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

Table 2
Density ranges of adult cholesterol and triglyceride in different conditions.

	Normal density	Marginal density of high-hazard	High-hazard density
Total cholesterol (non-fast)	<200 mg/dl	200–239 mg/dl	⩾40 mg/dl
Low density cholesterol (12 h fast)	<130 mg/dl	130–159 mg/dl	⩾160 mg/dl
Triglyceride (12 h fast)	<200 mg/dl	200–400 mg/dl	>400 mg/dl

asymptomatic sudden death; clinical symptoms include: chest distress, pressure on left front chest, thoracalgia, dyspnea or feeling dyspepsia, heart palpitation, cold sweat, complicated with dizziness, asthenia.

2.3.2.2. Hypertension. Blood pressure refers to pressure acting on unit area of blood vessel wall when blood flows pass, normally divided into systolic pressure and diastolic pressure. Normal systolic pressure is less than 140 mmHg, diastolic pressure is less than 90 mmHg. According to WHO standard, hypertension systolic pressure is higher than 160 mmHg and diastolic pressure is greater than 95 mmHg. If systolic pressure ranges between 140 and 160 mmHg and diastolic pressure is between 90–95 mmHg, then it is marginal hypertension. According to 1996 the US "Hypertension prevention, detection and treatment criteria," systolic pressure averages at 140 mmHg or diastolic pressure is over 90 mmHg Arauz-Pacheco, Parrott, & Raskin, 2002).

2.3.2.3. Hyperlipaemia. Hyperlipaemia refers to cholesterol and triglyceride in blood exceeding normal range. Dyslipidemia is the main cause of atherosclerosis (high hypercholestrolemia, high triglyceridehemia or combination of both), adding probability of coronary artery and heart diseases. The Department of Health formulated dyslipidemia classification for our country, as shown in Table 2 (Chimei, 2007).

2.4. Data mining technology

Data mining classification technology contains two parts: construction of classification model, and evaluation of model classification efficiency. In the first part, the adopted classification algorithm is trained by a classified training data set in order to build classification predictive model. In the second part, testing data set is used to test classification efficiency of this model. Every data in training data set or testing data set contains different number of attributes and a target class. This study employed 10-fold cross-validation in classification model construction and efficiency evaluation, and the classification algorithms adopted decision tree, Bayesian classifier and back propagation neural network.

2.4.1. Decision tree algorithm

Decision tree is a kind of classifying and predicting data mining technology, belonging to inductive learning and supervised knowledge mining technology. As decision tree is advantageous in fast construction and generating easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classification methods (Cabena et al., 1997; Kennedy, Lee, Roy, Reed, & Lippman, 1997).

Decision tree is a kind of tree diagram based method, the node on the top of its tree structure is root node, nodes in the bottom are leaf nodes, and one target class attribute is given to each leaf node. From root node to every leaf node, there is a path made of multiple internal nodes with attributes. This path generates rule required for classifying unknown data. Moreover, most of decision tree algorithms contain two-stage task, i.e., tree building and tree pruning. In tree building stage, a decision tree algorithm can use its unique approach (function) to select the best attribute, so as to split training data set. The final situation of this stage will be that data contained in the split training subset belong to only one certain target class. Recursion and repetition upon attribute selecting and set splitting will fulfill the construction of decision tree root node and internal nodes. On the other hand, some special data in training data set may lead to improper branch on decision tree structure, which is called overfitting. Therefore, after building a decision tree, it has to be pruned to remove improper branches, so as to enhance decision tree model accuracy in predicting new data (Quinlan, 1986; Witten & Frank, 2000).

