Intelligent Diagnosis of Diabetes based on Information Gain and Deep Neural Network

Lican Huang[1,2], Chuncheng Lu[1]

[1] School of Informatics, Zhejiang Sci-Tech University Hangzhou, 310018,China
[2] Domain Zones Technology Co,Ltd.,Hangzhou,China
licanhuang@zstu.edu.cn, chengguruchun@163.com,

Abstract: There are a great deal of diabetes patients all over the world, which has a severe effect on humans. It can cause a variety of serious complications. Therefore, early detection of diabetes is the key to reducing diabetes mortality. In this paper, an IG-DNN algorithm is proposed, which combines Information Gain (IG) and Deep Neural Network (DNN). In the proposed new method the information gain is applied to decrease the attributes to five ones ,which are input into DNN network. The new method presented in this paper has a classification accuracy of 90.26%, which is better than most previous research results.

Keywords: Diabetes Forecast; Deep Neural Network; Information Gain; Pima Indians Diabetes Dataset

1 Introduction

Insulin is a main anabolic hormone within the human body, which takes a key effect on converting sugar and other food items into the energy. Diabetes is disease caused by Insulin lack or other problems.

A report published in 2010 in the New England Journal of Medicine indicated that china has the world's largest population of diabetics and more than 92 million men and women have been diagnosed with the disease, an average of one out of 10 adults in china. The report says another 148 million Chinese citizens are pre-diabetic. At the same time, the economic loss of diabetes is astonishing. It is a serious threat to the health of people and the development of the national economy. Therefore, it is necessary to find a way to effectively control the onset and deterioration of diabetes. The study of diabetes has its necessity and practical significance in all aspects.

There are already some methods used for the diagnosis of diabetes. Sudha S. published a review paper about Disease Prediction by using Data Mining Technique[1]. In [2], the authors proposed the Homogeneity-Based Algorithm (or HBA) to optimally decrease the over-fitting and over-generalization behaviors of classification on Pima Indian diabetes data set (PIDD). In [3], Wu J., et.al. proposed a semi-supervised learning method, called as Laplacian support vector machine (LapSVM) to predict diabetes. Later, In [4], Hasan Temurtas , et.al. used a multilayer neural network which was trained by Levenberg Marquardt(LM) algorithm. In [5], Ergün U, et.al. presents a method for classification of MCA Stenosis in Diabetes by MLP and RBF Neural Network. In [6] , the authors used a recursive-rule extraction algorithm for the PIDD dataset. In [7] , Li Y. et.al used BP-network to construct fuzzy decision tree. Divya Tomar, et.al discussed WLSTSVM for the diagnosis of diabetes in [8]. In [9], Aslam M W, et.al used genetic programming for diabetes classification. In[10], Patil B M, et.al. proposed a hybrid prediction model for Type-2 diabetic patients. Erkaymaz O, et.al. analyzed the impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes[11].

In this paper, we combined with the Information Gain and Deep Neural Network to diagnosis diabetes and achieved a high classification accuracy.

2 Information Gain

Information gain can reflect the importance of attributes. The bigger value the information gain method gets, the more important the feature is. The Information gain based on the concept of entropy was applied to evaluate the quality of each attribute and choose those with high ranking by calculating the deference between post and prior entropy.

Prior entropy of X described by Eq(1), where X and Y are considered discrete variables and E is for Entropy:

$$\mathrm{E(X)} =-\sum_X P(X)\log_2 P(X) \qquad (1)$$

Here P(X) is the probability function of X. The conditional entropy of X given by post entropy Y is shown in Eq (2) and (3):

$$\mathrm{E(X|Y)} =-\sum_V P(Y)\mathrm{E(X|Y)} \qquad (2)$$

$$=-\sum_Y P(Y)\mathrm{E(X|Y)}\log_2 P(X|Y) \qquad (3)$$

The Information Gain (IG) is defined in Eq (4) and (5):

$$\mathrm{IG(X;Y) = Entropy(X) – Entropy\ (X|Y)} \qquad (4)$$

$$\mathrm{IG(X;Y)} = -\sum_X P(X)\log_2 P(X)- \sum_Y (-P(Y)\times \sum_X P(X|Y)\log_2 P(X|Y)) \qquad (5)$$

3 Deep Neural Network

Deep Neural Network (DNN) is a neural network with a lot of hidden layers. DNN can be split into three parts, input layer, hidden layers and output layer. In general, the first layer is the input layer, the last layer is the output layer, and a number of the middle layers is the hidden layers, as shown in Fig.1. And the DNN is based

978-1-5386-6005-8/18/$31.00 ©2018 IEEE

on forward propagation algorithm and back propagation algorithms.

