Available online at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Using support vector regression in gene selection and fuzzy rule generation for relapse time prediction of breast cancer

Hamid Mahmoodian *, *Leila Ebrahimian*

Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Esfahan, Iran

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
2 February 2016
Accepted 7 March 2016
Available online 23 March 2016

Keywords:
Support vector regression
Type 2 fuzzy logic
Relapse time

ABSTRACT

Gene expression profiles have been recently used in survival analysis, tumor classification and ER status identification. The prediction of breast cancer recurrence based on gene expression profile has been regarded in some previous studies in which the procedures were based on the concept of regression functions and fuzzy systems. In this study, a method based on the combination of these two concepts is presented; not only a method for gene selection, but also a systematic way to create fuzzy rules are going to be offered. Due to the ability of type-2 fuzzy systems in handling of uncertain systems, the proposed model is developed to type-2. The results show that this model has been improved in comparison to previous ones.

1. Introduction

Gene expression profiles have been recently used in survival analysis, tumor classification and ER status identification [1–7]. In many of these studies, a set of genes has been presented which is able to classify the samples into two categories. In this field, many algorithms have been developed based on a well-known dataset published in [1] to discriminate low and high risk breast cancer patients. In two previous studies [8,9], van't Veer dataset has been used to create a model for relapse time prediction.

In [8], predictor model has been developed based on support vector regression. Different methods of gene selection have been applied to select a reliable set of genes as features of SVR to increase the performance of the model. Actually, in [8], the procedure of gene selection has been totally separated from predictor model. On the other hand, because of a wide range of relapse time of the samples in the dataset (between 0.27 and 13.42 years) and their nonlinear nature, a simple regression model may not be able to accurately predict recurrence.

In [9], a fuzzy TSK model, which is able to describe a nonlinear model, has been used to predict the recurrence. In that study, correlation coefficients between gene expression values and relapse times along with a multi-step fuzzy rule mining have been applied for gene selection and the predictor model respectively. Although the gene selection process was completely separated from the creation of fuzzy rules, the nonlinear range of recurrence times has been divided into

* Corresponding author at: Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Esfahan, Iran.
E-mail addresses: H_mahmoodian@pel.iaun.ac.ir (H. Mahmoodian), ebrahimian_leila@yahoo.com (L. Ebrahimian).

http://dx.doi.org/10.1016/j.bbe.2016.03.003

many linear parts by TSK fuzzy rules. This could be the main reason for the superiority of this method over the mentioned one that was based on SVR.

In this study, to generate a model for predicting the relapse time of breast cancer, a combination method based on recursive feature elimination (RFE) and support vector regression (SVR) has been developed to select the significant genes and simultaneously realize the parameters of TSK fuzzy rules. The strengths of the two previous techniques can be implemented in the proposed model. Since type-2 fuzzy logic (compare to type-1) is usually better in handling of uncertainties in a system, the proposed model has been developed to type-2. Breast cancer dataset published by van't Veer has been applied for training the model. Moreover, 19 extra samples of this dataset and 37 lymph node negative samples of Vijver dataset [10] have been used for model validation.

This paper is organized as follows: in Section 2, mathematical frameworks of SVR, SVR_RFE for gene selection and type-2 fuzzy systems are briefly explained. Proposed model for predicting the relapse time and consequently simulation results are presented in Sections 3 and 4 respectively. The conclusion remarks are finally given in Section 5.

2. Mathematical frameworks

2.1. Support vector regression (SVR)

Support vector regression has been developed based on support vectors introduced by Vapnik [11]. Support vector machine (SVM) is a well-known method for classification which has been also used in gene selection field [12]. Such as SVM, SVR has been applied in various fields such as approximation, prediction and quadratic programming. The idea of SVR is to compute a linear function in a high dimensional feature space (in this work, features are gene expression values of gene expression profiles) where input data might be mapped via a nonlinear function. Similar to support vector machine, SVR is attempts to minimize an error function to increase the performance of regression. Support vectors have been determined from training data which have been used in minimization process.