Among developed decision tree algorithms, the commonly used ones include ID3 (Maher & Clair, 1993), C4.5 (Breiman, Friedman, Olshen, & Stone, 1984), CART (Kass, 1980) and CHAID (Quinlan, 1993). C4.5 was developed from ID3 (Iterative Dichotomiser 3) algorithm, it uses information theory and inductive learning method to construct decision tree. C4.5 improves ID3, which cannot process continuous numeric problem. CHAID algorithm is featured in using chi-square test to calculate p-value of node category in every splitting, so as to determine whether to allow decision tree to grow without pruning. CHAID cannot process continuous data, so it is not applicable to many medical issues with continuous numeric data. CART algorithm is a binary splitting method, applied in data whose attributes are continuous. Gini index is used to evaluate data discretion as basis of choosing splitting condition. Since this study is to process medical data with multiple attributes, C4.5 is chosen as the decision tree algorithm.

The decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis (Chang & Chen, 2009), predicting essential hypertension (Ture, Kurt, Kurum, & Ozdamar, 2005), and predicting cardiovascular disease (Eom, Kim, & Zhang, 2008).

2.4.2. Bayesian classifier

Theory of Bayesian classifier stems from Bayesian theorem in statistics, while presetting a hypothesis, i.e., every attribute is independent, so that the classifier can be simple and fast. According to Bayesian theorem, the probability of a set of data x_t belonging to c is:

$$P(C|X_t) = \frac{p(C)p(X_t|C)}{p(X_t)}$$

Based on above formula, Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression, it has the excellent interpretability same as decision tree, and is able to use previous data to build analysis model for future prediction or classification (Loether & McTavish, 1993). If the eigenvalues of data are continuous, there are two ways to process (Vapnik, 1982):

1. Suppose it be normal distribution and find (means, variances) of eigenvalues as likelihood.
2. Use splitting method to transfer continuous data into discrete data.

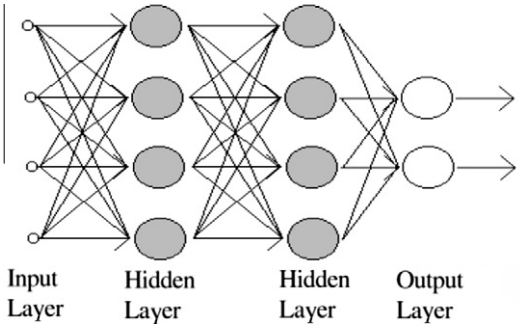

Fig. 1. Structure of multilayer neural network.

The Bayesian classifier has been applied in many medical issues, for examples, in measuring quality of care in psychiatric emergencies (Gustafson, Sainfort, Johnson, & Sateia, 1993), assisting diagnosis of breast cancer (Wang, Zheng, Good, King, & Chang, 1999) and medical cases (Kononenko, 1993).

2.4.3. Back propagation neural network (BPNN)

In artificial neural network, multilayer framework (see Fig. 1) is to enhance the capability of processing nonlinear problems. Multilayer perceptron, used in design of artificial neural network, is a three-layer framework, which comprising: input layer, one layer or above hidden layer and output layer. BPNN model is the most representative and mostly widely applied in current artificial neural network models (Cabena et al., 1997). During its operation, input layer employs linear transfer function, while hidden layer and output layer use nonlinear transfer function, among which, the most often applied transfer function is the sigmoid function.

The fundamental of BPNN model is to take advantage of the gradient steepest descent method to minimize error function. Its learning process is usually carried out in one training example at a time, until all training examples are learned. It is confirmed as one learning epoch. One network can learn from training example repeatedly, until network learning converges. BPNN belongs to supervised learning network, applicable to diagnosis, prediction problems. Back propagation neural network operation flow is as follows:

Step 1: set transfer function and network parameters (learning rate, inertia factor).
Step 2: use even random number to set network initial weight and initial bias.
Step 3: input training samples and target output.
Step 4: calculate output of every neuron in hidden layer, and estimated output of every neuron in output layer.
Step 5: calculate difference between output layer and hidden layer.
Step 6: calculate weight and bias correction of each layer.
Step 7: update weight and bias of each layer.
Step 8: repeat step 3 through step 7, till network convergence.

