Expert Systems with Applications 38 (2011) 5507–5513

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors

Cheng-Ding Chang [a], Chien-Chih Wang [b,*], Bernard C. Jiang [a]

[a] *Department of Industrial Engineering and Management, Yuan Ze University, Chung-Li 320, Taiwan*
[b] *Department of Industrial Engineering and Management, Ming Chi University of Technology, Taipei County 243, Taiwan*

ARTICLE INFO

Keywords:
Health evaluation center
Cardiovascular disease
Multi-feature selection
MARS

ABSTRACT

Many previous studies have employed predictive models for a specific disease, but fail to note that humans often suffer from not only one disease, but associated diseases as well. Because these associated multiple diseases might have reciprocal effects, and abnormalities in physiological indicators can indicate multiple associated diseases, common risk factors can be used to predict the multiple associated diseases. This approach provides a more effective and comprehensive forecasting mechanism for preventive medicine. This paper proposes a two-phase analysis procedure to simultaneously predict hypertension and hyperlipidemia. Firstly, we used six data mining approaches to select the individual risk factors of these two diseases, and then determined the common risk factors using the voting principle. Next, we used the Multivariate Adaptive Regression Splines (MARS) method to construct a multiple predictive model for hypertension and hyperlipidemia. This study uses data from a physical examination center database in Taiwan that includes 2048 subjects. The proposed analysis procedure shows that the common risk factors of hypertension and hyperlipidemia are Systolic Blood Pressure (SBP), Triglycerides, Uric Acid (UA), Glutamate Pyruvate Transaminase (GPT), and gender. The proposed multi-diseases predictor method has a classification accuracy rate of 93.07%. The results of this paper provide an effective and appropriate methodology for simultaneously predicting hypertension and hyperlipidemia.

1. Introduction

According to a World Health Organization (WHO) survey, Cardiovascular Disease (CVD) accounts for nearly one third of all deaths worldwide. Hypertension and hyperlipidemia are both indicators of the metabolic syndrome, and can potentially lead to CVD, cardiopathies, nephrosis, and other diseases (Kannel, 1990). Although many studies have investigated the risk factors of specific diseases and constructed corresponding prediction models, relatively little research considers multiple diseases. However, abnormalities in physiological indicators may indicate not only a single disease, but multiple diseases. Therefore, determining the common risk factors and developing a predictor model for multiple diseases is more important than doing so for only a single disease. For example, a patient with hypertension or hyperlipidemia is more likely to suffer from cardiovascular disease than a normal, healthy individual. Hypertension is also associated with hyperlipidemia (Bonna & Thelle, 1991). The purpose of this paper is to identify the common risk factors of hypertension and hyperlipidemia using data-mining techniques, and then, by applying the Multivariate Adaptive Regression Splines (MARS) method, to construct a predictive model for these two diseases.

In their examination of studies on hypertension and hyperlipidemia in the literature, Staessen, Wang, and Thijs (2001) found that hypertension is the most important risk factor for CVD. The National Library of Medicine defines hypertension as a Systolic Blood Pressure (SBP) value $\geqslant$140 mm Hg and/or a Diastolic Blood Pressure (DBP) value $\geqslant$90 mm Hg. They also reports that the risk factors for hypertension include old age, non-white race, high sodium and total fat intake, family history of hypertension, physical inactivity, excessive alcohol consumption, and smoking. The diabetes care guide of the American College of Physicians (ACP, chap. 10) shows that in many studies, lipid-lowering therapy leads to a 22–24% reduction in major cardiovascular events in patients with type 2 diabetes. Lee and Entzminger (2006) conducted a cross-sectional study of 1398 patients, and found that old age, Body Mass Index (BMI), and low educational attainment are statistically significant risk factors for hypertension. Wu, Lee, Hsu, and Lee (2003) defined hyperlipidemia as serum Total Cholesterol (T-CHO) $\geqslant$ 200 mg/dl, or Low-Density Lipoprotein (LDL) $\geqslant$ 130 mg/dl, or high-density lipoprotein (HDL) $\geqslant$ 200 mg/dl, in combination with either a T-CHO/HDL ratio of >5 or HDL <35 mg/dl. Silverstein et al. (2000) found that age, LDL, triglycerides, and HDL are risk factors of hyperlipidemia. The results of these studies show that the risk

[*] Corresponding author. Tel.: +886 2 29089899x4713; fax: +886 2 2904 1914.
E-mail address: ieccwang@mail.mcut.edu.tw (C.-C. Wang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.10.086

factors of hyperlipidemia are unlike those of hypertension, yet both conditions are causes of cardiovascular disease.

From the viewpoint of preventive medicine, monitoring a subject's risk factors with a predictive model might allow the patient to receive health care advice or early treatment that would prevent disease. Akdag et al. (2006) used the classification-tree method to determine the risk of hypertension among outpatients in a clinic in Denizli province, western Turkey, between January 2002 and July 2004. Their results show that BMI, waist-to-hip ratio, sex, serum triglycerides, serum total cholesterol, hypertension in first degree relatives, and saturated fat consumption are risk factors for hypertension. In their study of liver complaints, Young, So, and Chang (2003) used a growth curve analysis to construct a liver complaint predictor model. The three kinds of predictors in their study had 75.86%, 76.55%, and 78.62% accuracy rates. Armengol, Palaudaries, and Plaza (2001) identified the long term risk factors for diabetes, to predict complications, based on 370 cases. They achieved 100%, 90%, and 72.45% accuracy rates in predicting whether or not patients would suffer from apoplexy, amputation, or myocardial infarction, respectively.