Suppose $\{(X_1, y_1), \ldots, (X_n, y_n)\} \subset (\Re^m \times \Re)$ are training set where $X_i = [x_{1i}, x_{2i}, \ldots x_{mi}]^T$, (i = 1, 2, …, n) and y_i are denoted as input patterns and targets respectively (in this work, x_j (j = 1, 2, …, m) and y_i are gene expression values and relapse time of ith sample respectively). The goal of SVR is to find a function such as $f(X)$ that has at most ε deviation from the actually obtained targets for all training samples as flat as possible. Function $f(X)$ is in the form $f(X) = w^T X + b$ where b is a constant value, X is a training vector and w is a weighting vector which should be minimized as following:

$$\text{minimize} \frac{\|w\|^2}{2} \quad \text{subject to} \quad \begin{cases} y_i - (w^T X + b) \le \varepsilon \\ (w^T X + b) - y_i \le \varepsilon \end{cases} \tag{2.1}$$

Sometimes, some errors are allowed and minimization has been changed to the following equation:

$$\text{minimize} \frac{\|w\|^2}{2} + C \sum_i (\xi_i + \xi_i^*) \text{ subject to } \begin{cases} y_i - (w^T X + b) \le \varepsilon + \xi_i \\ (w^T X + b) - y_i \le \varepsilon + \xi_i^* \end{cases}$$

where ξ_i and ξ_i^* are nonnegative values, C is a positive value which determines the flatness of $f(x)$ and deviations around ε and i is the number of training samples. Using Lagrange multipliers and minimization procedures [11], yield two following relations:

$$w = \sum_i \beta_i X_i \quad \text{and} \quad f(x) = \sum_i \beta_i (X_i^T \cdot X) + b \tag{2.2}$$

where β_i are non-zero values for some x_i which are termed support vectors. Therefore, w is a linear combination of some training samples.

In this study, weighting vector w has an important role, not only in gene selection, but also in the consequent part (THEN-part) of fuzzy rules used in TSK model.

2.2. Gene selection

Similar to the SVM-RFE (Support Vector Machine-Recursive Elimination Features) method introduced in [13] for wrapper gene selection, in this study, SVR-RFE (Support Vector Regression-Recursive Elimination Features) has been suggested to select a final subset of genes with high performance in relapse time prediction. Based on relation $w = \sum_i \beta_i X_i$, absolute values of vector w can be considered as weighting for the genes involved in the training process. Smaller weights represent less important genes which are recursively removed. Like SVM-RFE, the number of genes deleted in each stage can be different.

2.3. Interval type-2 fuzzy logic system (IT2-FLS)

Type-2 fuzzy logic was originally proposed by Zadeh [14] and has been used in a widespread application in recent years [15–18]. Formally a type-2 fuzzy set $\tilde{A}$ is characterized by a type-2 membership function $\mu_{\tilde{A}}(x, u)$ such as:

$$\tilde{A} = \{((x, u), \mu_{\tilde{A}}(x, u)) \forall x \in X, \quad \forall u \in [0, 1], \mu_{\tilde{A}}(x, u) \in [0, 1]\} \tag{2.3}$$

To reduce the computational cost, interval type-2 fuzzy system which assumes interval membership grades for each type-2 fuzzy set, has been applied in this work [19]. Similar to T2-FLS, the membership functions of IT2-FLS include an uncertainty area, called footprint of uncertainty (FOU). For this reason, a type reducer block has been added to T1-FLS to defuzzify the fuzzy outputs into the crisp value.

IT2-FLS typically has two upper and lower boundaries for membership functions which will create two membership values shown by $\overline{\mu}_{\tilde{A}}(x)$ and $\underline{\mu}_{\tilde{A}}(x)$ respectively (Fig. 1). In Takagi–Sugeno–Kang (TSK) model, rth (r = 1,2,…,R) fuzzy rule is usually described by:

$$IF \ x_1 \text{ is } \mu_{\tilde{A}_1}^{-r} \quad \text{and} \ldots x_n \text{ is } \mu_{\tilde{A}_n}^{-r} \text{ Then } z^r = a_0^r + \sum_{j=1}^{n} a_j^r x_j \tag{2.4}$$

where a_j^r are consequent parameters of rth fuzzy rule which may themselves be uncertain in the range $[\underline{a}_j^r, a_j^{-r}]$. Type reduction and determination of the final output are based on solving the following optimization problem: