. Alternatively, median is used in MDI method to assure reliability since the mean is affected by the existence of the outliers. In the KNN impute method, the similarity between two instances is determined using a distance function. Besides these, the missing value imputation techniques are categorized as generic statistical methods and application specific modifications. Mean imputation, hot deck imputation, model based imputation, multiple imputation, and cold deck imputation are considered as generic statistical methods. In application specific modifications, quality issues or experimental designs are taken into account to impute the missing gene expression data. Moreover, existing algorithms are categorized according to the type of information, such as global approach (SVD imputation, Bayesian principal component analysis (BPCA)), local approach (KNNI, LLSI), hybrid approach (LinCmb), and knowledge assisted approach (Projection Onto Convex Sets (POCS)). In the global approach, algorithms search for a global covariance structure in all the genes. Examples of such algorithms include the SVD imputation (SVD impute) [11] and Bayesian principal component analysis