HOSTED BY

ELSEVIER

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

ORIGINAL ARTICLE

A novel hybrid intelligent system with missing value imputation for diabetes diagnosis

Rohollah Ramezani [a,*], **Mansoureh Maadi** [b], **Seyedeh Malihe Khatami** [c]

[a] *Department of Statistic, Faculty of Mathematics and Computer Science, Damghan University, Damghan, Semnan, Iran*
[b] *Department of Industrial Engineering, Faculty of Engineering, Damghan University, Damghan, Semnan, Iran*
[c] *Department of Computer Engineering, Faculty of Engineering, Damghan University, Damghan, Semnan, Iran*

Received 31 December 2016; revised 26 March 2017; accepted 29 March 2017

KEYWORDS

Diabetes;
Intelligent system;
Missing value;
ANFIS;
Logistic regression

Abstract Recently, diabetes becomes the widespread and major disease in the world. In this paper, we propose a novel hybrid classifier for diabetic diseases. The proposed hybrid classifier named Logistic Adaptive Network-based Fuzzy Inference System (LANFIS) is a combination of Logistic regression and Adaptive Network-based Fuzzy Inference System. Our proposed intelligent system does not use classifiers to continuous output, does not delete samples with missing values, and does not use insignificant attributes which reduces number of tests required during data acquisition. The diagnosis performance of the LANFIS intelligent system is calculated using sensitivity, specificity, accuracy and confusion matrix. Our findings show that the classification accuracy of LANFIS intelligent system is about 88.05%. Indeed, 3–5% increase in accuracy is obtained by the proposed intelligent system and it is better than fuzzy classifiers in the available literature by deleting all samples to missing values and applying traditional classifiers to different sets of features.

1. Introduction

Recently, diabetes becomes the widespread and major disease in the world. Several researchers have focused on this disease to reduce the growth of this problem. People become diabetics when their body either does not produce or appropriately use insulin. Indeed, the insulin as a hormone plays crucial role in this disease. Sugar and other food are converted into required energy by insulin. Genetics and environmental factors are important reasons of this disease. In 1980, 108 million have diabetes while 422 million people have diabetes worldwide [1–3]. Two types of diabetes are defined- juvenile diabetes and adult-onset diabetes [2]. The obesity is a main cause of adult diabetes that can be postponed or controlled with appropriate diet and exercise, but complete cure of diabetes isn't possible.

It is hard to diagnose the diabetes due to the presence of several factors. Doctors commonly judge by evaluating the current test results of a patient or by comparing the patient with other patients that had similar symptoms and test results. Consequently, recognition of diabetes is very complicated issue

[*] Corresponding author at: Daneshgah Sq, Damghan, Semnan Province 36716-41167, Iran.
E-mail addresses: r_ramezani@du.ac.ir (R. Ramezani), m_moadi@du.ac.ir (M. Maadi), m_khatami@du.ac.ir (S.M. Khatami).

Peer review under responsibility of Faculty of Engineering, Alexandria University.

http://dx.doi.org/10.1016/j.aej.2017.03.043
1110-0168 © 2017 Faculty of Engineering, Alexandria University. Production and hosting by Elsevier B.V.

for doctors [2]. Hence, scientists and researchers have tried to present an intelligent diagnostic system for diagnosing diabetes.

The aim of this paper is to introduce and investigate the method for making a novel robust intelligent diagnosis system based on training data with missing values and lower dimension of the clinical attributes. Among various fuzzy modeling techniques with continuous output such as Adaptive Network Based Fuzzy Inference System (ANFIS), this paper proposed a hybrid classifier based on Logistic regression and ANFIS named LANFIS with binary output to help the physician on diagnosis of diabetes disease. The diagnosis performance of the LANFIS intelligent system is estimated using sensitivity, specificity, accuracy and confusion matrix.

This paper is organized as follows. Section 2 contains a review of the related technical literature, a brief explanation of data preprocessing and mathematical theory used in the proposed approach. The modeling of the proposed method by novel hybrid classifier is described in Section 3. Experimental results and discussion to investigate the effectiveness of the proposed method are shown in Section 4. Finally, Section 5 contains a summary of conclusions.

