Research on Diabetes Management Strategy Based on Deep Belief Network

Yang Liu[1,2(✉)], Zhijie Zhao[1,2], Jiaying Wang[1,2], Ang Li[1,2], and Jialin Zhang[1,2]

[1] School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China
1249182744@qq.com

[2] Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin University of Commerce, Harbin 150028, China

Abstract. Diabetes is a chronic disease that seriously endangers human health. Early detection, early diagnosis and early treatment can reduce the possibility of diabetic complications and mortality, which can be solved effectively by prediction model, assisting doctors to make more comprehensive and reliable diagnosis and treatment decisions, and improving diabetes management strategies. Thus, a diabetes prediction model based on Deep Belief Network (DBN) is proposed. Based on the Pima Indians Diabetes data set, the relative strength between the input attributes and the output targets of the model is calculated by using the weight matrix among the layers of the DBN diabetes prediction model. The results showed that plasma glucose concentration, body mass index, diabetic pedigree function, gestational frequency and age are important indexes for diabetes diagnosis. Then, this paper proposes three management strategies, including diabetes prevention education, diabetes individual prevention and diabetes community prevention to improve the management and control of diabetes in China.

Keywords: Diabetes mellitus · Deep learning · DBN model · Diabetes prediction · Diabetes management strategy

1 Introduction

Diabetes is one of the most serious and critical health problems facing the world in the twenty-first Century. Data provided by the International Diabetes Federation in 2017 showed that 424.9 million people suffer from diabetes in the world, and that by 2045, the number of diabetics in the world could increase by 48% to 628.6 million [1].

Diabetes is highly correlated with various factors such as age, sex, obesity and heredity, most of which is hidden and likely to cause acute or severe long-term complications. Now it is very important to develop predictive models for assisting doctors to make more comprehensive and reliable diagnosis and treatment decisions [2]. Many kinds of models have been established for predicting the risk of Diabetes mellitus. Thirteen classification models are evaluated and Random Forest is confirmed to be the best performance, which is used to create a web application for predicting

M. Jia et al. (Eds.): WiSATS 2019, LNICST 281, pp. 177–186, 2019.
https://doi.org/10.1007/978-3-030-19156-6_17

disease risk classification [3]. The Cox proportional hazards regression method is used to construct a prediction model for type 2 diabetes mellitus [4]. Some researchers set up a prediction model from the perspective of drug failure. An efficient and effective ensemble of SVMs is proposed for the anti-diabetic drug failure prediction problem, which is confirmed that the prediction model is suitability and the accuracy is about 80% [5]. However, most of these articles use statistical methods or machine learning methods for modeling, whose ability to express complex functions is limited, and there are relatively few articles on factor analysis and pertinent suggestions. This paper combining the characteristics of diabetes diagnosis and data, constructs the diabetes prediction model by using DBN, which can identify complex patterns in data, to assist doctors to make more comprehensive and reliable diagnosis and treatment decisions. The relative strength between the input attributes (factors) and the output targets (diabetes) of the model is calculated by using the weight matrix among the layers of the DBN diabetes prediction model, and comprehensively analyzes the results to put forward and consummate diabetes management strategies for improving the status of managing and controlling diabetes in China.

2 DBN Prediction Modeling and Simulation Experiment

The DBN diabetes prediction model, the theoretical basis for its establishment is mainly based on literature [6–9].

The data is divided into two parts: training set and test set. For training set, the prediction model of diabetes based on DBN is established, combined with empirical formula, setting and adjusting the parameters to determine the optimal network structure of DBN. For test set, The DBN diabetes prediction model is validated by it. The establishment process is shown in Fig. 1.

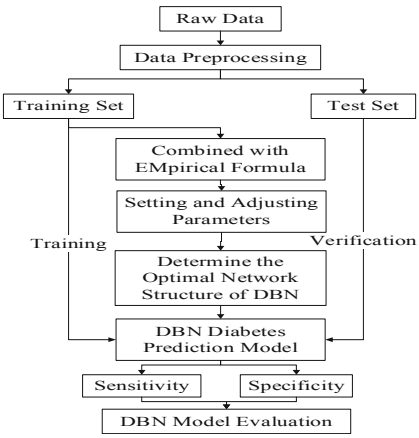

Fig. 1. Flow chart of diabetes prediction modeling based on DBN.

