Diagnosis of Diabetes by using Adaptive Neuro Fuzzy Inference Systems

Adem Karahoca[1], Dilek Karahoca[2], Ali Kara[3]

[1] Engineering Faculty, Bahcesehir University, İstanbul, Turkey.

[1] akarahoca@bahcesehir.edu.tr , [2]dilek.karahoca@bahcesehri.edu.tr ,
[3]ali.kara@stu.bahcesehir.edu.tr

Abstract

Most of discoveries indicate that the best way to overcome diabetes is to prevent the risks of diabetes before becoming a diabetic. With this opinion, we would like to find a way to estimate diabetes risk, according to some variables such as age, total cholesterol, gender or shape of the body. Due to having fuzzy input and output (glucose rate) values and because of that dependent variable have more than 2 values (unlike binary logic), ANFIS and Multinomial Logistic Regression should be executed for comparison. Then the results were benchmarked. As a result, in case of that there is a system which contains fuzzy inputs and output, ANFIS gives better results than Multinomial Logistic Regression for diabetes diagnosis.

1. Introduction

Medical researches declare that there are approximately 5 million diabetic patients in Turkey. But unfortunately most of diabetic patients either don't visit physician regularly or don't know he is already diabetic. Aim of this study is to develop a diabetes expert system to help these kinds of patients.

While diabetes is not actually a form of heart disease, it often contributes to heart disease. Diabetes occurs when a body is unable to produce or respond properly to insulin which is needed to regulate glucose (sugar). Besides contributing to heart disease, diabetes also increases the risks of developing kidney disease, blindness, nerve damage, and blood vessel damage [1].

Most of discoveries indicate that the best way to overcome diabetes is to prevent the risks of diabetes before becoming a diabetic. It has not mentioned about the people who are diabetic from birth. In order to prevent the risk of diabetes, possibility of diabetes should be predicted. At this point, the main question is how to predict possibility of diabetes. It can be encountered that some researches dealing with prediction of diabetes risks [2-5]. However they generally worked as binary classification (1=Healthy, 0=Diabetic). This study aims to move a step beyond to make a prediction based on fuzzy dependent variable (1=Hypoglycemic, 2=Low Risk of Hypoglycemia, 3=Healthy, 4= Low Diabetes Risk, 5=Diabetic), instead of binary dependant one. When it was researched about diabetes disease, it was found out that reasons of diabetes are more than one. So our statistical methods shouldn't be linear. Some of the reasons which will be the input parameters of prediction system are defined fuzzy. After these findings, this study had been formed.

Adaptive Neuro Fuzzy Inference System (ANFIS) is used as an estimation method which has fuzzy input and output parameters. Then it was benchmarked the standard error of ANFIS with Multinomial Logistic Regression (MLR) as a non-linear regression method. Results show that ANFIS is more efficient than MLR with diabetes dataset. In order to make a successful benchmarking, dataset was preprocessed. Then MATLAB v7.1 was used for calculating ANFIS. Fundamental objectives of this paper are

- Demonstrating efficiency of ANFIS in diagnosing of diabetes.
- Behavior of MLR with diabetes disease data.
- Benchmarking the standard errors of the methods with training and testing datasets.

In the 2nd section, it has been demonstrated that the preprocessing of the dataset. ANFIS and MLR are explained in detail of the Section 3. MLR is explained as an alternative estimation method in the 4th Section. In section 5, ANFIS has been compared with and MLR. You can find conclusions of this study in the Section 6.

2. Preparation of Dataset

The dataset consist of 4 variables of 470 subjects who were interviewed in a clinic in Istanbul. All subjects are known as diabetic and all of them are under diabetes treatment.

In this study, we are trying to find any relation between diabetes risk and age, gender, total cholesterol and a ratio that is called frame. The waist/hip ratio (frame) may be a predictor in diabetes. The dependent variable in the dataset is *Glucose rate* and independent variables are

- Age
- Gender
- Frame (Waist/Hip ratio)
- Total Cholesterol

Some variables of the dataset have already fuzzy values, such as frame. However some of them don't have fuzzy values such as age, glucose. The main purpose of preprocessing data is to make fuzziness of the variable in order to use them in ANFIS. Preprocessing is explained in subsections.