Its applications in medical issues could be found in literatures, for examples, García-Pérez, Violante, and Cervantes-Pérez (1998) conducted differential diagnosis of Alzheimer disease and vascular dementia, Übeyli and Güler (2003) analyzed internal carotid arterial Doppler signals, and Karabatak and Ince (2009) detected breast cancer.

2.4.4. Tenfold cross-validation method

According to original category data ratio, split experimental data set randomly into 10 equal data subsets, take nine data subsets for training data set, and the rest data subset as testing data set; repeat 10 times, allow every data subset to act as testing data in turn, and use average of results of 10 tests to evaluate predictive model efficiency.

3. Design of predictive model

The construction of predictive model was coupled with the 2007 cerebrovascular disease prevention and treatment program of a regional teaching hospital in central and southern Taiwan. The design is shown in the following.

3.1. Three-stage data mining framework

This study finished the construction of cerebrovascular disease predictive model in three stages (Berry & Linoff, 1997; Cabena et al., 1997; Kennedy et al., 1997):

Stage 1: Data pre-processing and attribute screening.
Stage 2: Classification model construction by decision tree, Bayesian classifier and back propagation neural network.
Stage 3: Classification efficiency comparison, optimum predictive model determination, and diagnosis classification rules extraction.

3.2. Data mining procedure

Step 1: Data collection and variable screening

Four hundred and ninety-three valid samples were acquired from the program. The physical exam results, blood test results, and diagnoses of the samples were collected for classification study. The physical exam data contained 24 variables, blood test data contained 30 variables, and diagnosis data contained 10 vari-

Table 3
Collection of the classification codes of 29 attributes

Diagnosis results		Blood test results		Physical exam results	
Disease name	Classification code	Item name	Classification code	Item name	Classification code
Diabetes mellitus	dm(D)	Heartbeat number	hr_la	Age	age
Hypertension	hp(B)	HbA1c	hba1c_1	Gender	sex
Immunoadsorption	imm(B)	Blood urea nitrogen	bun_1	Blood type	btype
Hyperlipemia	lip(B)	Cholesterol	cho_1	Marital condition	mar
Ischemic heart disease	hd(H)	Triglyeride	tri_1	Smoking habit	smoke
Myocardial infarction	mi(H)	Fasting glucose	glu_1	Drinking habit	drink
Arrhythmia	arr(H)	High density cholesterol	hdl_1	Body mass index	bmi_1
Cardiogenic shock	car(H)	Low density cholesterol	ldl_1	Percentage body fat	bfat_1
		Coagulation disorders	pt_1	Waist-and-hip ratio	wb_1
		Partial coagulation disorders	aptt_la		
		Albumin	alb_1		
		Urine Acid	ua_1		

Table 4
Normal ranges of blood test numeric data.

Classification code	Normal range	Classification code	Normal range
hr_1a	60–80	tri_1	50–150
hba1c_1	4–6	glu_1	70–110
bun_1	8–20	hdl_1	40–70
cho_1	0–200	ldl_1	0–130
pt_1	8–12	aptt_la	23.9–34.9
alb_1	3.7–5.2	ua_1	3.5–8.0

Table 5
An example of confusion matrix.

Predicted class real class	N	CD	BH	DM	SM	Total
N	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{R1}
CD	M_{21}	M_{22}	M_{23}	M_{24}	M_{25}	M_{R2}
BH	M_{31}	M_{32}	M_{33}	M_{34}	M_{35}	M_{R3}
DM	M_{41}	M_{42}	M_{43}	M_{44}	M_{45}	M_{R4}
SM	M_{51}	M_{52}	M_{53}	M_{54}	M_{55}	M_{R5}
Total	M_{P1}	M_{P2}	M_{P3}	M_{P4}	M_{P5}	M

ables. Based on literature review and confirmation by the doctors of this hospital, 8 factors of cerebrovascular related disease, 12 major attributes of blood test and 9 major attributes of physical exam were selected. Classification code of each attribute is listed in Table 3.