2.1 Data Source

The Data is about diabetes diagnosis information of native American women from Pima Indian heritage near phoenix, Arizona. They are above the age of 21 per capita. Eight risk factors related to diabetes are extracted, including six quantitative and continuous features, composed of a variety of clinical trial results, the remaining two quantitative and discrete features. The 9 features and sample data, including the classification codes, are shown in Tables 1 and 2 respectively. In Table 1, Body Mass Index (BMI) is calculated by formula, being recorded as $BMI = wei \div hei^2$, where *wei* represents weight, unit kg and *hei* represents height, unit m. In Table 2, Column 1 is the patient number. Columns 2–9 are the eigenvalues of the diabetes checking, as the attribute of the input data. The last column shows whether a given individual will suffer from diabetes within five years: 1 represents the positive in diabetes, that is, the individual will have diabetes in five years, a total of 268 cases. 0 represents the negative in diabetes, that is, the individual will not have diabetes within five years, a total of 500 cases. The Pima Indians Diabetes data set can be achieved in the UCI machine learning database.

Table 1. Data feature list (including category).

Feature number	Feature descriptions
Feature 1	Number of times pregnant
Feature 2	Plasma glucose concentration
Feature 3	Diastolic blood pressure (mm Hg)
Feature 4	Triceps skin fold thickness (mm)
Feature 5	2-Hour serum insulin (mu U/ml)
Feature 6	Body mass index
Feature 7	Diabetes pedigree function
Feature 8	Age in years
Classification codes	Binary class variable

Table 2. Original data sample (including category).

Number	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Classification codes
1	6	148	72	35	0	33.6	0.627	50	1
2	8	183	64	0	0	23.3	0.672	32	1
3	1	85	66	29	0	26.6	0.351	31	0
4	1	89	66	23	94	28.1	0.167	21	0
…	…	…	…	…	…	…	…	…	…

2.2 Training Set and Test Set

After considering the problem about the number of diabetes data sets and model training, this paper adopts K-fold Cross Validation method to determine the training set

and test set, so as to avoid over-learning and under-learning effectively. The basic idea of K-fold Cross Validation is to divide the available data into K parts (K $\geq$ 3), each representing a subset. Arbitrary K-1 subset is combined into a training set, and the remaining subset is used as a test set [10]. Thus, different training sets and corresponding test sets of K groups could be obtained. In the implementation, making K = 3, 768 pieces of input data, which are normalized by [0, 1] of the interval normalization method, are divided into three parts. 256 pieces of each part represent a subset, and three groups of training sets and corresponding test sets are obtained.

2.3 DBN Network Structure

Only four important parameters including the number of input layer nodes, hidden layer nodes, hidden layer nodes and output layer nodes are identified to determine the DBN network structure before the prediction model can be established. The number of input layer nodes is the attribute number of the dataset and the number of output layer nodes is equal to the class number in the dataset. The empirical formula for selecting the number of hidden layers and nodes is adopted to determine its approximate range, which can avoid the blindness and increase the effectiveness of selection [11]:

$$S = \sqrt{mn} + \mathrm{k}/2 \tag{1}$$

Here, S is the number of nodes in the hidden layer, m is the number of nodes in the input layer and n is the number of nodes in the output layer. K is constant belonging to 1–10.

Under the condition that m, n, k are known, the number of nodes in the first hidden layer S_1 can be calculated to have p values according to formula (1), which are recorded as $S_1 = [S_{11}, S_{12}, S_{13}, \cdots, S_{1P}]$. Assuming that the number of nodes in the first hidden layer has been determined, the number of nodes in the first hidden layer is equal to the number of nodes in the input layer, for example, if $S_1 = S_{1P}$, then $m = S_{1P}$, and the number of nodes in the output layer n remains unchanged. According to formula (1), the number of nodes in the second hidden layer S_2 can be calculated to have q values, which are recorded as $S_2 = \left[S_{21}, S_{22}, S_{23}, \cdots, S_{2q}\right]$. By analogy, under the different values of the number of nodes in the first hidden layer S_1, the number of nodes in the second hidden layer S_2 can be calculated respectively, which should be merged, and are recorded as $S_2 = \left[S_{21}, S_{22}, S_{23}, \cdots, S_{2q}, \cdots, S_{2(q+a)}\right]$. This calculation idea of simultaneously determining the number of hidden layers and the number of corresponding hidden layer nodes can be continued and determined experimentally.

2.4 DBN Network Parameters

In the process of training DBN model, parameters need to be set and adjusted to improve the classification accuracy of it. The sigmoid function is set as the hidden layer and the output layer neuron activation function. The maximum number of network cycles is 1000 times and the training data per batch is 128. The value of learning rate is set to 1 and the value of momentum factor is set to 0, which are contained in the RBM

training parameters. BP training parameters include learning rate, the value of which is set to 2, and momentum factor, the value of which is set to 0.9.