Integration of the literature reveals that hypertension and hyperlipidemia not only cause many diseases, but are also themselves caused by some common risk factors, such as age, T-CHO, and triglycerides. Patients suffering from hyperlipidemia are at a higher risk of developing hypertension. This paper uses a two-stage analysis procedure to analyze a database of 2048 subjects from a physical examination center in Taipei. The first stage uses data mining classifier techniques, including logistic regression analysis, discriminant analysis, and C5.0, CHAID, and Exhaustive CHAID, to separately determine the risk factors of hypertension and hyperlipidemia. The second stage uses the Multivariate Adaptive Regression Splines (MARS) method developed by Friedman (1991) to build a predictive model that can simultaneously predict these two diseases. Most previous studies use risk factors to construct a predictive model for a specific disease. However, this type of model can only be used to predict the likelihood of a subject acquiring one disease. The predictive model proposed in this paper can predict if a subject is at risk of both hypertension and hyperlipidemia.

2. Methodology

This paper proposes a two-stage analysis procedure that uses data mining techniques and mathematical approaches to determine common risk factors and predictive models for hypertension and hyperlipidemia. Stage I first selects the risk factors of hypertension and hyperlipidemia using six data mining approaches: logistic regression analysis, C5.0 decision tree, Classification And Regression Tree (CART), Chi-squared Automatic Interaction Detector (CHAID), exhaustive CHAID, and discriminant analysis. Stage II then uses the MARS approach to construct a predictive model for hypertension and hyperlipidemia based on the common risk factors of these two diseases.

2.1. Common risk factor selection

Although each disease might be caused by different abnormal physiological indicators, physiological indicator abnormality can indicate a number of interrelated diseases. Therefore, this study uses six data mining approaches to individually screen out the key physiological indicators of hypertension and hyperlipidemia. After identifying the risk factors of these two diseases, the proposed approach arranges and compares the risk factors of each disease to determine common risk factors of hypertension and hyperlipidemia. The following section describes the six data mining methodologies used in this study.

2.1.1. Logistic regression

Logistic regression can be used when the target variable is a categorical variable. Examples of logistic regression includes factors such as living or dead, has a disease or does not have a disease, etc. The logistic model formula computes the probability of the selected disease y ($y = 0$ if the subject does not suffer from the disease, otherwise, $y = 1$) as a function of the values of the predictive risk factors. If the subject suffers from this disease, the conditional probability is given by $p(y = 1|\mathbf{X}) = p(\mathbf{X})$, and the logistic model formula takes the following form:

$$y = \log\left[\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{1}$$

where $\mathbf{X} = (x_1, x_2, \ldots, x_k)$ represents the vector of k's risk factors selected by the logistic regression approach. This paper uses the forward stepwise method to increase the significance (p value, <0.05) of the medical test indicators and calculate the correct classification rate. Finally, this study obtains a logistic regression model with the best correct predictive rate, and the indicators in this model are the risk factors of the disease.

2.1.2. C5.0

The C5.0 decision tree originates from the ID3 decision tree. The ID3 decision tree cannot be used for continuous variables, but C5.0 overcomes this limitation and boosts the accuracy rate. Let $\mathbf{X}$ represent the vector of m medical test indicators $\mathbf{X} = (x_1, x_2, \ldots, x_k)$ in the database. Let each x_j be a separator of v subgroups, and let s_{jv} represent the number of subjects of the subgroup v in the medical test indicator x_j. If the medical test indicator x_j selects the test variable, the entropy can then be calculated as

$$E(x_j) = \sum_{r=1}^{v} \frac{s_{1jr} + s_{2jr}}{s} I(s_{1jr}, s_{2jr}). \tag{2}$$

The smaller the entropy is, the higher the purity of the medical test indicator. The information expected to enable the classification of a subject under y_i can be given as:

$$I(s_{1jr}, s_{2jr}) = -\sum_{i=1}^{2} p_{ijr} \log_2(p_{ijr}). \tag{3}$$

Thus, $p_{ijr} = s_{ijr}/|s_{jr}|$ is the probability that the subjects in the subgroup s_j belong to y_i. Using the C5.0 decision tree method makes it possible to reduce impurities during assessment at the branch node. The definition of a reduction in impurities is as follows:

$$Gain(x_j) = I(s_1, s_2) - E(x_j). \tag{4}$$

The study also considers the gain value to decide the best branch nodes for classification, and each node represents the risk factors for the disease.

2.1.3. Classification and regression tree (CART)

The CART uses the Gini ratio to estimate diversity of physical condition when the subjects are divided by a medical test indicator. One of the advantages of CART is that it automatically adjusts the tree model to minimize the effects of the measured impurities and determines the best node for classification. CART can be used to analyze either continuous or discrete data. However, the response variable must be in binary format.

Consider a database that includes N patients and m medical test indicators $X_1, X_2, \ldots, X_m$. Every indicator could divide the patient group into two subgroups (S_i, $i = 1,2$) when the indicator used is a node. Let n_{ij} represent the number of patients in subgroup i at the node X_j. The Gini function is then given as

$$Gini = \sum_{i=1}^{2} P_i(1 - P_i), \tag{5}$$