Journal of Southeast University (English Edition) Vol. 24, No. 3, pp. 267 – 271 Sept. 2008 ISSN 1003—7985

Neural network based online hypertension risk evaluation system

Ma Guangzhi Lu Yansheng Song Enmin Nie Shaofa Jing Weifeng Zhang Wei

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Since the previous research works are not synthetic and accurate enough for building a precise hypertension risk evaluation system, by ranking the significances for hypertension factors according to the information gains on 2 231 normotensive and 823 hypertensive samples, totally 42 different neural network models are built and tested. The prediction accuracy of a model whose inputs are 26 factors is found to be much higher than the 81.61% obtained by previous research work. The prediction matching rates of the model for "hypertension or not", "systolic blood pressure", and "diastolic blood pressure" are 95.79%, 98.22% and 98.41%, respectively. Based on the found model and the object oriented techniques, an online hypertension risk evaluation system is developed, being able to gather new samples, learn the new samples, and improve its prediction accuracy automatically.

Key words: hypertension prediction; neural network; information gain

Hypertension factors are studied through different aspects such as demography, dietary habit, and physical health care by using methods such as statistical analysis, regression analysis, and meta-analysis[1-4]. Without dealing with a large amount of samples and factors, previous results are not accurate enough for building a precise prediction model. However, an artificial neural network (ANN) is found to be more accurate than the methods mentioned before[5]. By using age, gender, and diastolic blood pressure (DBP) in the past 24 h as inputs, Poli et al.[6] built up an ANN model to diagnose and to offer clinical treatments for patients. Also by taking factors such as DBP, systolic blood pressure (SBP), and body mass index (BMI) as inputs, Ning et al.[7] set up an ANN cardiovascular risk stratifying model whose prediction accuracy reached 81.61%.

Given sufficient samples and adequate factors, the ANN model can achieve higher accuracy in prediction. To make use of the ANN model's self-study ability and to reach higher prediction accuracy, this research intends to collect a large amount of samples through the Internet, quantitatively analyze the significances of different factors with respect to hypertension, build a high accurate hypertension prediction model, and develop a hypertension risk evaluation system. Our research is based on a broad survey in Yichang city of the Three Gorges area of Hubei Province, China. Tongji Medical School of Huazhong University of Science and Technology (HUST) has designed the survey questionnaire which has been adopted in gathering samples.

The survey covered people whose ages are 35 to 92 and is carried out by well trained clinical doctors, who take charge of enquiring and filling in/out the questionnaire, examining the health of the volunteers, and recording their SBP and DBP. Altogether 3 054 valid samples are collected and 133 factors are considered. The factors refer to many aspects such as demography, medical history, smoking habit, drinking habit, dietary habit, physical activity, economy status, educational background, medical insurance, and family genetic history.

To accurately study the multiple factors that lead to hypertension, we adopt information gain and neural network methods. The research procedure is as follows: 1) Analyzing 42 pure objective factors by computing information gain and ranking the significances for these factors that yield to hypertension; 2) Setting up 42 neural network models with different factors according to the ranking results and evaluating the prediction precision of "hypertension or not", SBP, and DBP for each model; 3) Choosing a neural network model for predicting hypertension risk, then developing an online hypertension risk evaluation system.

1 Relevance of the Factors with Respect to Hypertension

The general idea behind the relevant analysis is to calculate some measures used to quantify the relevance of a factor yield to a given class. Statistics, machine learning, fuzzy set, and rough set methods can be used to analyze the relevance of a factor. By reducing factors with relevant analysis, we can simplify a model used for classification and prediction. The relevant measures of factors include information gain, Gini index, correlation, coefficient etc. Among these measures, the information gain[8] is based on statistics and broadly used for factor selection in building classification and prediction models.

Before the information gain has been computed for each factor, the samples stored in a database are cleaned in order to construct precise prediction models. In the database, each survey questionnaire is taken as a row, and each factor is taken as a column of the row. The last three columns of the row are the sphygmomanometer measured values of SBP, DBP, and the diagnostic result "hypertension or not". These columns are used as three predictable variables.