2.5 DBN Network Training

Under the same training parameters, it is necessary to combine the p value of S_1 with the $q + a$ value of S_2, and K-time DBN training should be carried out by using the K-group training set and test set, totaling $K * p * (q + a)$ experiments. The average accuracy of each classification under K-time DBN training is calculated as the final classification accuracy. The higher the classification accuracy rate, the better the prediction effect of DBN model.

2.6 DBN Experiment Results and Analysis

Seeing the data characteristics of Pima Indians Diabetes, it is known that the number of input nodes m is 8 and the number of output categories n is 2. Using the formula (1) and the idea of the corresponding algorithm, which are mentioned in Sect. 2.3, experiments are conducted to determine that the number of hidden layers is 2 layers. The number of nodes in the first hidden layer S_1 has 6 values, recorded as $S_1 = [4, 5, 6, 7, 8, 9]$, and the number of nodes in the second hidden layer S_2 has 9 values, recorded as $S_2 = [2, 3, 4, 5, 6, 7, 8, 9, 10]$. Three DBN network training are conducted for three groups of training sets, and 6 values of S_1 and 9 values of S_2 are combined in two to conduct each DBN training, in total of 3 * 6 * 9 experiments. The experimental results are shown in Table 3.

Table 3. DBN experiment results.

S2	S1					
	4	5	6	7	8	9
2	0.7591	0.7630	0.7721	0.7669	0.7697	0.7656
3	0.7630	0.7630	0.7565	0.7552	0.7578	0.7682
4	0.7734	0.7682	0.7318	0.7617	0.7642	0.6966
5	0.7708	0.7643	0.7617	0.7617	0.7513	0.7656
6	0.7656	0.7747	0.7396	0.7669	0.7708	0.6862
7	0.7565	0.7096	0.7357	0.7552	0.7695	0.7318
8	**0.7760**	0.7617	0.7591	0.7669	0.7305	0.7708
9	0.7734	0.7656	0.7620	0.7539	0.7539	0.7500
10	0.7565	0.7396	0.7591	0.7578	0.7669	0.7305

Table 3 shows that under the same training parameters, the number of nodes in two hidden layers is different, resulting in different classification accuracy of DBN, but the overall effect of the model is better than 65%, and the classification accuracy of some DBN network structures is as high as 80%. When the number of nodes in the input layer is 8, the number of nodes in the first hidden layer is 4, the number of nodes in the

second hidden layer is 8, and the number of nodes in the output layer is 2, the diabetes prediction model is best and the accuracy is 77.60%.

For verifying the accuracy of DBN model prediction, BP Neural Network and Support Vector Machine (SVM) are used to comparative experiments. The experimental results are shown in Table 4, which can be seen that the accuracy of the three diabetes prediction models is above 75%, indicating that they can predict diabetes accurately for the experimental data. The accuracy of DBN model is the highest, reaching 77.60%, which shows that it performs better than the other two models in the term of prediction accuracy.

Table 4. Accuracy comparison results of several models.

Model	Acc (%)
DBN	77.60
BP	76.3
SVM	76.56

3 Diabetes Prevention and Management Strategies

Diabetes is a complex lifelong disease, with the continuous development of artificial intelligence and in-depth learning technology, prediction accuracy of which could be improved, leading to be diagnosed more effectively and treated earlier. According to the experimental results of the DBN prediction model in Sect. 2, this paper discusses the effect of the input attributes on the output goals, analyzes the influencing factors of diabetes mellitus, and improves the prevention and management strategies of diabetes mellitus on the basics of the final analysis results.

Since DBN is a deep neural network, this paper evaluates the relationship between input attributes and output targets by the following formula [12]:

$$Y_{ji} = \frac{\sum_{h=0}^{n} (W_{hi} \times W_{jh})}{\sum_{i=0}^{m} \sum_{h=0}^{n} |W_{hi} \times W_{jh}|} \tag{2}$$

Here, W_{hi} represents the weights between the hth hidden nodes and the ith input nodes. W_{jh} represents the weights between the jth output nodes and the hth hidden nodes. Y_{ji} is the relative strength between the ith input and the jth output variable, representing the ratio of the relationship strength between the ith input and the jth output variable to the total relationship strength of all input and output variables.

Obtaining the weight matrix of each layer in the DBN-based diabetes prediction model, and combining with the formula (2), the relative strength between the eight influencing factors and diabetes mellitus was calculated. The results are shown in Table 5.