2.1. Cleaning the Dataset

After the cleaning of the data for lacking value, our dataset decreased from 470 instances to 390. We determined 300 instances for training and 90 instances for checking.

9781-4244-3428-2/09/$25.00 ©2009 IEEE

2.2. Preprocessing the Dataset

The dataset has both fuzzy and continuous values. In order to use the data in ANFIS, we need fuzzy values to make successful estimation. We searched earlier works over expert systems about disease diagnosis and saw that most of them are interested in binary logic instead of fuzzy [14-16]. We categorized the variables which have continuous value, and then gave them fuzzy values as "very low (1)", "low (2)", "medium (3)", "high (4)" and "very high (5)". The table 1 shows the fuzzy values according the categories.

Table 1: Clustering variables

Age	Gender	Frame	Chol.	Gluc.
0–24	M	Small	< 200	< 60
25 – 49	F	Medium	201 – 240	60-89
50 – 74		Large	over 240	90 -120
75 – 99				121- 300
>= 100				over 300

Input A.G.E. columns and output glucose are classified by five classes. They are converted from continuous structure to fuzzy by this way.

The function *exhsrch* in MATLAB performs an exhaustive search within the available inputs to select the set of inputs that most influence the diabetes diagnosis. The first parameter to the function specifies the number of input combinations to be tried during the search.

The results that are shown in Table 2 indicates that it's better to make a set of age, frame and total cholesterol. Because all other alternative sets of inputs have more standard error of training data. Our purpose is to have estimation with the least standard error, so it's better to use the set with the least error. Essentially, exhsrch builds an ANFIS model for each combination and trains it for one epoch and reports the performance achieved.

Table 2: Result of preprocessing

INPUT			Train Error	Check Error
1	2	3	0.7154	0.7921
1	2	4	0.7139	0.7940
1	3	4	0.7122	0.8033
2	3	4	0.7346	0.7974

1. Age, 2.Gender, 3. Frame (Waist/Hip) ,4. Total Cholesterol.

3. Adaptive Neuro Fuzzy Inference System (ANFIS)

A Fuzzy Logic System (FLS) can be seen as a non-linear mapping from the input space to the output space. The mapping mechanism is based on the conversion of inputs from numerical domain to fuzzy domain with the use of fuzzy sets and fuzzifiers, and then applying fuzzy rules and fuzzy inference engine to perform the necessary operations in the fuzzy domain [6]. The result is transformed back to the arithmetical domain using defuzzifiers. The ANFIS approach uses Gaussian functions for fuzzy sets and linear functions for the rule outputs. The parameters of the network are the mean and standard deviation of the membership functions (antecedent parameters) and the coefficients of the output linear functions (consequent parameters).

The last node (rightmost one) calculates the summation of all outputs. Sugeno fuzzy model was proposed (Sugeno et al., 1988; Takagi et al. 1985) proposed fuzzy if-then rules are used in the model. A typical fuzzy rule in a Sugeno fuzzy model has the format

If x is A and y is B then x = f(x,y),

where A and B are fuzzy sets in the antecedent; z =f (x ,y) is a crisp function in the consequent. Usually f(x, y) is a polynomial in the input variables x and y, but it can be any other functions that can appropriately describe the output of the system within the fuzzy region specified by the antecedent of the rule. When f (x, y) is a first-order polynomial, we have the first-order Sugeno fuzzy model, which was originally proposed in [7,8]. When f is a constant, we then have the zero-order Sugeno fuzzy model, which can be viewed either as a special case of the Mamdani fuzzy inference system [9] where each rule's consequent is specified by a fuzzy singleton, or a special case of Tsukamoto's fuzzy model [10] where each rule's consequent is specified by a membership function of a step function centered at the constant. Moreover, a zero order Sugeno fuzzy model is functionally equivalent to a radial basis function network under certain minor constraints [6]. Consider a first-order Sugeno fuzzy inference system which contains two rules:

Rule 1: If X is A_1 and Y is B_1, then $f_1 = p_1x + q_1y + r_1$
Rule 2: If X is A_2 and Y is B_2, then $f_2 = p_2x + q_2y + r_1$

The fuzzy reasoning mechanism is summarized in Figure 1 [6]. Weighted averages are used in order to avoid extreme computational complexity in defuzzification processes.

