International Journal of Nanomedicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

SHORT REPORT

Effective heart disease prediction system using data mining techniques

This article was published in the following Dove Press journal:
International Journal of Nanomedicine

Poornima Singh[1]
Sanjay Singh[2]
Gayatri S Pandi-Jain[1]

[1]L. J. Institute of Engineering and Technology, Gujarat Technological University, [2]Institute of Life Sciences, School of Science and Technology, Ahmedabad University, Ahmedabad, Gujarat, India

Abstract: The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, an effective heart disease prediction system (EHDPS) is developed using neural network for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The EHDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge, eg, relationships between medical factors related to heart disease and patterns, to be established. We have employed the multilayer perceptron neural network with backpropagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

Keywords: data mining, neural network, multilayer perceptron neural network, backpropagation, disease diagnosis

Introduction

Among various life-threatening diseases, heart disease has garnered a great deal of attention in medical research. The diagnosis of heart disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, family history of heart disease, obesity, high blood pressure, and lack of physical exercise.

A major challenge faced by health care organizations, such as hospitals and medical centers, is the provision of quality services at affordable costs.[1] The quality service implies diagnosing patients properly and administering effective treatments. The available heart disease database consists of both numerical and categorical data. Before further processing, cleaning and filtering are applied on these records in order to filter the irrelevant data from the database.[2] The proposed system can determine an exact hidden knowledge, ie, patterns and relationships associated with heart disease from a historical heart disease database. It can also answer the complex queries for diagnosing heart disease; therefore, it can be helpful to health care practitioners to make intelligent clinical decisions. Results showed that the proposed system has its unique potency in realizing the objectives of the defined mining goals.

Methods

The experiment was carried out on a publicly available database for heart disease. The dataset contains a total of 303 records that were divided into two sets, training set

Correspondence: Poornima Singh
L. J. Institute of Engineering and Technology, Gujarat Technological University, S.G. Road, Ahmedabad, Gujarat 382210, India
Email poornimasingh0806@gmail.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/IJN.S124998

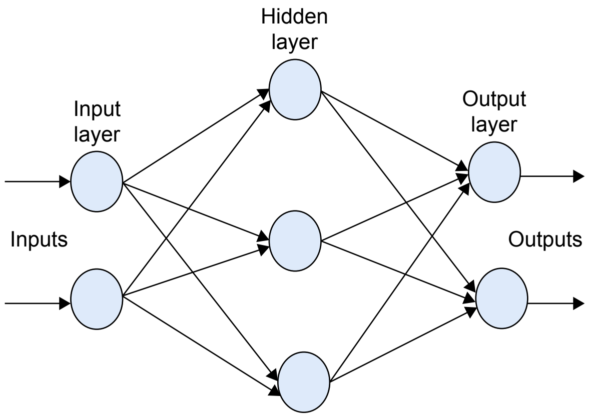


Figure 1 Multilayer perceptron neural network.

(40%) and testing set (60%). A data mining tool named Weka 3.6.11 was used for the experiment. Additionally, multilayer perceptron neural network (MLPNN) with backpropagation (BP) was used as the training algorithm.

MLPNN

MLPNN is one of the most significant models in artificial neural network. The MLPNN consists of one input layer, one or more hidden layers and one output layer.[3] In MLPNN, the input nodes pass values to the first hidden layer, and then nodes of first hidden layer pass values to the second and so on till producing outputs as shown in Figure 1.

BP network

The BP algorithm has served as a useful methodology to train multilayer perceptron for a wide range of applications.[4] The BP network calculates the difference between real and predicted values, which is circulated from output nodes backwards to nodes in previous layer. The BP learning algorithm can be divided into two phases, propagation and weight update.[4]

First, this learning algorithm provides training data to the network and compares the actual and desired outputs. Then, it calculates the error in each neuron. Based on this, the algorithm calculates what output should be for each neuron and how much higher or lower output must be adjusted for desired output and finally adjusts the weights. The overall process is done to improve weights during processing.

Results and discussion

In order to predict the probability of patients having heart disease, a confusion matrix (Table 1) was created, where A denotes patients with heart disease, and B denotes patients with no heart disease.

Table 1 A confusion matrix

	A (patients with heart disease)	**B (patients with no heart disease)**
A (patients with heart disease)	TP	FN
B (patients with no heart disease)	FP	TN

Abbreviations: TP, true positive; FN, false negative; FP, false positive; TN, true negative.