From Table 5, we can see that the relative strength values of the factors influencing diabetes are plasma glucose concentration, body mass index, diabetes pedigree function, pregnancy frequency and age in turn. The strongest effect on diabetes is plasma

glucose concentration, with a relative strength of 0.3407, suggesting that in glucose testing, the higher the plasma glucose concentration 2 h after oral administration, the greater the likelihood of diabetes. The relative effect of BMI on diabetes is 0.2637, which suggests that people with high BMI, or obesity, are more likely to suffer from diabetes. Relative strength values of the two factors to diabetes are large, indicating that they are important factors for diabetes, which need to be controlled by personal daily living habits and self-management ability, so as to delay the onset of diabetes. Among the 8 attributes, the correlation strength of diabetes pedigree function is 0.1332, which is in the third place, indicating that diabetes is related to heredity. The relative intensity of pregnancy frequency and age are not more than 0.1, indicating that they are not the main cause of diabetes, but also have a certain impact on it. The three factors are almost uncontrollable, but also closely related to diabetes, which need persons to increase the knowledge of diabetes and improve self-prevention awareness. Although the relative strength values of the other three factors are all negative, it does not mean that they had no effect on diabetes. It may be affected and limit by the data itself, which didn't reflect. All the above analyses are based on the Pima Indians Diabetes dataset. According to the analysis, the paper proposes management strategies of diabetes prevention and treatment besides basic drug treatment, which are summarized from the following aspects: diabetes prevention education management strategy, diabetes individual prevention and treatment management strategy, diabetes community prevention and treatment management strategy.

Table 5. Relative strength values.

Feature	Diabetes
Number of times pregnant	0.0904
Plasma glucose concentration	0.3407
Diastolic blood pressure (mm Hg)	−0.0549
Triceps skin fold thickness (mm)	−0.0261
2-Hour serum insulin (mu U/ml)	−0.0021
Body mass index	0.2637
Diabetes pedigree function	0.1332
Age in years	0.0534

(1) Diabetes prevention education management strategy

Studies have shown that diabetes is very common in clinical practice. Strengthening health education for diabetic patients is not only conducive to promoting patients' awareness of the disease, improving patients' compliance with treatment, but also can avoid wasting medical resources and lighten the social burden [13–15]. The strategies of diabetes prevention education management are mainly targeted at three kinds of people. For the non-diabetic patients, introduce the basic knowledge of diabetes to them, helping them strengthen awareness of prevention. For diabetics, they not only need to master basic knowledge of diabetes, but also need to know how to control it. Considering the signs on recession of understanding and memory in elderly diabetic

patients, family members or primary caregivers should be included to play a role of supervision and reminder, who have a direct impact on the rehabilitation effect of patients [16]. For medical professionals, education and training should be planned to enhance the professional knowledge and skills of diabetes.

(2) Diabetes individual prevention and treatment management strategy

Good living habits and self-management ability are the basis of diabetes treatment, throughout the entire diabetes treatment process. Studies at home and abroad have confirmed that increasing physical activities, diet control and weight loss can reduce or delay the onset of diabetes [17, 18]. Diabetes personal prevention and treatment management strategies are proposed for the details of daily life of diabetic patients, including self-blood glucose testing, diet therapy and exercise intervention. Firstly, small and fast blood glucose meters are gradually popularized, which can be used for blood glucose detection. Secondly, reasonable and effective diet treatment and control are conducive to weight loss, so as to preventing and treating diabetes. The last but not the least, regular and appropriate exercise intervention is also an important method, long-term exercise is an example, which can reduce weight, enhance physical fitness, and improve the body's disease resistance, so as to prevent and control the occurrence and process of diabetes.

(3) Diabetes community prevention and treatment management strategy

As a chronic and lifelong disease, diabetes patients need to return to family and society when their condition is stable, if they can't be effectively managed after discharge, they will increase the risk of readmission and the incidence of complications, which means that community plays an important role in the prevention and treatment of chronic diseases [19, 20]. There are many forms of diabetes prevention and management activities available in the community. Firstly, regular lectures on diabetes health education and publicity brochures are given to improve the level of awareness of diabetes. Secondly, the hospital-community integration model should be established by taking advantages of hospitals to provide professional training for community hospitals and diagnose, treat, guide patients, and making use of community to establish community prevention and monitoring websites to the registration, follow-up and various intervention activities, which is evaluated systematically to prove that it could enhance the awareness and reduce the incidence of diabetes [21]. Thirdly, the DBN-based diabetes prediction model can be embedded into the diabetes community prevention and monitoring website to strengthen the real-time prediction function for improving the management. Finally, according to the predicted results of model, a primary warning is given to the population with a disease probability of less than 30%, an intermediate warning is given to the population with a disease probability of more than 30% and less than 60%, and a severe warning is given to the population with a disease probability of more than 60%. Then, different strategies for prevention and treatment are proposed for different early-warning groups.