2. Methods and data

2.1. Literature review

Many studies have been done in the diagnosis and classification of disease specially diabetes and their complications [4–9]. In these researches, the information taken from patients and decisions of intelligent systems significantly influences the diagnosis of diabetes. Akram et al. [10] proposed a system consisting of a hybrid classifier for the detection of retinal lesions due the diabetic disease. The proposed system consists of data preprocessing, extraction of candidate lesions, feature set formulation, and classification. Abawajy et al. [11] presented a method for identifying cardiovascular autonomic neuropathy category in diabetic data with missing values. Recently, the researchers [12–15] have focused on fuzzy modeling as a proper technique for diagnosing disease. In Ref. [12], a fuzzy classifier named ARTMAP-IC neural network for medical diagnosis is proposed. This model improves the ARTMAP search algorithm to allow the network to encode inconsistent cases, and combines instance counting during training with distributed category representation during testing to obtain probabilistic predictions. They have reported 81% classification accuracy using ARTMAP-IC with conventional validation method (one training and one test) on diabetic diseases. In Ref. [2], Polat et al. presented a cascade learning system based on Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) to diagnose diabetes disease. They have reported 78.21% classification accuracy using LS-SVM with 10-fold cross-validation (10× FC). Dogantekin et al. [13] achieved 84.61% classification accuracy using Linear Discriminant Analysis (LDA) and Adaptive Network Based Fuzzy Inference System (ANFIS): LDA-ANFIS for diagnosing diabetic diseases. Temurtas et al. [16] have also reported 82.37% classification accuracy (conventional valid) on Pima Indian diabetes disease diagnosis using a multilayer neural network structure which was trained by Levenberg–Marquardt (LM) algorithm. There have been several other reports focusing on diagnosing diabetic diseases with accuracy between 59.5% and 80.5%. The accuracy values of these reports can be seen in Section 5.

All previous papers only circumvented the problem of missing values by ignoring missing values, deleting all attributes with missing values, removing all records with missing values from the training set and applying various classifiers. Now, we introduce and apply a novel intelligent system for handling diabetic data with missing values. It utilizes multiple imputation techniques to increase the accuracy of the obtained results.

2.2. Diabetes dataset

In this paper, Pima Indians diabetes dataset from the UCI Machine Learning Archive is used for modeling and describing our proposed approach. The proposed method can be applied to any other diabetes dataset. The dataset is selected from a larger dataset held by the National Institutes of Diabetes and Digestive and Kidney Diseases [17]. In this dataset, all patients are Pima-Indian women at least 21 years old and living near Phoenix and Arizona states in USA. Table 1 presents the eight clinical predictor attributes included in the Pima Indians diabetes dataset. The binary response attribute takes the values ‘0’ or ‘1’, where ‘1’ means a positive test for diabetes and ‘0’ is a negative test for diabetes. There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’. Missing values and high-dimensional data are the main deficiencies in Pima Indians diabetes dataset. These problems cause inaccurate results in classification techniques. In the next sections, some methods to overcome these difficulties have been introduced.

The concept of missing values is an important issue in mathematical modeling of data. Breault [18] noted that the five of attributes listed in Table 1 exhibit biologically implausible zero values while the zero value is considered for displaying the missing values. Pearson [19] represented a brief study on missing data in this dataset. He showed that the presence of a small concentration of disguised missing values can have serious consequences. In fact, this metadata is incorrect and has missing data. On this dataset, previous papers approached the problem of missing values by disregarding missing values, vanishing all attributes with missing values, eliminating all records with missing values from the training set and applying various classifiers [12,13,16,20].

Pima Indians diabetes dataset is high-dimensional data and caused some computational difficulties in proposed method. In this dataset, all the measured attributes are not “important”

Table 1 The eight clinical predictor attributes included in Pima Indians diabetes dataset.

No.	Attribute name	Description
1	NPG	The number of period pregnant
2	PGL	The plasma glucose concentrations
3	DIA	The diastolic blood pressures (mm Hg)
4	TSF	The triceps skin fold thickness (mm)
5	INS	The serum insulin concentrations
6	BMI	Body mass index
7	DPF	Diabetes pedigree function
8	AGE	Age in years

for specifying diabetes disease. Therefore, OT (Orthogonal Transformation) method is applied to determine insignificant attributes in Pima Indians diabetes dataset. These attributes will be neglected.