Step 2: Attribute symbolization

A. Disease diagnosis data with one certain disease were expressed with symbol "Y", or "N" if having no such disease. Class code of data having only cerebrovascular disease but no any other disease was CD; seven kinds of diseases, hypertension, rheumatic immune blood disease, hyperliaemia, stenocardia, arrhythmia, cardiac asthenia and myocardial infarction were classified as cardiovascular disease, if having one of these diseases, the class code was BH; if having diabetes mellitus, class code was DM; if having two kinds of cardiovascular diseases or diabetes mellitus or above, class code was SM.
B. In part of blood test numeric data, according to preset standard of inspection instrument, the data below normal range is expressed with symbol "L", the data within normal range is expressed with symbol "N", and the data above normal range is expressed with symbol "H". The normal ranges are shown in Table 4.

Step 3: Input data splitting

The attribute data was listed in Table 3. The input form of attribute data was divided into three modes, T_1, T_2 and T_3, representing as follows:

T_1: disease diagnosis data + physical examination data + blood test numeric data;

T_2: disease diagnosis data + physical examination data;
T_3: disease diagnosis data + blood test numeric data.

Steps 4–6: Using three classification algorithms to construct cerebrovascular disease classification models, respectively

The 10-fold cross-validation method was used to input training data set and testing data set into decision tree, Bayesian classifier and back propagation neural network algorithms, respectively, to construct classification model and evaluate classification efficiency. In each attribute input mode, the same training and testing processes were repeated 10 times. Finally, the average classification efficiency of each classification model was calculated.

In this study, decision tree adopted C4.5 algorithm. In BPNN, number of hidden neurons was set at 5, 10, 20, ... until 100 to find the number of hidden neurons with the best accuracy, and learning rate and momentum coefficient were set as 0.1, 0.5 and 0.9, respectively. Number of trainings was set as 1000, 2000 3000, ..., until artificial neural network converges.

Step 7: Classification efficiency analysis and comparison

The classification efficiency averages of three classification models were compared and analyzed to decide the optimum cerebrovascular disease predictive model, which has the best classification efficiency. Then, the best attribute input mode was also confirmed. Two classification efficiency indicators commonly used in medical field, sensitivity and accuracy, were employed in this study to evaluate efficiency of classification model. Based on the classification confusion matrix in Table 5, sensitivity and accuracy are defined by:

$$Sensitivity = M_{ii}/M_{Ri}, \quad i = 1, 2, 3, 4, 5$$

$$Accuracy = (M_{11} + M_{22} + M_{33} + M_{44} + M_{55})/M,$$

where M_{ii} (i = 1, 2, 3, 4, 5) denotes the number of real classes with diseases accurately determined by classification model, e.g., M_{22} denotes the number of persons having diabetes mellitus (DM) that can be recognized to have diabetes mellitus; M_{ij} (i, j = 1, 2, 3, 4, 5, $i \neq j$) denotes the number of cases whose actual disease class differs from recognized disease class, e.g., M_{34} denotes the number of persons who have a kind of cardiovascular disease(BH) but were recognized to have diabetes mellitus (DM); in addition, M_{R1} denotes the total of persons having no any disease (N); M_{R2} denotes total of persons having CD coded disease; M_{R3} denotes total of persons having BH coded disease; M_{R4} denotes total of persons having DM coded disease; M_{R5} denotes total of persons having SM coded disease; finally, M denotes total of persons of studied samples.

Step 8: Extraction of diagnosis classification rules

From the optimum predictive model, diagnosis classification rules were extracted, and confirmed by cerebrovascular doctors to be effective in diagnosis and prediction of cerebrovascular disease.

4. Empirical study and analysis

This study adopted Weka-toolkit as the work platform. After conducting the procedures shown above, all the results are presented in the following.

4.1. Basic data of studies samples

Among disease diagnosis data, data related with studied samples with diseases are detailed in Table 6.

