Expert Systems with Applications 38 (2011) 8970–8977

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A predictive model for cerebrovascular disease using data mining

Duen-Yian Yeh [a], Ching-Hsue Cheng [b,*], Yen-Wen Chen [b]

[a] Department of Information Management, Transworld University, Yunlin, Taiwan

[b] Department of Information Management, National Yunlin University of Science & Technology, Yunlin, Taiwan

ARTICLE INFO

Keywords:
Cerebrovascular disease
Data mining
Decision tree
Predictive model

ABSTRACT

Cerebrovascular disease has been ranked the second or third of top 10 death causes in Taiwan and has caused about 13,000 people death every year since 1986. Once cerebrovascular disease occurs, it not only leads to huge cost of medical care, but even death. All developed countries in the world put cerebrovascular disease prevention and treatment in high priority, and invested considerable budget and human resource in long-term studies, in order to reduce the heavy burden. As the pathogenesis of cerebrovascular disease is complex and variable, it is hard to make accurate diagnosis in advance. However, in perspective of preventive medicine, it is necessary to build a predictive model to enhance the accurate diagnosis of cerebrovascular disease. Therefore, coupled with the 2007 cerebrovascular disease prevention and treatment program of a regional teaching hospital in Taiwan, this study aimed to apply the classification technology to construct an optimum cerebrovascular disease predictive model. From this predictive model, cerebrovascular disease classification rules were extracted and used to improve the diagnosis and prediction of cerebrovascular disease.

This study acquired 493 valid samples from this cerebrovascular disease prevention and treatment program, and adopted three classification algorithms, decision tree, Bayesian classifier and back propagation neural network, to construct classification models, respectively. After analyzing and comparing classification efficiencies – sensitivity and accuracy, the decision tree constructed model was chosen as the optimum predictive model for cerebrovascular disease. In this model, the sensitivity and accuracy were 99.48% and 99.59%, respectively, and eight important influence factors of predicting cerebrovascular disease and 16 diagnosis classification rules were extracted. Five experienced cerebrovascular doctors assessed these rules, and confirmed them to be useful to the current clinical medical condition.

1. Introduction

Cerebrovascular disease is a disease threatening human health seriously; it has "four-high" features: high prevalence, high fatality rate, high disability rate and high recurrence rate. In Taiwan, cerebrovascular disease has been ranked the second or third place of top 10 death causes since 1986. For example, among the top 10 death causes in 2007 published by the Department of Health in October 2008 (shown in Table 1), cerebrovascular disease ranked the third; 12,875 people died from it in that year (DOH, 2007). Cerebrovascular disease not only leads to high medical care expenditure, but also a heavy burden of mid-to-long term medical care expenditure and cost on families and communities. In light of this, all advanced countries in the world listed cerebrovascular disease prevention and treatment at high priority in health medical care, and invested considerable budget and human resources into cerebrovascular disease research and education, so as to lower its morbidity rate, fatality rate and sequela, as well as its burden on individuals, families, communities and countries (Pogue, Ellis, Michel, & Francis, 1996).

Elderly population is vulnerable to cerebrovascular disease. As early as in 1993, Taiwan had been concluded by the world health organization (WHO) as an ageing society. Thus, how to discover and prevent cerebrovascular disease as early as possible has become a critical issue for Taiwan. As the pathogenesis of cerebrovascular disease is complex and variable, doctors need to rely on profound medicine knowledge and rich clinical experience to predict the probability of patient contracting cerebrovascular disease. On the other hand, in clinical practice, cerebrovascular disease occurrence is so abrupt and fierce that it is hard to make early and accurate diagnosis and prediction beforehand. Hence, in perspective of preventive medicine, it is indeed necessary to build a predictive model to help doctors diagnosing cerebrovascular disease accurately, so as to improve the treatment quality and contribute to cerebrovascular disease prevention and treatment.

Along with great progress of information technology, computer can search for large amounts of data; the technology of detecting relation and knowledge from data is called data mining. Its main

* Corresponding author. Address: 123 University Road, Section 3, Douliou, Yunlin, Taiwan.

E-mail address: chcheng@mis4k.mis.yuntech.edu.tw (C.-H. Cheng).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.114

Table 1
Taiwan top 10 death causes in 2007.

Ranking	Cause of death	Death No.
1	Malignant neoplasm	40,360
2	Heart disease	13,003
3	Cerebrovascular disease	12,875
4	Diabetes mellitus	10,231
5	Accidents and adverse effects	7130
6	Pneumonia	5895
7	Chronic liver disease and cirrhosis	5160
8	Nephritis, nephritic syndrome and nephrosis	5099
9	Suicide	3933
10	Hypertensive disease	1977

use and purpose are defined: seeking unknown, effective and feasible rule or knowledge from large amounts of data. In business, data mining has been applied to extract decision-making procedure or rule from history data stored in information system, to assist company in improving decision quality and enhancing competitiveness (Berry & Linoff, 1997; Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1997). With the development of data mining technology, it is not only extensively applied in commercial purposes, but also successfully applied in many medical tasks, for examples, in intensive care medicine analysis (Ganzert & Guttmann, 2002), time dependency patterns mining in clinical pathways (Lin, Chou, & Chen, 2001), breast cancer screening (Ronco, 1999), diagnosis of ischaemic heart disease (Kukar, Kononenko, & Groselj, 1999).