4 Conclusion

Diabetes is a complex lifelong disease, which could be predicted effectively by DBN. Based on the Pima Indians Diabetes dataset, the DBN prediction model is established and the accuracy is 77.60%. BP and SVM are used to establish model for comparison experiments, the results of which show that DBN is superior to the other two algorithms in predicting diabetes. Furthermore, the weight matrix of each layer of DBN prediction model is used to calculate the relative strength between each influencing factor and diabetes mellitus, and they are analyzed. The results showed that the relative strength values of plasma glucose concentration, body mass index, diabetic pedigree function, gestational frequency and age are 0.3407, 0.2637, 0.1332, 0.0904 and 0.0534 respectively, and on the basis of it, three management strategies, namely, diabetes education, personal prevention and community control, are put forward and improved.

Acknowledgements. This research is supported by the Harbin Science and Technology Bureau outstanding subject leader fund project (2017RAXXJ055) and the Humanities and social sciences research projects of the Ministry of Education (18YJAZH128).

References

1. International Diabetes Federation (IDF) Diabetes Atlas Eighth Edition poster Homepage. http://diabetesatlas.org/resources/2017-atlas.html. Accessed 24 Sept 2018
2. Kandhasamy, J.P., Balamurali, S.: Performance analysis of classifier models to predict diabetes mellitus. Proc. Comput. Sci. **47**, 45–51 (2015)
3. Nai-Arun, N., Moungmai, R.: Comparison of classifiers for the risk of diabetes prediction. Proc. Comput. Sci. **69**, 132–142 (2015)
4. Su, P., Yang, Y., Yang, Y., et al.: Prediction models on the onset risks of type 2 diabetes among the health management population. J. Shandong Univ. (Health Sci.) **55**(6), 82–86 (2017)
5. Kang, S., Kang, P., Ko, T., et al.: An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Syst. Appl. **42**(9), 4265–4273 (2015)
6. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
7. Li, H., Li, X., Ramanathan, M., et al.: Identifying informative risk factors and predicting bone disease progression via deep belief networks. Methods **69**(3), 257–265 (2014)
8. Lim, K., Lee, B.M., Kang, U., et al.: An optimized DBN-based coronary heart disease risk prediction. Int. J. Comput. Commun. Control **13**(4), 492–502 (2018)
9. Sun, Z., Xue, L., Xu, Y., et al.: Overview of deep learning. Appl. Res. Comput. **29**(8), 2806–2810 (2012)
10. Hu, J., Zhang, G.: K-fold cross-validation based selected ensemble classification algorithm. Bull. Sci. Technol. **29**(12), 115–117 (2013)
11. Yang, X.: Study of early warning for cerebrovascular risk based on deep beliefs networks. Beijing Jiaotong University (2016)
12. Lee, S., Choeh, J.Y.: Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Syst. Appl. **41**(6), 3041–3046 (2014)

13. Gao, Y., Hu, Y.: Influence of different health education modes on self-management level of diabetic patients. Jilin Med. J. **35**(3), 616 (2014)
14. Wu, S.: Application effect of health education in clinical nursing of diabetic patients. J. Tradit. Chin. Med. Manag. **25**(11), 1948–1949 (2014)
15. Zhao, F., Luo, J., Wang, Y., et al.: Analysis of the influencing on clinical health education management effect of diabetes. Int. J. Nurs. **35**(24), 3387–3392 (2016)
16. Scollan-Koliopoulos, M., O'Connell, K.A., Walker, E.A.: The first diabetes educator is the family: using illness representation to recognize a multigenerational legacy of diabetes. Clin. Nurse Spec. **19**(6), 302–307 (2005)
17. Li, G., Zhang, P., Wang, J., et al.: The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. Lancet **371**(9626), 1783–1789 (2008)
18. Saaristo, T., Moilanen, L., Korpihyövälti, E., et al.: Lifestyle intervention for prevention of type 2 diabetes in primary health care. One-year follow-up of the Finnish National Diabetes Prevention Program (FIN-D2D). Diabetes Care **33**(10), 2146–2151 (2010)
19. Xu, L., Liu, S., Chen, S., et al.: Readiness for hospital discharge and its influencing factors among diabetes patients. J. Nurs. Sci. **33**(10), 12–15 (2018)
20. Fang, W., Li, Y.: Correlation analysis between the readiness for hospital discharge and social support status in diabetic patients. Chin. J. Mod. Nurs. **22**(25), 3558–3561 (2016)
21. Fang, R., Xia, X.: Effect evaluation of whole course diabetes health education in hospital community integration. Chin. Community Doct.: Med. Spec. **11**(18), 251 (2009)