2.3. Multiple imputation method for missing values

The multiple imputation analysis is a statistical technique for analyzing incomplete datasets to obtain the missing values. In [21] it has been shown that if the method creates imputations properly, the resulting inferences will be statistically valid. In this paper, this method is applied to Pima Indians diabetes dataset in three steps as follows:

(I) *Imputation:* The missing values of the incomplete datasets are imputed m times. The results of this step are m complete datasets. In the following, we use $m = 5$ as a conservative choice.

(II) *Analysis:* Each m completed dataset is analyzed. Then, their performance is measured by MLP Neural Network model.

(III) *Pooling:* All m datasets are integrated into a final dataset. Therefore, the weighted mean of predicted values in m completed datasets is considered as imputed missing values. The weights are computed based on the performance in step II.

In stage I, the regression method is used to assign the new values. In the regression method, a regression model is fitted for each attribute with missing values while the previous attributes are considered as covariates. Based on the resulting model, a new regression model is fitted for assigning the missing values for each attribute ([22], pp. 166–167). If the dataset has a monotone missing data pattern, the process is repeated sequentially for attributes with missing values. Therefore, for attribute X_j with missing values, the following model is assigned with the non-missing observations.

$$X_j = c_0 + c_1X_1 + c_2X_2 + \ldots + c_{j-1}X_{j-1} \tag{1}$$

The fitted model estimates the regression parameters $(\hat{c}_0, \hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{j-1})$ and obtains the associated covariance matrix $\sigma_j^2 V_j$, where V_j is equal to $(X'X)^{-1}$ and $X = [X_1, X_2, \ldots, X_{j-1}]$.

New parameters $(c_{i0}, c_{i1}, \ldots, c_{i(j-1)})$ and σ_{ij}^2 are calculated from the posterior predictive distribution of the missing data for $i = 1, \ldots, m$ imputation. Indeed, they are simulated by using $(c_{i0}, c_{i1}, \ldots, c_{i(j-1)}), \sigma_{ij}^2, V_j$. The missing values are then replaced by

$$x_{ij,mis} = c_{i0} + c_{i1}x_1 + c_{i2}x_2 + \ldots + c_{i(j-1)}x_{j-1} + z_i\sigma_{ij}^2, \tag{2}$$

where z_i is a simulated value of a normal standard distribution. The analysis of missing value patterns does not cause any particular obstacles to multiple imputations.

2.4. Dimension reduction method

OT was used to make a classifier system more accurate and efficient. Therefore, before modeling, OT is applied to determine insignificant attributes in Pima Indians diabetes dataset. Removing these attributes of dataset will cause the dimension of diabetes dataset to reduce. Diabetes disease dataset is represented as a vector consisting of 8 attributes.

OT is a linear dimension reduction method based on mean-square error [23]. It is proved that OT has the best performance among all other possible orthogonal transform methods that are used to de-correlate the components of a given input. The most common inference of OT is in a standardized linear projection, such as $Y = W^TX$, which maximizes the variance in the projected space Y. For a given n-dimensional dataset X, the p principal axes $w_1, w_2, \ldots, w_p$, where $1 \leqslant p \leqslant n$ are orthogonal axes, and the memorized variance is maximum in the projected space. Generally, $w_1, w_2, \ldots, w_p$ can be defined by the p leading eigenvectors of the sample covariance matrix

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T, \tag{3}$$

where $x_i \in X$, m is the sample mean, and N is the number of samples, so that

$$Sw_i = \lambda_i w_i, \quad 1 \leqslant i \leqslant p \tag{4}$$

where λ_i is the ith largest eigenvalue of S. The p principal components of a given observation matrix X are given as below:

$$Y = [y_1, y_2, \ldots, y_p] = [w_1^TX, w_2^TX, \ldots, w_p^TX] = W^TX \tag{5}$$

2.5. ANFIS model

In this section, the basic theory of ANFIS model is initially presented. This model acquires both artificial neural network and fuzzy logic as ANFIS scheme (Fig. 1). ANFIS consists of if–then rules and couples of input–output. In addition, learning algorithms of neural network are used for ANFIS training. To simplify the explanations, two inputs (x and y) and one output (z) are considered for the fuzzy inference system. The ANFIS architecture with two inputs and one output is depicted in Fig. 1. This scheme is formed by using five layers and nine if–then rules as follows:

