Applied Soft Computing 14 (2014) 47–52

Contents lists available at ScienceDirect

Applied Soft Computing

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

Hybrid intelligent modeling schemes for heart disease classification

Yuehjen E. Shao [a], Chia-Ding Hou [a,*], Chih-Chou Chiu [b]

[a] Department of Statistics and Information Science, Fu Jen Catholic University, 510, Chung-Cheng Road, Xinzhuang District, New Taipei City 24205, Taiwan, ROC
[b] Department of Business Management, National Taipei University of Technology, Taipei City 106, Taiwan, ROC

ARTICLE INFO

Article history:
Received 13 March 2013
Received in revised form 17 August 2013
Accepted 24 September 2013
Available online 9 October 2013

Keywords:
Hybrid
Logistic regression
MARS
Artificial neural network
Rough sets
Heart disease

ABSTRACT

Heart disease is the leading cause of death among both men and women in most countries in the world. Thus, people must be mindful of heart disease risk factors. Although genetics play a role, certain lifestyle factors are crucial contributors to heart disease. Traditional approaches use thirteen risk factors or explanatory variables to classify heart disease. Diverging from existing approaches, the present study proposes a new hybrid intelligent modeling scheme to obtain different sets of explanatory variables, and the proposed hybrid models effectively classify heart disease. The proposed hybrid models consist of logistic regression (LR), multivariate adaptive regression splines (MARS), artificial neural network (ANN), and rough set (RS) techniques. The initial stage of the proposed process includes the use of LR, MARS, and RS techniques to reduce the set of explanatory variables. The remaining variables are subsequently used as inputs for the ANN method employed in the second stage. A real heart disease data set was used to demonstrate the development of the proposed hybrid models. The modeling results revealed that the proposed hybrid schemes effectively classify heart disease and outperform the typical, single-stage ANN method.

1. Introduction

The heart can be viewed as the body's engine: it is responsible for pumping life-sustaining blood via a network of vessels. Although most people know that the heart must be properly cared for, heart disease has risen steadily over the last century and has become the leading cause of death for people in the United States [1]. In addition, due to the necessity for the prevention of heart disease, the UK government provides 169 million European dollars to fund research on coronary heart disease [2].

Several studies have been devoted to using some single classification algorithm for the classification of eye diseases [3] and of brain diseases [4,5]. Different from most studies, this paper aims to propose a novel hybrid scheme for the classification of heart diseases. The heart disease data sets used in the present study were real data obtained from a UCI machine learning benchmark repository [6]. Due to its importance to mankind, many studies [7–9] on modeling procedures for heart disease classification have been conducted. Although heart disease data sets have 75 explanatory variables and one dependent variable (i.e., the presence or absence of heart disease), almost every study uses 13 explanatory variables to predict or classify the dependent variable. Although one can use the aforementioned variables to classify heart disease through the use of logistic regression (LR) techniques or machine learning approaches, the true relationship between these measurements and heart disease is not easy to determine. Besides, by using a single technology to address all of the classification problems may not always be possible [10–13]. To overcome the aforementioned limitations and maintain the classification accuracies of existing approaches for heart disease, the purpose of the present study was to determine the classification performance of hybrid modeling schemes that integrate the techniques of logistic regression, multivariate adaptive regression splines (MARS) and rough sets (RS) with artificial neural networks (ANN).

The LR model is a forecasting or classification technique that is widely used in many practical applications. However, the LR model is sometimes criticized for its strong assumptions such as variation homogeneity. As a result, the LR model has limited applications. MARS is typically able to effectively reveal important data patterns and relationships within the complex data structures that often hide in high-dimensional data. In addition to LR and MARS techniques, ANN has become a fruitful alternative in modeling classification problems due to its ability to capture complex nonlinear relationships among variables. Consequently, ANN has superior classification capabilities compared to regression techniques [14–23]. However, ANN is criticized for its long training

* Corresponding author.
E-mail address: stat0002@mail.fju.edu.tw (C.-D. Hou).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.09.020

process in the design of optimal network topologies and difficulties in identifying the relative importance of potential input variables [24,25].

Rough set theory is a new effective classification tool for dealing with vagueness and uncertainty information [26]. Attribute reduction is one of the most important concepts of rough set theory. Irrelevant and redundant attributes are removed from the decision without any a priori information. Meanwhile, the knowledge mined by rough set theory can be expressed and saved as a rule. Due to these advantages, rough set theory has been widely used in many fields [26]. Although rough set theory has been successfully used to find data dependencies and feature subsets in various works, its running time, which is polynomial, is relatively long.

The procedure for hybrid modeling is to initially use LR, MARS and RS techniques to model heart disease data sets. Because there is no theoretical approach for determining the best input variables for an ANN model, LR, MARS and RS can be implemented to determine a good subset of input variables when many potential variables are considered as input vectors of the designed ANN model. The resulting fewer but more significant explanatory variables are used as inputs to the ANN model. In terms of classification capability, the present study compares the traditional single stage of the MARS model, RS model, ANN model, and the proposed hybrid model for heart disease application. The superior classification capability of the proposed hybrid approach is addressed.