4.2. Classification efficiency comparison

Table 7 shows classification efficiency indicator values of classification models constructed with three algorithms under three attribute input modes. In Table 7, sensitivity value is the average of four sub-indicators (i = 1, 2, 3, 4).

As shown in Table 7, after excluding T_3 attribute input mode, whether in perspective of sensitivity or accuracy indicator, both classification models built with decision tree and BPNN had good classification efficiency. The resulted classification efficiency values were 94.68%, 98.01% and 93.80%, 97.87%, respectively, which were significantly better than Bayesian classifier classification efficiency ($p < 0.05$). In terms of accuracy indicator, decision tree and

Table 6
Sample number of each disease class code

Disease class code	N	CD	BH	DM	SM	Total
Sample number	109	175	43	70	96	493

BPNN had considerable classification efficiency; while in terms of sensitivity indicator, decision tree generated better classification efficiency than BPNN. Moreover, in terms of standard deviation, decision tree had the smallest standard deviation. In all, decision tree could construct classification model with stable classification efficiency. Hence, decision tree is the optimum classification algorithm in this study.

4.3. The optimum attribute input mode

If taking sensitivity indicator and standard deviation for criteria, as shown in Table 7, T_1 is the best attribute input mode for this study. In other words, to have the optimum classification efficiency, all the 29 items, physical exam results, blood test results, and diagnoses, are listed as major input attributes.

4.4. The optimum cerebrovascular disease predictive model

Taking decision tree for classification algorithm and T_1 for attribute input mode, in the process of 10-fold cross-validation, the model with the best classification efficiency was chosen as the optimum cerebrovascular disease predictive model. It has the features as follows.

4.4.1. Classification efficiency

This optimum predictive model has five sensitivity indicators, 100%, 100%, 100%, 100%, 97.92%, respectively, and accuracy indicator = 99.59%.

4.4.2. Tree diagram of the optimum predictive model

The tree diagram is shown in Fig. 2. In this tree, 18 leaf nodes and 35 paths are contained.

4.4.3. Major influence factors

According to Fig. 2, in decision tree constructed optimum cerebrovascular disease predictive model, eight major factors that will influence the diagnosis of cerebrovascular disease can be obtained, as shown in Table 8. They are diabetes mellitus, hypertension, myocardial infarction, cardiogenic shock, hyperlipemia, arrhythmia, ischemic heart disease and body mass index.

4.4.4. Confusion matrix

The confusion matrix of the optimum predictive model is shown in Table 9. Obviously, two cases of SM class were misidentified as BH class and DM class, respectively.

Table 7
Classification efficiency comparisons of the three models

	T_1		T_2		T_3	
	Sensitivity (SD)	Accuracy (SD)	Sensitivity(SD)	Accuracy (SD)	Sensitivity (SD)	Accuracy (SD)
Decision tree	95.29% (1.81%)	98.01% (1.61%)	94.68% (2.01%)	98.01% (1.61%)	62.81% (5.54%)	66.93% (5.31%)
Bayesian classifier	87.10% (3.61%)	91.30% (3.46%)	86.30% (3.75%)	91.36% (3.52%)	66.60% (4.40%)	71.83% (4.36%)
BPNN	94.82% (2.57%)	97.87% (2.41%)	93.80% (2.85%)	98.05% (2.40%)	64.21% (5.62%)	69.32% (5.41%)

SD stands for standard deviation.