In Taiwan, cerebrovascular disease prevention and treatment has been listed in annual health medical care priorities, and the Department of Health allocates high budgets for cerebrovascular disease studies every year. In 2007, a regional teaching hospital in the central and southern of Taiwan implemented the cerebrovascular disease prevention and treatment program, which targeted residents in central and southern Taiwan. This program aimed at obtaining data on the patients, including their physical exam results, blood test results and diagnosis data. Then, these data were gathered, stored, and analyzed to contribute to the prevention and treatment of cerebrovascular disease. Therefore, the purpose of this study was to coordinate with the 2007 cerebrovascular disease prevention and treatment program of this regional teaching hospital and construct an optimum cerebrovascular disease predictive model. This study utilized case data of this program and employed classification techniques in data mining, such as decision tree, Bayesian classifier and back propagation neural network, to construct three classification models. After analyzing and comparing classification efficiency, the model with the highest efficiency was chosen as the optimum predictive model, and diagnosis classification rules would be extracted from it. The results would be evaluated by professional cerebrovascular doctors and confirmed to be effective and accurate in diagnosing and predicting cerebrovascular disease.

This study acquired 493 valid samples from the prevention and treatment program database. The data of the patients, their physical exam results, blood test results, and diagnoses, were divided into three attribute input modes, T_1, T_2 and T_3, in order to construct the classification models and analyze and compare the classification efficiency. After 10-fold cross-validation, the decision tree in T_1 attribute input mode was found to construct a classification model with stable classification efficiency, and thus, chosen as the optimum classification algorithm of this study. The constructed optimum cerebrovascular disease predictive model has 99.48% upon sensitivity and 99.59% upon accuracy, and from this predictive model, 8 important factors of predicting cerebrovascular disease were selected, and 16 diagnosis classification rules were extracted. The results were confirmed by five cerebrovascular doctors, and were conformable with the current clinical medical condition and had reference value.

2. Literature review

2.1. Cerebrovascular disease

Cerebrovascular disease is a type of pathological change in brain blood vessels, and a general artery sclerosis complication. Cerebrovascular sclerosis leads to vascular stenosis or accidental peel off of atherosclerotic plaque, and further blocks remote brain blood vessels, and even leads to cerebrovascular embolism, infarction, or cerebrovascular break and hemorrhage. In fact, patients of cerebrovascular disease usually have other chronic diseases to variable extents, such as hypertension, stenocardia, hyperlipaemia, hyperuricemia, diabetes mellitus, and obesity. In addition, cerebrovascular disease and cardiovascular disease interact as both cause and effect. A clinical research titled "REACH Registry" traced over 67,800 high-risk sclerotic arterial thrombosis outpatients in 44 countries for a year, it was found that 40% of cerebrovascular disease patients had cardiovascular or peripheral vascular embolism; 25% of coronary artery patients had cerebrovascular embolism or peripheral arterial embolism (Carmen, 2007). Moreover, if the patient has heart-related disease, such as coronary sclerosis, ventricular fibrillation arrhythmia, or valvular heart disease, the condition is also easily complicated by cerebrovascular disease, such as cerebral embolism or cerebral infarction.

2.2. Risk factors of cerebrovascular disease

As mentioned above, cerebrovascular disease risk factors can be divided into major risk factors: elder age, hypertension, heart disease, diabetes mellitus, temporal cerebral ischemia seizure and cerebrovascular disease history, and subordinate risk factors: hyperlipaemia, obesity, polycythemia, smoking, drinking, family heredity, oral contraceptive and other medicine.

2.3. Major diseases related with cerebrovascular disease

2.3.1. Diabetes mellitus

Diabetes mellitus is a kind of systemic metabolism disorder. It is because internal insulin excretion insufficiency or dysfunction causes metabolic disorder of nutrients, such as carbohydrates, protein and fat, leading to excessive glucose in blood which is then discharged out of body with urine through kidney, resulted in sugar in urine. Diabetes mellitus diagnosis depends primarily on glucose density in the blood, fasting blood-glucose of normal adult is 70–110 mg/dl, and the blood glucose in 2 h after meal shall be less than 140 g/dl (Chimei, 2007). Mortality of insulin-dependent diabetes mellitus patients is about six times of male at the same age, and 10 times of female at the same age (Science News, 1990). Obesity stems from excessive visceral fat buildup, thus it is positively correlated with diabetes mellitus, and atherosclerosis. In addition, high BMI, WHR, and waist circumference have strong correlation with diabetes mellitus occurrence rate, and those with high BMI and WHR are vulnerable to diabetes mellitus (Faster, 1983; Pouliot et al., 1994).

2.3.2. Cardiovascular disease

2.3.2.1. Heart-related diseases. Heart disease is defined as insufficient blood flow due to vascular tissue anomaly or occlusion, and the heart cannot receive enough oxygen. Thus, some cardiac muscles will lack oxygen or die, impacting on general functioning. Clinical symptoms due to undersupply of cardiac muscles include: arrhythmia, angina, cardiovascular occlusion, heart failure and