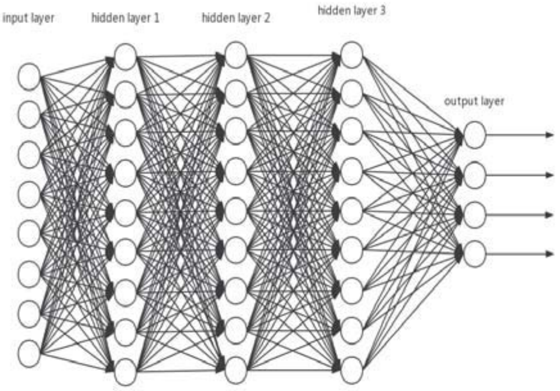

Fig1 the structure of DNN

3.1 Forward Propagation Algorithm

The so-called forward propagation algorithm uses the weight coefficient matrix ((W), the bias matrix (b) and the input value (x) to conduct a series of linear operations and activation operations. From the input layer, it is calculated forwards from one level to the next until it reaches the output level. And the output results are obtained. The steps is as follows:

```
Input: total number of layers ( $M$ ), The matrix corresponding to the hidden layer and the output layer ($W$), Bias vector($b$), output($x$);
Process：
1：Initialization: $a^1 = x$
2：For m = 2 to $M$, calculate:
        $a^m = \sigma(z^m) = \sigma(W^m a^{m-1} + b^m)$
Output : $a^m$
```

3.2 Back Propagation Algorithm

back propagation algorithm is presented below:

```
Input: total number of layers ($M$), Activation function, loss function, Iterative step( $a$ ), Maximum number of iterations ( $MAX$ ) and Stop iterative threshold ( ε ), training samples{$(x_1,y_1)$,$(x_2,y_2)$,...,$(x_n,y_n)$}
(1) Initialize the linear relationship between hidden layer and output layer. Set the value of coefficient matrix $W$ and bias vector $B$ a random value;
(2) for iter from 1 to $MAX$:
(2-1) for $i$ = 1 to $n$:
        (a) input of DNN $a^1$ set $x_i$
        (b) for $m$ = 2 to $M$, use forward propagation algorithm to calculate:
            $a^{i,m} = \sigma(z^{i,m}) = \sigma(w^m a^{i,m-1} + b^m)$
        (c) Calculation of the output layer by loss function: $\delta^{i,M}$;
        (d) for $m$ = $M$ to 2, Calculation of back propagation algorithm:
            $\delta^{i,m} = (W^{m+1})^T \delta^{i,m+1} \odot \acute{\sigma}(z^{i,l})$
(2-2) for $m$ = 2 to $M$, Update the $m$ layer $W^m$,$b^m$ :
            $W^m = W^m - a \sum_{i=n}^{n} \sigma^{i,m} (a^{i,m-1})^T$
            $b^m = b^m - a \sum_{i=1}^{n} \sigma^{i,m}$
(2-3) if the change of $W$, $b$ is less than $\varepsilon$ , Then jump out of the iteration loop to step (3);
(3) Output: get $W$ and $b$
```

4 Data and Method

4.1 Dataset of diabetes

The Pima Indian diabetes is one of the most popular datasets for testing the quality of diabetes classification algorithms. This dataset includes records of 786 female patients of Pima Indian heritage. The challenge is to predict whether a new patient is positive or negative for diabetes. The eight clinical features for this population are as the following.

1. Number of times pregnant(NP)
2. Plasma glucose concentration after 2h in an OGTT（PGC）
3. Diastolic blood pressure(mmHg)(DBP)
4. Triceps skinfold thickness(mm)(TSFT)
5. Two-hour serum insulin(μ U/mL)(2HSI)
6. BMI(BMI)
7. Diabetes pedigree function(DPF)
8. Age(years)(AGE)

The dataset information has been shown in Table I.

Table I the brief statistical analysis of PIDD

Attribute	Mean	Standard Deviation
NP	3.8	3.4
PGC	120.9	32.0
DBP	69.1	19.4
TSFT	20.5	16.0
2HSI	79.8	115.2
BMI	32.0	7.9
DPF	0.5	0.3
AGE	33.2	11.8

4.2 Methods

The method presented here combines the Information Gain (IG) method and Deep Neural Network (DNN). The function of IG is to choose the first-rank attributes. After IG step, the first-rank attributes will be input into the DNN network and train the attributes. Then ,we use test dataset to evaluate the accuracy. The flow chart of the proposed method is shown in Fig.2 .