By taking the column "hypertension or not" as the class label of the samples, we calculate the information gains for all the 133 factors, and rank the significances for 56 objective factors according to their information gains. The top 42 objective factors are listed in Tab. 1, which show that our results are consistent with Geleijnse's most points of view[4]. However, there is one difference in the rank of the significances relevant to "salt intake" and "physical activity". Our results show that the " physical activity " is more important

Received 2008-04-15.

Biographies: Ma Guangzhi (1964—), male, associate professor, maguangzhi@hust.edu.cn; Lu Yansheng (corresponding author), male, professor, LYS@hust.edu.cn.

Foundation item: The National High Technology Research and Development Program of China (863 Program) (No. 2006AA02Z347).

Citation: Ma Guangzhi, Lu Yansheng, Song Enmin, et al. Neural network based online hypertension risk evaluation system[J]. Journal of Southeast University (English Edition), 2008, 24(3): 267 – 271.

Tab. 1 Pathogenic factors and information gains

Factor	Information gain	Factor	Information gain
Age	0. 102 277 31	Body weight	0. 042 279 17
Body height	0. 037 058 29	Physical activity type	0. 035 433 12
Education degree	0. 034 275 19	BMI	0. 025 710 76
Health status	0. 022 879 15	Drink times last month	0. 022 587 29
Age at the first smoke	0. 022 339 12	Activity problem degree last half-month	0. 021 844 96
Marriage status	0. 018 947 21	Exercise-over-half-hour days a week	0. 013 573 49
Health change degree last half-month	0. 013 359 24	Drink-alcohol-over-150g times last month	0. 013 295 38
Legume intake days a week	0. 011 451 52	BMI group	0. 010 122 74
Social activity degree last half-month	0. 010 071 07	Number of cigarettes a day	0. 009 108 01
Egg intake days a week	0. 008 924 14	Body pain degree last month	0. 006 610 77
Salty food intake days a week	0. 006 490 32	Occupation	0. 006 441 16
Oil of fried dishes	0. 006 255 96	Fish/meat intake days a week	0. 005 416 46
Physical exercise times last month	0. 004 892 73	Gross income of family	0. 004 886 24
Smoke or not	0. 004 056 27	Per capita living area	0. 003 645 58
Sweet food intake days a week	0. 003 514 33	Fat intake days a week	0. 003 406 86
Feeling type last half-month	0. 003 343 68	Fumigated food intake days a week	0. 003 313 45
Number of cigarettes over 100 or not	0. 002 907 14	Major reason of smoke give-up	0. 002 822 64
Kinsfolk hypertension or not	0. 002 657 93	Family smoke or not	0. 002 519 64
Physical exercise or not	0. 002 493 23	Milk product intake days a week	0. 002 422 57
Vegetable/fruit intake days a week	0. 002 205 98	Smoke days last month	0. 002 177 61
drink alcohol or not	0. 002 093 99	Medical care type	0. 001 531 56

than the "salt intake" in causing "hypertension or not". We suppose that the "physical activity" may help in metabolizing salt by drinking large amounts of water.

2 The ANN Prediction Model for Hypertension

Since neural networks generally accept normalized input and output values, we need to normalize our sample data before the ANN models are trained. We build and train 42 ANN models subject to the 42 objective factors studied previously. After all the models have been evaluated by doing the cross validation, we choose one model with high prediction accuracy and high generalizing ability as our hypertension prediction model.

2.1 The structure of the ANN models

In our ANN models, we use tanh and sigmoid as the transfer functions for the hidden and the output layer nodes, respectively. Therefore, the input values need to be normalized into the interval $[-1, 1]$, and the output values need to be normalized into the interval $[0, 1]$. The inputs of ANN models are not allowed to be empty. Thus, to normalize a column as the input of an ANN model, it is necessary to change the empty or null value into a certain new value. We first replace the null value with a different value other than all the normal values of this column, then we normalize the values of this column into the interval $[-1, 1]$.