$$\begin{aligned} f_1 &= p_1x + q_1y + r_1 \\ f_2 &= p_2x + q_2y + r_2 \end{aligned} \quad \Rightarrow \quad \begin{aligned} t &= \frac{w_1 + f_1 + w_2 f_2}{w_1 + w_2} \\ &= \overline{w_1} f_1 + \overline{w_2} f_2 \end{aligned} \qquad (1)$$

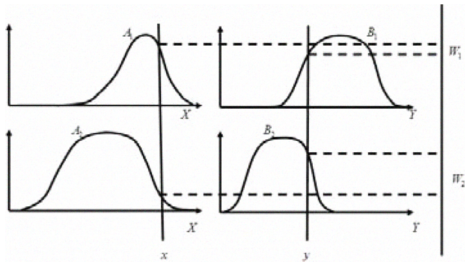

Figure 1: First-Order Sugeno Fuzzy Model

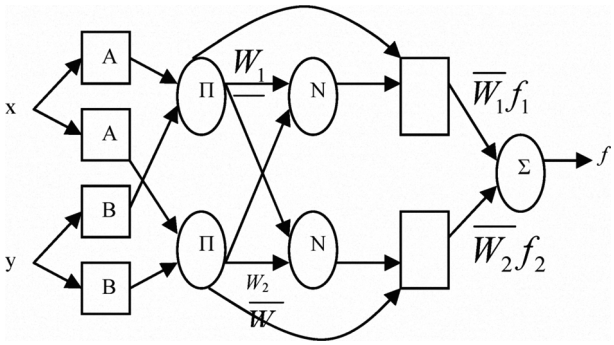

Figure 2: ANFIS Architecture

Figure 1 illustrates graphically the fuzzy reasoning mechanism to derive an output f from a given input vector [x,y]. The firing strengths w_1 and w_2 are usually obtained as the product of the membership grades in the premise part, and the output f is the weighted average of each rule's output. To facilitate the learning (or adaptation) of the Sugeno fuzzy model, it is convenient to put the fuzzy model into the framework of adaptive networks that can compute gradient vectors systematically. The resultant network architecture, called ANFIS (Adaptive Neuro-Fuzzy Inference System), is shown in Figure 2, where node within the same layer performs functions of the same type, as detailed below.

4. Multinomial Logistic Regression(MLR)

MLR is used when the dependent variable in question is nominal and consists of more than two categories. In this study, MLR would be appropriate, because of factors predict glucose level causing diabetes disease.

The MLR model assumes that data are case specific; that is, each independent variable has a single value for each case. The multinomial logistic model also assumes that the dependent variable cannot be perfectly predicted from the independent variables for any case.

$$\Pr(y_i = j) = \frac{\exp(X_i\beta_j)}{1 + \sum_j^J \exp(X_i\beta_j)} \qquad (2)$$

$$\Pr(y_i = 0) = \frac{1}{1 + \sum_j^J \exp(X_i\beta_j)}, \qquad (3)$$

According to MLR model, which is defined in (1) and (2), the ith is individual, y_i is the observed outcome and X_i is a vector off explanatory variables. The unknown parameters β_j are typically estimated by maximum likelihood [11].

5. ANFIS versus MLR

In this section, it have been evaluated the performance of ANFIS, comparing it with the performance of MLR. After preparation of dataset as explained in the section 2, we run ANFIS for training and checking datasets. Result of ANFIS process is shown by Figure 2.

```
ANFIS info:
    Number of nodes: 78
    Number of linear parameters: 108
    Number of nonlinear parameters: 27
    Total number of parameters: 135
    Number of training data pairs: 300
    Number of checking data pairs: 90
    Number of fuzzy rules: 27


Start training ANFIS ...

   1     0.709445    0.872509
   2     0.709445    0.87411

Designated epoch number reached --> ANFIS training completed at epoch 2.
```

Figure 3: ANFIS Results: Trained by 300 instances and checked by 90 instances.

Although it was used 300 instances for training and 90 ones for checking, ANFIS reached the results at just epoch 2. With the same training and checking datasets, we run MLR (MLR) model. We reached the results as listed in Table 3.

Table 3: Results of benchmarking

Method	epoch	RMSE	Data Type
ANFIS	2	**0.1418**	Train
ANFIS	2	**0.1745**	Check
MLR	300	**0.1417**	Train
MLR	90	**0.2343**	Check

We would like to have your attention to two important points while evaluating the results.