The remainder of the study is organized as follows: the methodologies of LR, MARS, RS and ANN are discussed in Section 2, and a literature review is provided, as well. The designed LR, MARS, RS and ANN models are presented in Section 3, and a real heart disease data set is used to verify the proposed and typical models. The final section addresses the research findings and concludes the present study.

2. Methodologies

Heart disease is one of the most serious diseases for human beings; thus, heart disease classification is an important issue. In the present study, we employed the techniques of LR, MARS, RS, and ANN and the proposed hybrid LR–ANN, MARS–ANN and RS–ANN models to classify heart disease. These methodologies are addressed in the following sections.

2.1. Logistic regression

LR is one of the most common statistical methods for modeling real applications. The modeling process involves the setup of relationships between one dependent or response variable and several independent or explanatory variables. The performance of LR is typically acceptable, as long as the required assumptions have been met. However, the assumptions of the LR model (for example, variation homogeneity) often confine its application.

The framework of LR can be simply described as follows. Let Y_i represent dependent variables in case i and let $Y_i = 0$ or $Y_i = 1$, where 0 denotes the absence of heart disease and 1 denotes the presence of heart disease. Let

$$\begin{aligned} P_r[Y_i = 1 \ \ | X_{1i}, X_{2i}, X_{3i}, \ldots, X_{ni}] &= \pi_i \\ P_r[Y_i = 0 \ \ | X_{1i}, X_{2i}, X_{3i}, \ldots, X_{ni}] &= 1 - \pi_i \end{aligned} \tag{1}$$

be the probabilities of $Y_i = 1$ and $Y_i = 0$ under a set of given independent variables $(X_{1i}, X_{2i}, ..., X_{ni})$. Therefore, the logistic regression model has the following form:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_n X_{ni} \tag{2}$$

A collinearity diagnosis should be initially implemented to exclude variables exhibiting high collinearity. Consequently, the remaining variables can be employed for LR modeling. In addition, the Wald forward method was used to recognize explanatory variables with significant influence.

2.2. Multivariate adaptive regression splines

MARS has been generally applied in many fields [25,27–34]. The general MARS function can be represented as follows [34]:

$$\hat{f}(x) = b_0 + \sum_{m=1}^{M} b_m \prod_{k=1}^{K_m} \left[S_{km}(x_{v(k,m)} - t_{km})\right] \tag{3}$$

where b_0 and b_m are parameters, M is the number of basis functions (BF), K_m is the number of knots, S_{km} takes on values of either 1 or −1 and indicates the right or left sense of the associated step function, $v(k, m)$ is the independent variable, and t_{km} is the knot location.

The optimal MARS model was chosen using a two-stage process. First, we set up a large number of basis functions to fit the data. Second, the basis functions with the least contributions were deleted using generalized crossvalidation (GCV) criterion. A measure of variable importance can be obtained by observing the decrease in the calculated GCV values when a variable is removed from the model. The GCV can be expressed as follows:

$$LOF(\hat{f}_M) = GCV(M) = \frac{1/N\sum_{i=1}^{N}[y_i - \hat{f}_M(x_i)]^2}{[1 - (C(M)/N)]^2} \tag{4}$$

2.3. Artificial neural network

In recent years, ANN has been widely applied in engineering, education, social science, medical research, business and forecasting [35–41]. ANN nodes are divided into three layers, including the input, output and hidden layers. The structure of ANN can be briefly described as follows. For each neuron j in the hidden layer and neuron k in the output layer, the net inputs are given as:

$$net_j = \sum_i w_{ji} \times o_i, \quad \text{and} \quad net_k = \sum_j w_{kj} \times o_j, \tag{5}$$

where i (j) is a neuron in the previous layer, o_i (o_j) is the output of node i (j) and w_{ji} (w_{kj}) is the connection weight from neuron i (j) to neuron j (k). The neuron outputs can be described as:

$$o_i = net_i \tag{6}$$

$$o_i = \frac{1}{1 + \exp^{-(net_i + \theta_i)}} = f_i(net_i, \theta_i) \tag{7}$$

$$o_k = \frac{1}{1 + \exp^{-(net_i + \theta_i)}} = f_i(net_k, \theta_k) \tag{8}$$

where net_j (net_k) is the input signal from the external source to node $j(k)$ in the input layer and $\theta_j(\theta_k)$ is a bias. The transformation function shown in Eqs. (7) and (8) is a sigmoid function and is the most commonly utilized function to date. Thus, the aforementioned sigmoid function was used in the present study.

2.4. Rough set

The concept of rough set was introduced in the early 1980s [26,42]. Rough set theory is an extension of the set theory used for the study of intelligent systems characterized by inexact, uncertain or vague information and can serve as a new mathematical tool for soft computing [43]. General elements engaged in rough set theory can be described as outlined in the following section.