By applying the normalization presented above, the trained ANN models will have a quality of tolerating faulty data when they are used for prediction. The min-max method is used to normalize the input and output values. Assume that for an input A, the minimal value is min$A$ while the maximal value is max$A$. Also assume that the original value v for the input A is transformed into a new value v', and the new min value for v' is newmin$A$ in the range $[-1, 1]$. The normalization formula is as follows:

$$v' = \frac{2(v - \min A)}{\max A - \min A} + \text{newmin}A \tag{1}$$

Our ANN models learn samples based on a back-propagation (BP) training algorithm. The most common BP training algorithm SDBP is introduced by Rumelhart et al. [9]. Later, some improved algorithms (such as the MOBP algorithm[10] which adds a momentum term, the VLBP algorithm[11] which has a variable learning speed, and the CGBP algorithm[12] which uses the conjugate gradient method) are also introduced. Among the four algorithms, the CGBP algorithm is the fastest one and has the feature of quadratic convergence. Moreover, the CGBP algorithm does not require a second derivative to calculate the Hesson matrix, thus it does not require a large amount of storage to store the Hesson matrix. So we use the CGBP algorithm to train our ANN models.

Our ANN models have the structure of three-layer feed forwards (i. e., the input layer, the hidden layer, and the output layer). Although we can build three single models each with only one output, we would rather build a composite model with three outputs. The reason is that building such a composite model could share the input and hidden layers and, thus, reduce the amount of neural cells and amortize the training cost, and, as a result, reduce the model training time.

As shown in Fig. 1, the input layer has n inputs which depends on the factors of the questionnaire, and the output layer contains three outputs y_1, y_2 and y_3 which are correspondent to the three predictable columns: "hypertension or not", SBP and DBP. According to our experience, the proper number of nodes in the hidden layer is equal to $4\sqrt{m \times 3}$. For each node in the hidden and the output layers, its net input value is a weighted sum of all its input values, which in turn is taken as the input of the transfer function of that node. Thus, the ANN model's output vector $\boldsymbol{Y}$ is described as follows:

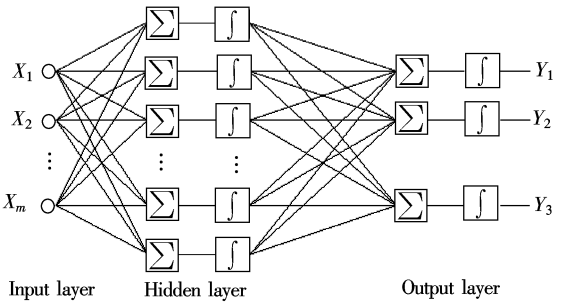

Fig. 1 Neural network structure for hypertension prediction

$$\boldsymbol{Y} = f_o(\boldsymbol{W}_o f_h(\boldsymbol{W}_h \boldsymbol{X} + \boldsymbol{B}_h) + \boldsymbol{B}_o) \tag{2}$$

where $\boldsymbol{X}$ is the input vector concerned with the hypertension factors; $\boldsymbol{W}_h$ and $\boldsymbol{W}_o$ are the weight vectors of the hidden and the output layers, respectively; $\boldsymbol{B}_h$ and $\boldsymbol{B}_o$ are the biases of the nodes in the hidden and the output layers, respectively; f_h is the transfer function (i. e. tanh) of the hidden layer nodes, whose output value is in the range $[-1, 1]$; f_o is the transfer function (i. e. sigmoid) of the output layer nodes, whose output value is in the range $[0, 1]$.

To choose an ANN model suitable for prediction, we build 42 ANN models in the following way: The first ANN model takes the first factor in Tab. 1 as its input, while the second ANN model takes the first two factors in Tab. 1 as its inputs. We repeat the training and cross validation process 100 times for each model. Each time we randomly divide the total 3 054 samples into two groups: the first group which contains 2 750 samples is used for training, and the second group which covers the remaining 304 samples is used for cross validation.