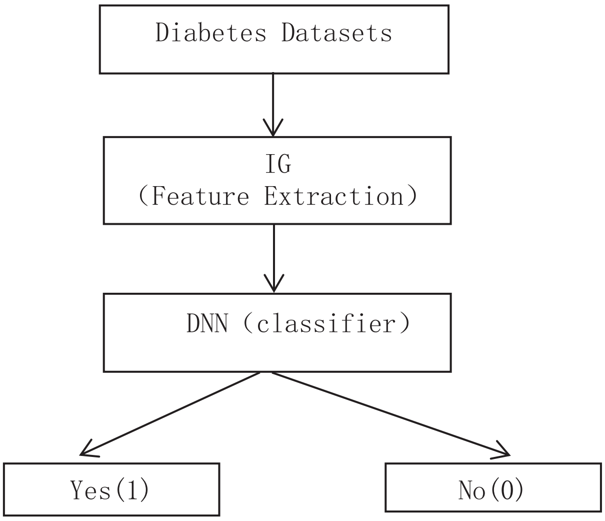

Fig2 IG-DNN flow diagram

5 Experiments and Results Analysis

Our proposed method mainly divided into two stages.

Firstly, we divide the dataset into training set and testing set, of which 90% is the training set and the remaining 10% is the training set. Then , we use the information gain method in the Waikato Environment for Knowledge Analysis to select the optimal attributes. Among the eight attributes, five attributes are chosen to obtain the highest classification accuracy. The selected attributes are following: Plasma glucose、Body mass index、Age、Two hour serum insulin(muU/ml)、Triceps skin fold thickness(μ U/ml) . The attribute ranking is shown in Fig.3.

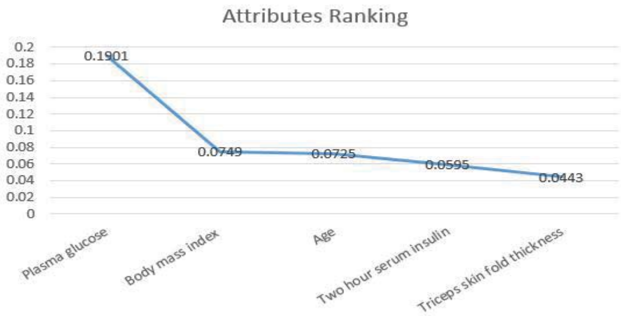

Fig3 Attributes selected by IG ranking

In the second part, we use Tensorflow to build a 5-layer DNN network with 12, 30, 50, 30, 12 units respectively. We choose the Rectified Liner Uints (RELU) as the activation function. The network structure is shown as Fig 4. The, we used this model to train the diabetes dataset from UCI, and after about 3200 iterations, a higher classification accuracy was achieved.

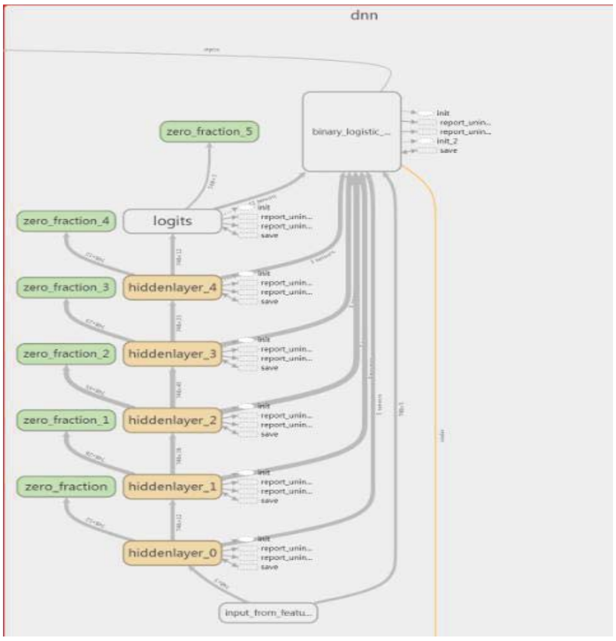

Fig4 DNN structure

The accuracy rate of the classification and the loss of the result are shown in Fig 5 and Fig 6.