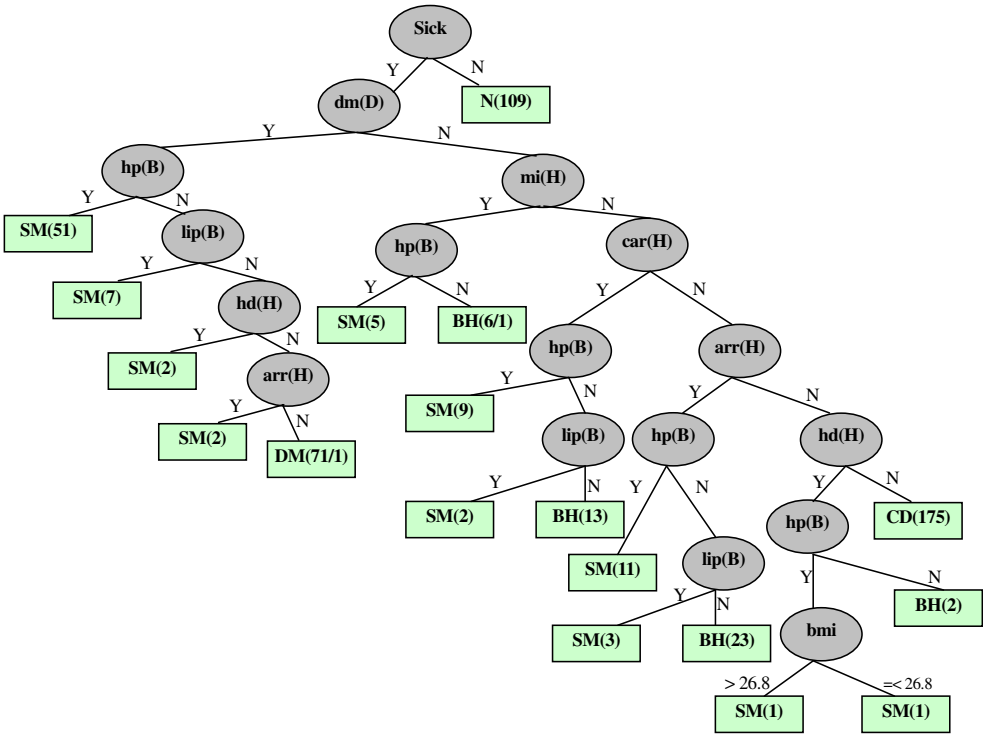

Fig. 2. Tree diagram of the optimum cerebrovascular disease predictive model.

Table 8
Classified orders of all nodes on the optimum predictive model.

Order	Classification code	Attribute name
1	dm(D)	Diabetes mellitus
2	hp(B)	Hypertension
2	mi(H)	Myocardial infarction
4	car(H)	Cardiogenic shock
4	lip(B)	Hyperlipemia
6	arr(H)	Arrhythmia
6	hd(H)	Ischemic heart disease
8	bmi	Body mass index

Table 9
The confusion matrix of the optimum predictive model.

Predicted class real class	N	CD	BH	DM	SM	Total
N	109	0	0	0	0	109
CD	0	175	0	0	0	175
BH	0	0	43	0	0	43
DM	0	0	0	70	0	70
SM	0	0	1	1	94	96
Total	109	175	44	71	94	493

4.5. Extraction of diagnosis classification rules

In this study, class codes of patients having cerebrovascular disease and other related disease simultaneously were DM, BH and SM. Based on the three classes, 16 diagnosis classification rules could be extracted from this optimum predictive model, as shown in Table 10. The decimal numbers in the most right column are precision rates of this optimum predictive model.

Diagnosis classification rules are interpreted as following examples:

[Rule 1]: Patients having diabetes mellitus (dm(D)) could be induced by this optimum predictive model that 71 persons have cerebrovascular disease. Precision rate reaches 0.9859.
[Rule 6]: Patients having diabetes mellitus (dm(D)) and hypertension (hp(B)) simultaneously could be induced by this optimum predictive model that 51 persons have cerebrovascular disease. Precision rate reaches 1.0000.

The same interpretation could apply to the other 14 rules. Finally, as discussed and confirmed by five cerebrovascular doctors, these 16 diagnosis classification rules comply with current clinical medical condition and have reference value in the diagnosis and prediction of cerebrovascular disease.