Layer-1: Every node i is a square node with a node function.

$$\begin{aligned} O_{1,i} &= \mu_{A_i}(x), \quad i = 1, 2, 3, \\ O_{1,i} &= \mu_{B_{i-3}}(x), \quad i = 4, 5, 6, \end{aligned} \tag{6}$$

where x and y are inputs for node i, and A_i and B_i are linguistic labels for x and y inputs, respectively. In addition, $\mu_{A_i}(x)$ and $\mu_{B_i}(x)$ are the membership functions of A_i and B_i, respectively. Typically, $\mu_{A_i}(x)$ and $\mu_{B_i}(x)$ are chosen to be bell-shaped with maximum equal to 1 and minimum equal to 0, such as

$$\mu_{A_i}(x), \mu_{B_i}(x) = \exp\left(-\left(\frac{x_i - c_i}{a_i}\right)\right)^2, \tag{7}$$

where a_i and c_i are premise parameters.

Layer-2: In this layer, every node is a circle node labeled by Π which multiplies the incoming signals and sends the product out. For instance,

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_{i-3}}(y), \quad i = 1, \ldots, 9. \tag{8}$$

An output node represents the firing strength of a rule.

Please cite this article in press as: R. Ramezani et al., A novel hybrid intelligent system with missing value imputation for diabetes diagnosis, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2017.03.043

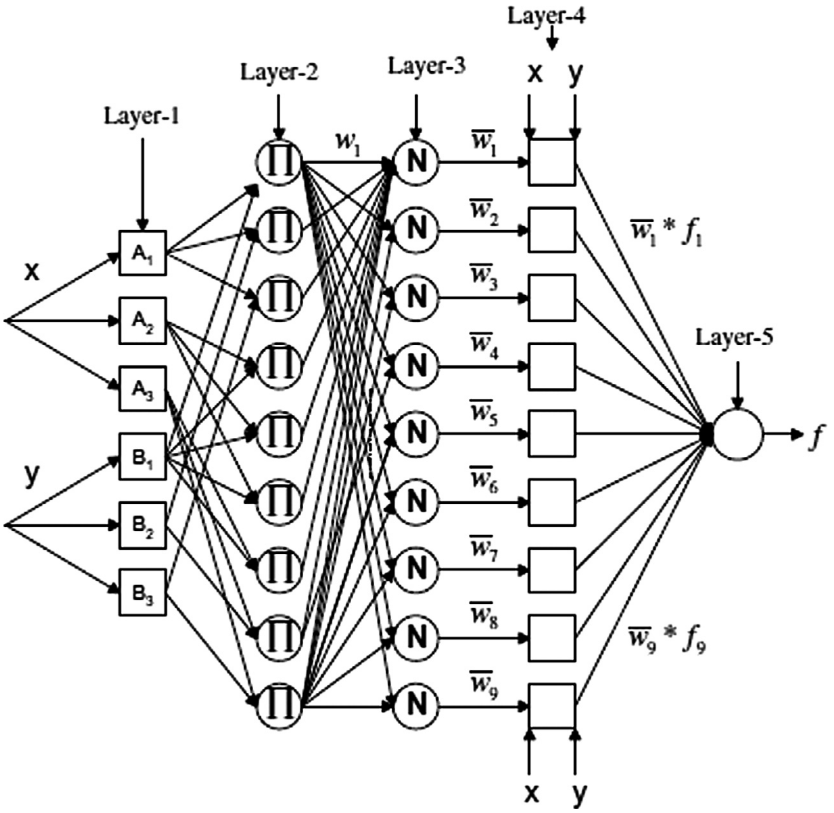

Figure 1 ANFIS architecture of two inputs and nine rules.

Layer-3: In this layer, every node is a circle node labeled by N. The ith node calculates the ratio of the ith rules firing strength to the sum of all rules firing strengths:

$$O_{3,i} = \bar{w}_i = \frac{w_1}{w_1 + \ldots w_9}, \quad i = 1, \ldots 9 \tag{9}$$

Layer-4: In this layer, node i is a square node with a node function

$$O_{4,i} = \bar{w}_i . f_i = w_i(p_i x + q_i y + r_i), \quad i = 1, \ldots 9 \tag{10}$$

where w_i is the output of layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set. These parameters will be referred to as consequent parameters in this layer.