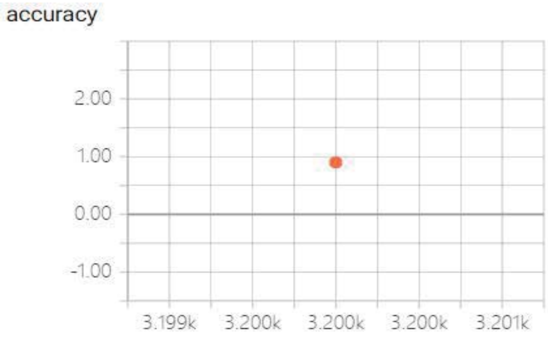

Fig5 Accuracy achieved

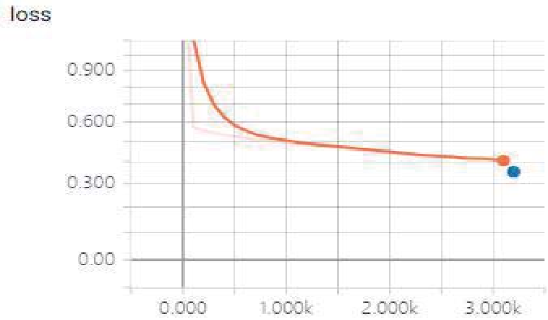

Fig6 Loss of result

We used k-fold cross validation technique to evaluate the classification accuracy of our method. We set k as 10 in this experiment. The dataset was split into ten subsets, where 9 subsets form a training set while remaining one was used as test set. At last the average error across all ten trails is computed. The classification accuracy for IG-DNN to diagnose diabetes disease was calculated and the result is 90.26%, which is higher than most previous studies.

$$\text{Accuracy (A)} = \frac{\sum_{i-1}^{|A|} assess(\boldsymbol{a_i})}{|\boldsymbol{A}|},\ a_i \in A \tag{6}$$

$$\text{Assess (a)} = \begin{cases} 1,\ if\ classify(a) = a.c \\ 0,\ otherwise, \end{cases} \tag{7}$$

In equation (6) and (7), “A” represents test-set of data items to be classified, $a_i \in$ A; $a.c$ indicates item “a” class and classification of a_i is labeled value, which is same as DNN classifier returns .

The classification accuracy by this paper and best values achieved by the other studies for pima-diabetes disease dataset are displayed in table II.

Table II comparison with earlier diagnostic methods

Method	Accuracy(%)	Year/Refs
Levenberg-Marquardt	77.08	2003/3
GDA-LS-SVM	79.16	2006/3
DT-HBA	91.6	2008/2
LapSVM	82.29	2009/3
MLNN with LM	82.37	2009/4
HFS + WLSTSVM	89.71	2014/8
Extreme Learning Machine	77.63	2015/6
Re-Rx with J48graft	83.83	2016/6
Small-World ANN	91.66	2016/13
IG-DNN	90.26	This study

6 Conclusions

A new method for predict diabetes is introduced in this paper. In the first step, we apply the information gain method to select the first-rank attributes, then apply these attributes to the DNN network to diagnose diabetes. The accuracy of IG-DNN is 90.26%, better than most other researches. In the future, we will gather more data of diabetes cases from the hospitals and try to test different network models and hyper parameters to get better results.

Acknowledgements

The paper is a part of the project “The function development of Semantic P2P networks”, which is financially supported by Domain Zones Technology Co., Ltd,Hangzhou,China.

References

[1] Sudha S. Disease Prediction in Data Mining Technique – A Survey[J]. International Journal of Computer Applications & Information Technology, 2013, 2(1):189-195..

[2] Pham H N A, Triantaphyllou E. Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization[M]// Computer and Information Science. 2008:11-26.

[3] Wu J, Diao Y B, Li M L, et al. A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis[J]. Interdisciplinary Sciences Computational Life Sciences, 2009, 1(2):151-155.

[4] Temurtas H, Yumusak N, Temurtas F. A comparative study on diabetes disease diagnosis using neural networks[J]. Expert Systems with Applications, 2009, 36(4):8610-8615.

[5] Ergün U, Barýþçý N, Ozan A T, et al. Classification of MCA Stenosis in Diabetes by MLP and RBF Neural Network[J]. Journal of Medical Systems, 2004, 28(5):475.

[6] Hayashi Y, Yukita S. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset[J]. Informatics in Medicine Unlocked, 2016, 2:92-104.

[7] Li Y, Wang X Z, Hua Q. Using BP-network to construct fuzzy decision tree with composite attributes[C] International Conference on Machine Learning and Cybernetics. IEEE, 2004:1791-1795 Vol.3.

[8] Tomar D, Agarwal S. Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing Breast Cancer, Hepatitis, and Diabetes[M]. Hindawi Publishing Corp. 2015.

[9] Aslam M W, Zhu Z, Nandi A K. Feature generation using genetic programming with comparative partner selection for diabetes classification[J]. Expert Systems with Applications, 2013, 40(13):5402-5412.

[10] Patil B M, Joshi R C, Toshniwal D. Hybrid prediction model for Type-2 diabetic patients[J]. Expert Systems with Applications, 2010, 37(12):8102-8108.

[11] Erkaymaz O, Ozer M. Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes[J]. Chaos Solitons & Fractals the Interdisciplinary Journal of Nonlinear Science & Nonequilibrium & Complex Phenomena, 2016, 83:178-185