It's clear to see from Figure 1, RMSE of *training* datasets for both of the methods are very similar. However same parameter of *checking* datasets shows that MLR has much bigger RMSE than ANFIS.

Another important point is difference between learning durations of the methods. As seen from Table 3, the learning duration of ANFIS is shorter than MLR. ANFIS training could be completed at epoch 2; but MLR should evaluate the whole dataset.

6. Conclusions

We tried to determine an estimation method to predict glucose rate in blood which indicates diabetes risk. We initially designated continuous values in the dataset, and converted them to fuzzy values. We didn't made glucose rate (dependent variable) fuzzy, instead of binary. Binary values have high accuracy, but don't have enough information about diabetes risk.

After preprocessing dataset, we run ANFIS method and MLR method. Table 3 summarizes the results of benchmark we made between ANFIS and MLR.

We found out that learning duration of ANFIS is much shorter than MLR's duration. When a more sophisticated system with a huge data is imagined, the use of ANFIS instead of MLR would be more useful to overcome faster the complexity of the problem.

In training of the data, ANFIS and MLR gave quite similar results with standard error. However, when the trained parameters were applied to checking data, standard error of ANFIS is smaller than that of MLR. This shows that ANFIS is a better and faster learning method than MLR.

Consequently it could be said, if we have a system which contains fuzzy inputs and output, ANFIS is better system than MLR for diabetes diagnosis.

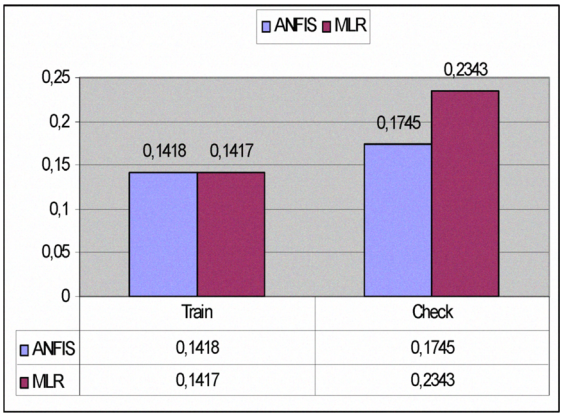

Figure 4: Root Means Squared Errors

7. References

[1] Polat K, Günes S: *An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease*. ELSEVIER, Digital Signal Processing 17 (2007) 702–710

[2] R. Prasad, K. R. Ranjan, A.K. Sinha, "*AMRAPALIKA: An expert system for the diagnosis of pests, diseases, and disorders in Indian mango*", Elsevier, Knowledge-Based Systems 19 (2006) 9–2.

[3] Willems JP, Saunders JT, DE Hunt, JB Schorling: *Prevalence of coronary heart disease risk factors among rural blacks: A community-based study.* Southern Medical Journal 90:814-820; 1997

[4] Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: *A trial of church-based smoking cessation interventions for rural African Americans*. Preventive Medicine 26:92-101; 1997

[5] Bin Othman MF, Moh Shan Yau T: *Neuro Fuzzy Classification and Detection Technique for Bioinformatics Problems*. Proceedings of the First Asia International Conference on Modeling and Simulation; 2007

[6] Jang J. S. R., Sun C. T. and Mizutani E: "*Neuro- fuzzy and soft computing. A computational approach to learning and machine intelligent*". United States of America. Prentice Hall International; 1997.

[7] SUGENO, M., KANG, G.T., Sturcture identification of fuzzy model. Fuzzy sets and Systems, 28(1988), pp.15-33.

[8] TAKAGI, T. and SUGENO, M., Fuzzy identification of systems and its application to modeling and control. IEEE Trans. On Systems, Man & Cybernetics, 15(1985), pp.116-132.

[9] MAMDANI, E. H. and ASSILIAN, S., An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies, 7(1), (1975), pp.:1-13

[10] TSUKAMATO, Y., An approach to fuzzy reasoning method. In M. M. Gupta, R. K.Ragade, and R. R. Yager, editors, Advances in Fuzzy Set Theory and Applications, (1979), pp:137-149.

[11] Krishnapuram B, Carin L, Figueiredo MAT, Hartemink A: *Sparse MLR*: *Fast Algorithms and Generalization Bounds.* IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 27, No. 6, JUNE 2005