5. Conclusions

This study carried out along with the 2007 cerebrovascular disease prevention and treatment program of a regional teaching hospital in Taiwan, and used data mining technology to construct an optimum cerebrovascular disease predictive model. A total of 493 valid sample patients were acquired from this prevention and treatment program database, the data on the patients were collected for classification study, which included their physical exam results, blood test results, and diagnoses. Data mining technologies adopted in this study were decision tree, Bayesian classifier and back propagation neural network.

In comparison of data mining technology, this study used sensitivity and accuracy indicators to evaluate classification efficiency of different algorithms. Over all, in T_1 mode, decision tree's sensitivity and accuracy were 95.29%, 98.01%, respectively; Bayesian classifier's sensitivity and accuracy were 87.10%, 91.30%, respectively; BPNN sensitivity and accuracy were 94.82%, 97.87%, respectively. Decision tree had comparable classification efficiency to BPNN. After comparing standard deviation, with the more stable classification efficiency, decision tree was the best classification algorithm in this study. This result is similar to those of Lim, Loh, and Shih (1997) and Liu, Bowyer, and Hall (2004). On the other hand, among T_1, T_2 and T_1 attribute input modes, T_1 is the best attribute input mode in this study, which contained physical exam results, blood test results, and diagnosis data on the patients, 29 major attributes in total.

The optimum cerebrovascular disease predictive model obtained in this study adopts decision tree as classification algorithm, T_1 as attribute input mode, and its classification efficiency: sensitivity indicator = 99.48% and accuracy indicator = 99.59%. Eight major influence factors, diabetes mellitus, hypertension, myocardial infarction, cardiogenic shock, hyperlipemia, arrhythmiaischemic heart disease and body mass index, were recognized for accurately predicting cerebrovascular disease. In addition, 16 diagnosis classification rules were extracted from this predictive model, and confirmed by five cerebrovascular doctors to be conformable with current clinical medical condition and have reference value in diagnosis and prediction of cerebrovascular disease.

Table 10
The 16 diagnosis classification rules of the optimum predictive model.

Serial No. of rule	dm(D)	mi(H)	hp(B)	car(H)	lip(B)	arr(H)	hd(H)	bmi	Prediction results
1	Y								DM(71), 0.9859
2		Y							BH(6), 0.8333
3				Y					BH(13), 1.0000
4						Y			BH(23), 1.0000
5							Y		BH(2), 1.0000
6	Y		Y						SM(51), 1.0000
7	Y				Y				SM(7), 1.0000
8	Y						Y		SM(2), 1.0000
9	Y					Y			SM(2), 1.0000
10		Y	Y						SM(5), 1.0000
11			Y	Y					SM(9), 1.0000
12				Y	Y				SM(2), 1.0000
13			Y			Y			SM(11), 1.0000
14					Y	Y			SM(3), 1.0000
15			Y				Y	⩽26.8	SM(1), 1.0000
16			Y				Y	>26.8	SM(1), 1.0000

References

Arauz-Pacheco, C., Parrott, M., & Raskin, P. (2002). The treatment of hypertension in adult patients with diabetes. *Diabetes Care, 25*, 134–147.

Berry, M. J. A., & Linoff, G. (1997). *Data mining techniques: For marketing sales and customer support*. New Jersey: John Wiley & Sons Inc.. pp. 286–334.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1997). *Discovering data mining: From concept to implementation*. New Jersey: Prentice Hall.

Carmen, S. (2007). Baseline characteristics of patients with cerebrovascular disease in the REACH Registry: The Spanish contribution. *Cerebrovascular Diseases, 24*(Suppl. 1), 89–95.

Chang, C.-L., & Chen, C.-H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications, 36*(2), 4035–4041.

Chimei (2007). Chimei Hospital Reports. <http://www.chimei.org.tw/97_newindex/layers2/dissertation97.html>.

DOH (2007). Department of Health, Executive Yuan, ROC (Taiwan). <http://www.doh.gov.tw/statistic/data>.

Eom, J.-H., Kim, S.-C., & Zhang, B.-T. (2008). AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. *Expert Systems with Applications, 34*(4), 2465–2479.