2.2 Cross-validation and error estimate for our ANN models

Having all of the 42 ANN models trained, we execute cross validations to estimate the errors between the original values of the predictable columns of the samples and the output values of the ANN models. As the output values of our ANN models are normalized values, while the original values of the predictable columns of the samples are discrete values, we need to map the continuous output values y_1, y_2, and y_3 to certain discrete values or to the contrary.

For convenience, we use superscripts to identify the training and cross validation times. Thus, the sample set for the first cross validation is written as s^1, and the value of a factor column c of a sample r in s^i is written as $s^i_{r,c}$. Similarly, the related output value of a predictable column c in a sample s^i_r is written as $y^i_{r,c}$, where $c = 1, 2, 3$. The output $y^i_{r,1}$, which is related to the predictable column "hypertension or not", has the binary values 0 and 1, where 0 represents normotensive and 1 represents hypertensive. We regard the predicted "hypertension or not" as normotensive if the output value $y^i_{r,1} \leqslant 0.5$. By doing the cross validation a 100 times and each time using 304 samples, the formula of the average error estimated for y_1 can be written as

$$\bar{E}(y_1) = \frac{1}{100 \times 304} \sum_{i=1}^{100} \sum_{r=1}^{304} (s^i_{r,1} \neq (y^i_{r,1} > 0.5)) \tag{3}$$

As for the output values $y^i_{r,2}$ and $y^i_{r,3}$, which represent SBP and DBP, respectively, we first find column $s^i_{r,c}$ with respect to $y^i_{r,2}$ and $y^i_{r,3}$, and then we normalize $s^i_{r,c}$ to $v^i_{r,c}$ by applying formula (1). Then we estimate the cross validation error for $y^i_{r,2}$ and $y^i_{r,3}$. By doing the cross validation a 100 times, the formula of the average error estimated for y_2 and y_3 can be written as follows:

$$\bar{E}(y_o) = \frac{1}{100 \times 304} \sum_{i=1}^{100} \sum_{r=1}^{304} \left(\frac{| v^i_{r,o} - y^i_{r,o} |}{v^i_{r,o}} \right) \qquad o = 2, 3 \tag{4}$$

The average matching rates of y_1, y_2 and y_3 are defined as $\mathrm{AMR}(y_o) = 1 - \bar{E}(y_o)$, where $o = 1, 2, 3$. By executing cross validations for the 42 ANN models, we obtain the average matching rates of y_1, y_2 and y_3 for each ANN model, which are shown in Figs. 2, 3 and 4, respectively.

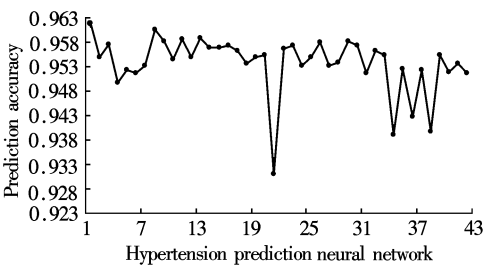

Fig. 2 Hypertension prediction matching rates of 42 ANNs

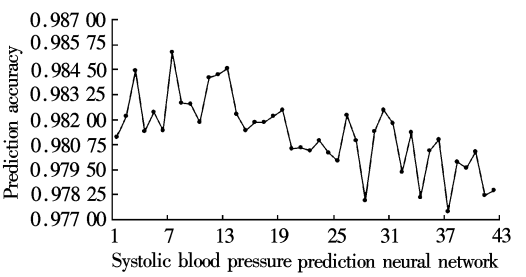

Fig. 3 SBP prediction matching rates of 42 ANNs

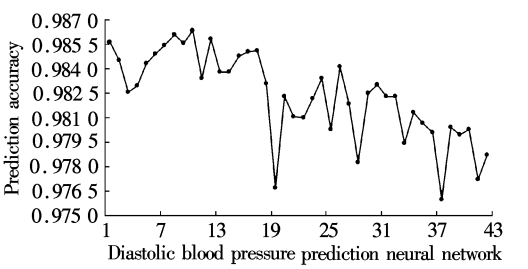

Fig. 4 DBP prediction matching rates of 42 ANNs

According to the principle of Ockham's razor, the ANN model structure should be as simple as possible under the condition that the prediction precision is acceptable. This means that fewer input ANN models should be adopted first. We suppose that the SBP of any person has, to a certain extent, a relationship with the DBP; i. e. the two factors SBP and DBP should have some functional association in an ANN model which predicts SBP and DBP at the same time. For example, the distribution coincidence curves of the prediction matching rates of SBP and DBP should be matched. Once the distribution coincidence curves of the prediction matching rates of SBP and DBP are matched, then

this indicates that the ANN model has reached a stable status in training or has been trained well.