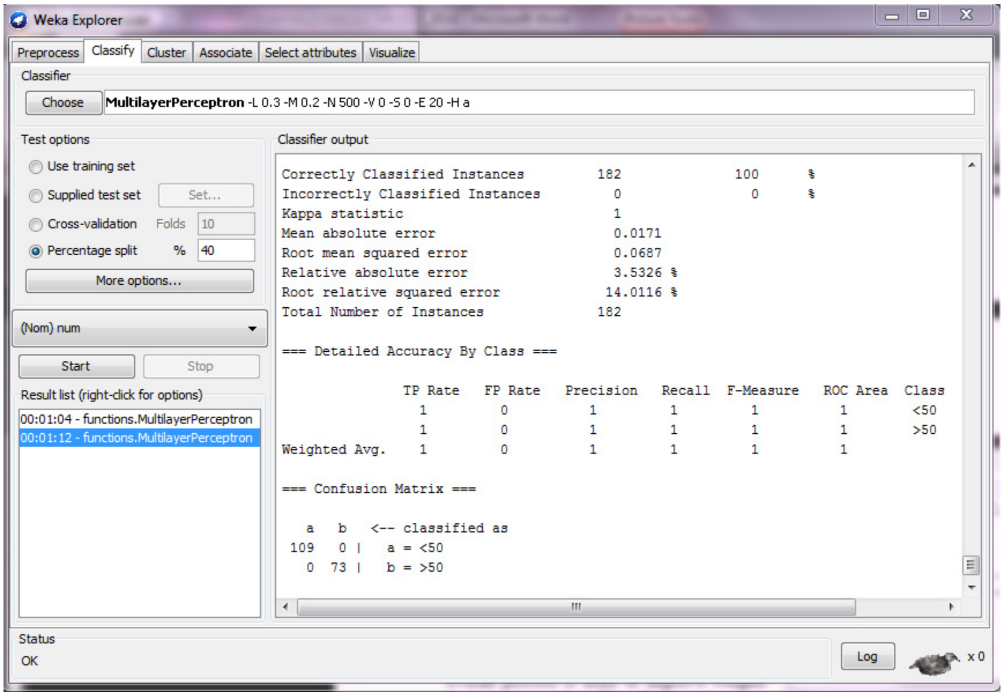
Table 2 Description of 15 used parameters

S no	**Parameters**	**Parameter description**	**Values**
1	age	Age in years	Continuous
2	sex	Male or female	1= male 0= female
3	thestbps	Resting blood pressure	Continuous value in mmHg
4	cp	Chest pain type	1= typical type 1 2= typical type angina 3= non-angina pain 4= asymptomatic
5	chol	Serum cholesterol	Continuous value in mm/dL
6	fbs	Fasting blood sugar	1≥120 mg/dL 0≤120 mg/dL
7	restecg	Resting electrographic results	0= normal 1= having ST-T wave abnormal 2= left ventricular hypertrophy
8	thalach	Maximum heart rate achieved	Continuous value
9	old peak	ST depression induced by exercise relative to rest	Continuous value
10	exang	Exercise induced angina	0= no 1= yes
11	ca	Number of major vessels colored by fluoroscopy	0–3 value
12	slope	Slope of the peak exercise ST segment	1= unsloping 2= flat 3= downsloping
13	thal	Defect type	3= normal 6= fixed 7= reversible defect
14	obes	Obesity	1= yes 0= no
15	num	Diagnosis of heart disease	0%≤50% 1%>50%

Table 3 Results for neural network showing 100% accuracy

	A (patients with heart disease)	**B (patients with no heart disease)**
A (patients with heart disease)	109 (TP)	0 (FN)
B (patients with no heart disease)	0 (FP)	73 (TN)

Abbreviations: TP, true positive; FN, false negative; FP, false positive; TN, true negative.



Classifier: MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Test options:
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 40

(Nom) num

Result list (right-click for options)
00:01:04 - functions.MultilayerPerceptron
00:01:12 - functions.MultilayerPerceptron

Classifier output

```
Correctly Classified Instances         182               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0171
Root mean squared error                  0.0687
Relative absolute error                  3.5326 %
Root relative squared error             14.0116 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        <50
                 1         0          1         1         1          1        >50
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

   a   b   <-- classified as
 109   0 |   a = <50
   0  73 |   b = >50
```

Status
OK

Figure 2 Results showing accuracy for 15 parameters using Weka tool.

A confusion matrix contains information about real and predicted classifications done by a classification system. The data in the matrix are evaluated to know the performance of such systems.

The confusion matrix contains the following four entries:

*TP (true positive): The number of records classified as true while they were actually true.

*FP (false positive): The number of records classified as true while they were actually false.

*FN (false negative): The number of records classified as false while they were actually true.

*TN (true negative): The number of records classified as false while they were actually false.

The overall process of effective heart disease prediction system (EHDPS) is based on the following three steps:

1. Data collection
2. Data pre-processing and
3. The classification of data.

The data are collected from a standard dataset that contains 303 records. The 15 parameters, such as age, sex, chest pain type (CP), and cholesterol (chol), with some domain values associated with them, considered to predict the probability of heart disease are shown in Table 2.

The collected data were used to create a structured database system. The pre-processing was done by identifying the associated fields and removing all the duplications. After that, all the missing values were filled, and the data were coded according to the domain value.

After applying neural networks on training dataset, the results show that there are zero FN or FP entries (Table 3), suggesting that the system predicts heart disease with 100% accuracy. Figure 2 shows the actual work done by Weka 3.6.11 tool.

Conclusion

In this study, an EHDPS has been presented using data mining techniques. From ANN, an MLPNN together with BP algorithm is used to develop the system. The MLPNN model proves the better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient. This system performs realistically well even without retraining. Furthermore, the experimental results show that the system predicts heart disease with ~100% accuracy by using neural networks.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. *Int J Comput Sci Net Secur*. 2008;8:343–350.
2. Sayad AT, Halkarnikar PP. Diagnosis of heart disease using neural network approach. *Int J Adv Sci Eng Technol*. 2014;2:88–92.

3. Gudadhe M, Wankhade K, Dongre S. Decision support system for heart disease based on support vector machine and Artificial Neural Network. In: Computer and Communication Technology (ICCCT), 2010 International Conference on, 2010:741–745.

4. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating error. *Nature*. 1986;323:533–536.