Layer-5: The single node in this layer is a circle node that computes the overall output as the summation of all incoming signals:

$$O_{5,i} = \text{over all output} = \sum_i \bar{w}_i \cdot f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}. \tag{11}$$

3. Proposed method

The proposed method consists of two stages: data preprocessing and classification. The missing value analysis and feature selection in data preprocessing stage are important parts of robust pattern diagnosis. If the missing values do not manage properly, it may result an inaccurate inference about real class of the patients. If the dimension of dataset is not reduced by selecting the important features, the model will be more complex. Therefore, the accuracy of classifier is decreasing through class detection. In the classification stage, a novel hybrid classifier named LANFIS is applied to input data. The input data have lower dimension and no missing values are observed with effective information from the original data. The block diagram of the LANFIS intelligent system for diagnosing diabetes disease used in this paper is given in Fig. 2.

3.1. LANFIS hybrid classifier

One of the problems is that the output in ANFIS model is continuous while the response attribute is binary in Pima Indians diabetes dataset. Hence, the LANFIS intelligent diagnosis system will be introduced to solve this problem. This new model based on Logistic regression and ANFIS model can be integrated to classify Pima Indians diabetes dataset.

If the probabilities π_i depend on a vector of observed covariates x_i, π_i are initially considered as an ANFIS output as follows:

$$\pi_i = ANFIS(\mathbf{x}_i) \tag{12}$$

The challenge of this model is that the probability π_i on the left-hand-side should between zero and one while the output of ANFIS on the right-hand-side could take any continuous value. Therefore, there is no guarantee that the predicted values will be within the correct range unless complex restrictions are imposed. In this approach, a solution is transforming the

Please cite this article in press as: R. Ramezani et al., A novel hybrid intelligent system with missing value imputation for diabetes diagnosis, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2017.03.043

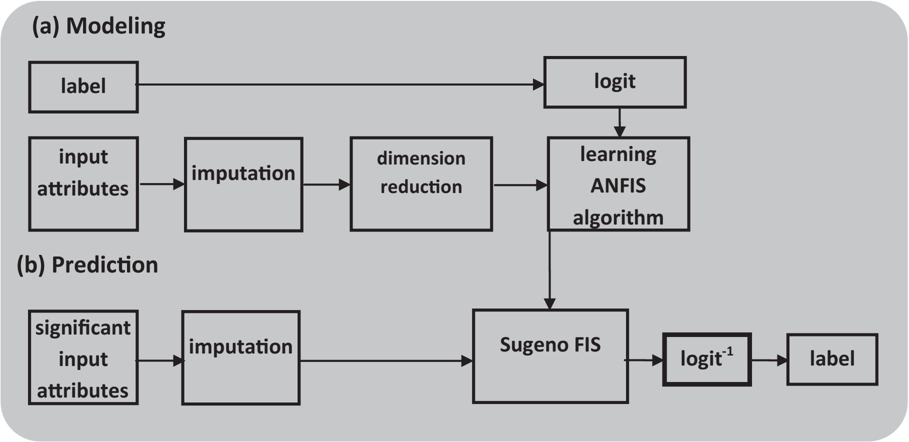

Figure 2 The block diagram of the LANFIS intelligent system for diagnosing diabetes disease.

probability to remove the range of restrictions. This will be done in two steps. First, new attribute $odds_i$ is defined as follows:

$$odds_i = \frac{\pi_i}{1-\pi_i}. \tag{13}$$

If the probability of an event is a half, *odds* are one-to-one. If the probability is 1/3, *odds* are one-to-two. If the probability is very small, *odds* are said to be long. In some contexts, the language of *odds* is more conventional than the language of probabilities.