As shown in Fig. 3 and Fig. 4, from the 21st to the 31st ANN model, the distribution coincidence curves of the prediction matching rates of SBP and DBP are almost matched, and the 26th ANN model has almost the highest prediction accuracy on SBP and DBP among these models. After the 26th ANN model, the prediction accuracy becomes lower and lower, which indicates that more samples are required to train the following models. Since the following models have more parameters relative to their input factors, they require more samples to achieve the same accuracy as the accuracy of the 26th ANN model. As a result, we decide to choose the 26th ANN model as our prediction model.

The prediction accuracies of our chosen model for "hypertension or not", SBP, and DBP are 95.79%, 98.22%, and 98.41%, respectively. They are much higher than 81.61% offered by Ning et al.[7], since our model has many more samples and far more factors taken as inputs.

3 Online Hypertension Risk Evaluation System

Our hypertension risk evaluation system is developed under the integrated developed environment of visual studio 2005 with C# and ASP. NET. We take IIS 6.0 as our web server and SQL server 2005 as our database server. All the samples, the ANN models, and the training results are stored by the database server.

We use the object-oriented method in designing and developing our risk evaluation system. As shown in Fig. 5, our model contains five classes used for data storing and data mining: DataBase, DataSource, DataSourceView, MiningStructure, and MiningModel. The attributes and the methods of these classes are presented and also, the associations between these classes are given.

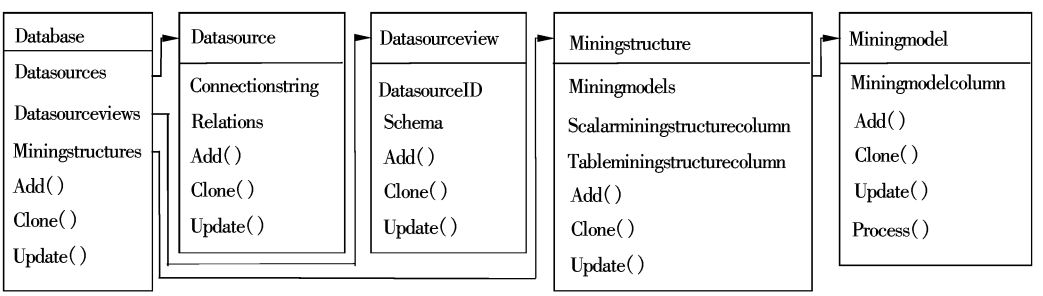

Fig. 5 Mining objects of our system

The instances of the class DataBase are used to store mining structures, data source views, and data sources. The factors of 3 054 samples are described by the instance of the class DataSourceView. However, which factors should be used to construct an ANN model is defined by the instances of the class MiningStructure, and the training algorithm is characterized by the instance of the class MiningModel. By setting up the training parameters or using the default parameters, the method process of the class MiningModel can be called on to train ANN models.

After the online hypertension risk evaluation has been deployed onto a web server, the users can log on to the home page of the online risk evaluation system through the Internet. They can register the basic information of their age, gender, occupation etc., and fill in/out the survey questionnaire with all the kinds of their hypertension factors. The system will automatically evaluate the hypertension risk for the users with the well-trained ANN models and offer the users some suggestions on how to prevent and treat the hypertension.

The online hypertension risk evaluation system has been checked and used by Tongji Medical School of HUST. While predicting hypertension according to the data input by the registered users, our system can automatically store the data into its database and use these data as new samples for its ANN model retraining. By using the hypertension risk evaluation system through the Internet in this way, it can automatically collect more samples and firmly improve its prediction accuracy continuously.