Faster, D. W. (1983). Diabetes mellitus. In K. J. Isselbacher (Ed.), *Harrison's principles of internal medicine* (10th ed.). New York: McGraw-Hill.

Ganzert, S., & Guttmann, J. (2002). Analysis of respiratory pressure–volume curves in intensive care medicine using inductive machine learning. *Artificial Intelligence in Medicine, 26*(1–2), 69–86.

García-Pérez, E., Violante, A., & Cervantes-Pérez, F. (1998). Using neural networks for differential diagnosis of Alzheimer disease and vascular dementia. *Expert Systems with Applications, 14*(1–2), 219–225.

Gustafson, D. H., Sainfort, F., Johnson, S. W., & Sateia, M. (1993). Measuring quality of care in psychiatric emergencies: Construction and evaluation of a Bayesian index. *Health Services Research, 28*(2), 131–158.

Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications, 36*(2), 3465–3469.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Kennedy, L., Lee, Y., Roy, V. B., Reed, C. D., & Lippman, R. P. (1997). *Solving data mining problems through pattern recognition*. New Jersey: Prentice Hall.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence, 7*, 317–337.

Kukar, M., Kononenko, I., & Groselj, C. (1999). Analyzing and improving the diagnosis of ischaemic heart disease with machining learning. *Artificial Intelligence in Medicine, 16*(1), 25–50.

Lim, T.-S., Loh, W.-Y., & Shih, Y.-S. (1997). *An empirical comparison of decision trees and other classification methods*. UW, Madison: Department of Statistics. TR979.

Lin, F. R., Chou, S. P., & Chen, Y. (2001). Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics, 62*(1), 11–25.

Liu, X., Bowyer, K. W., & Hall, L. O. (2004). Decision trees work better than feed-forward back-propagation neural nets for a specific class of problems. *IEEE International Conference on Systems, Man and Cybernetics, 6*, 5969–5974.

Loether, H. J., & McTavish, D. G. (1993). *Descriptive and inferential statistics: An introduction* (4th ed.). Needham Heights, MA: Allyn and Bacon.

Maher, P. E., & Clair, D. S. (1993). Uncertain reasoning in an ID3 machine learning framework. In *Proceedings of the 2nd IEEE international conference on fuzzy systems, FUZZ-IEEE'93* (Vol. 1, pp. 7–12).

Pogue, V. A., Ellis, C., Michel, J., & Francis, C. K. (1996). New staging system of the fifth joint national committee report on the detection evaluation, and treatment of high blood pressure (JNC-V) alters assessment of the severity and treatment of hypertension. *Hypertension, 28*, 713–718.

Pouliot, M., Despres, J. P., Lemieux, S., Moorjani, S., Bouchard, C., & Tremblay, A. (1994). Waist circumference and abdominal sagittal diameter: Best simple anthropometric indexes of abdominal visceral adiposetissue accumulation and related cardiovascular risk in men and women. *The American Journal of Cardiology, 73*, 460–468.

Quinlan, J. R. (1986). Induction of decision tree. *Machine Learning, 1*(1), 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann.

Ronco, A. L. (1999). Use of artificial neural networks in modeling association of discriminant factors: Towards an intelligent selective breast cancer screening. *Artificial Intelligence in Medicine, 16*(3), 299–309.

Science News (1990). Smoking boosts death risk for diabetics. *Science News, 138*(4), 61.

Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications, 29*(3), 583–588.

Übeyli, E. D., & Güler, I. (2003). Neural network analysis of internal carotid arterial Doppler signals: Predictions of stenosis and occlusion. *Expert Systems with Applications, 25*(1), 1–13.

Vapnik, V. N. (1982). *Estimation of dependencies based on empirical data, addendum*. New York: Springer-Verlag.

Wang, X.-H., Zheng, B., Good, W. F., King, J. L., & Chang, Y.-H. (1999). Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54*, 115–126.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools with java implementations*. San Francisco: Morgan Kaufmann.