4 Conclusion

By using the information gain, we have finished a quantitative analysis of all the hypertension factors and made a ranking of the significances for 42 pure objective factors. The ranking results can be used as a reference in hypertension prevention and hypertension diagnosing. Also we have built and tested 42 ANN models based on the 42 objective factors, and have chosen an ANN model which contains 26 objective factors as our risk evaluation model. The prediction matching rate of our risk evaluation system exceeds 95%, which means that it can be effectively used for the risk evaluation and the clinical diagnosis of hypertension.

Acknowledgments This study is done in the Chinese Education Ministry Key Laboratory of Image Processing and Intelligent Control. We heartily thank Tongji Medical School of HUST for providing the hypertension samples and checking and using the hypertension risk evaluation system.

References

[1] Walker J, MacKenzie A D, Dunning J. Does reducing your salt intake make you live longer? [J]. *Interactive CardioVascular and Thoracic Surgery*, 2007, **6**(6): 793 − 798.

[2] He F J, MacGregor A. Effect of modest salt reduction on blood pressure: a meta-analysis of randomized trials. Implications for public health[J]. *Journal of Human Hypertension*, 2002, **16**(1): 761 − 770.

[3] Bacquer D D, Clays E, Delanghe J, et al. Epidemiological evidence for an association between habitual tea consumption and markers of chronic inflammation [J]. *Atherosclerosis*, 2006, **189**(2): 428 - 435.
[4] Geleijnse J M, Kok F J, Grobbee D E. Impact of dietary and lifestyle factors on the prevalence of hypertension in Western populations[J]. *European Journal of Public Health*, 2004, **14**(3): 235 - 239.
[5] Lisboa P J G. Neural networks in medical journals: current trends and implications for biopattern[C]//*Proc of the First European Workshop on Assessment of Diagnostic Performance* (*EWADP*). Milan, 2004: 99 - 112.
[6] Poli R, Cagnoni S, Livi R, et al. A neural network expert system for diagnosing and treating hypertension[J]. *Computer*, 1991, **24**(3): 64 - 71.
[7] Ning G, Su J, Li Y, et al. Artificial neural network based model for cardiovascular risk stratification in hypertension [J]. *Medical and Biological Engineering and Computing*, 2006, **44**(3): 202 - 208.
[8] Kent J T. Information gain and a general measure of correlation [J]. *Biometrika*, 1983, **70**(1): 163 - 73.
[9] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. *Nature*, 1986, **323**(9): 533 - 536.
[10] Rumelhart D E. *Parallel distributed processing*[M]. Cambridge, MA: MIT Press, 1986: 318 - 362.
[11] Vogl T P, Mangis J K. Accelerating the convergence of the back-propagation method [J]. *Biological Cybernetics*, 1988, **59**(3): 256 - 264.
[12] Patric P. Minimization method for training feedforward neural networks[J]. *Neural Networks*, 1994, **7**(1): 1 - 11.

基于神经网络的高血压在线风险评估系统

马光志　卢炎生　宋恩民　聂绍发　靖伟峰　张　魁

(华中科技大学计算机科学与技术学院,武汉 430074)

摘要:由于先前的研究工作不够综合和精确,不足于建立准确的高血压风险评估系统,根据 2 231 个正常样本及 823 个高血压样本计算的信息增益,对高血压致病因素的重要程度进行了排序,总共建立和测试了 42 个不同的神经网络模型,发现了一个输入为 26 个致病因素的神经网络模型,其预测精度远高于先前研究取得的 81.61%,该模型关于"是否高血压"、"收缩压"、"舒张压"的预测符合率分别为 95.79%,98.22% 和 98.41%. 基于发现的神经网络模型及面向对象的技术,开发了一个能自动收集新样本、学习新样本并能改进预测精度的高血压风险在线评估系统.

关键词:高血压预测;神经网络;信息增益

中图分类号:TP183