Second, we apply logarithms function for calculating the logit or log-odds as follows:

$$\eta_i = logit(\pi_i) = log\frac{\pi_i}{1-\pi_i}, \tag{14}$$

This operation removes the restriction in low numbers. This point notes that the probability goes down to zero when the odds tends to zero and the logit approaches to $-\infty$. In addition, the probability becomes close to one when the odds and lower tend to $+\infty$. Thus, logits operations transform probabilities from the specific range $(0,1)$ to the entire real number. Note that if the probability is 1/2 then the *odds* are even and the logit is zero. Negative logits represents probabilities below 0.5 and positive logits correspond to probabilities above 0.5. Fig. 3 clearly illustrates the logit transformation. The logit transformation is one-to-one function. The inverse transformation is called the antilogit and allows applicant to transform logits to probabilities. According to Fig. 3, π_i could be obtained from Eq. (14) as follows

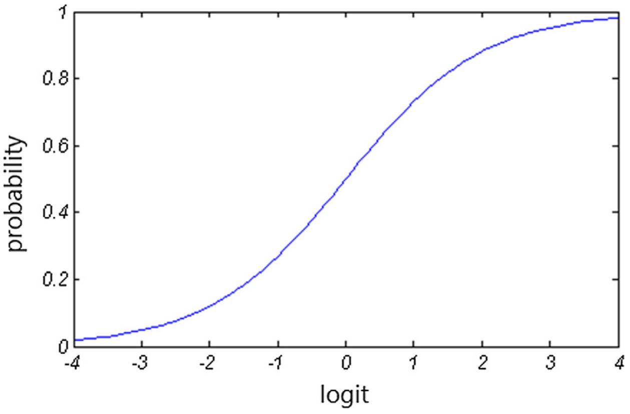

Figure 3 The logit transformation.

$$\pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{15}$$

Therefore, the LANFIS model is summarized as follows:

$$\eta_i = ANFIS(\boldsymbol{x}_i), \quad \pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{16}$$

4. Result

4.1. Measurement of intelligent diagnosis system performance

It is significant to know the efficiency of the present model. In this section, accuracy, k-fold cross validation and confusion matrix will be discussed. These subjects determine the efficiency of the proposed method.

(a) Accuracy

The accuracy of obtained results by applying LANFIS model is defined according to the following equations:

$$\text{class prediction accuracy } (N) = \frac{\sum_{i=1}^{|N|} assess(n_i)}{|N|}, \quad n_i \in N \tag{17}$$

$$assess(n) = \begin{cases} 1 & \text{if LANFIS}(n) = n_c \\ 0 & \text{otherwis} \end{cases} \tag{18}$$

where N is the set of data items to be classified (the test dataset), $n \in N$, n_c is the real class of the item n and LANFIS(n) determines the predicted class.

In addition, sensitivity and specificity are done to evaluate the robustness of the method. These measures are presented as follows:

$$Sensitivity = \frac{TP}{TP+FN}(\%) \tag{19}$$

$$Specificity = \frac{TN}{TP+TN}(\%) \tag{20}$$

where TP, TN, FP and FN denote respectively:

- TP (True positives): classifies Health as Health.
- TN (True negatives): classifies No Health as No Health.
- FP (False positives): classifies No Health as Health.
- FN (False negatives): classifies Health as No Health.

(b) *k*-Fold cross validation

Cross validation is one of effective methods to improve the evaluation and comparison of learning algorithms by dividing data into k segments. In each iteration, one of the k segments is used to examine a model and the other $k-1$ segments are put together to form a training set. We used 3-fold cross-validations in our proposed LANFIS intelligent classifier system since it reduces the bias associated with random sampling.

(c) Confusion matrix

A confusion matrix contains information about both correct and incorrect predicted classes. Table 2 shows the confusion matrix for a two classifier. In the next section, the confusion matrix of the proposed method is computed.

4.2. Result of missing value analysis

Table 3 defines the descriptive statistics for missing and valid data in Pima Indians diabetes dataset. The results of *t*-test show significant difference between means of NPG, DIA, TSF and INS attributes in two states and assign missing values as zero in data (total data) and not considering as data (valid data) (see Table 3). Pearson [19] confirmed that missing values of NPG attribute are a challenge in Pima Indian diabetes dataset because the zero values are used to code missing observations in other attributes while NPG attribute could be zero. Thus, it is not obvious either the zero value for NPG is used to code missing data or it is the real value of this attribute. Therefore, the *t*-student test is done for the comparison of the proportion of diabetic diseases in two groups of NPG = 0 and NPG > 1. The value of t-test statistic is determined ($t = -0.158$) and its significant level value is 0.875.

Table 2 Representation of confusion matrix.

	Predicted	
Observed	Health	Diseases
Health	α	β
Diseases	γ	δ

Therefore, there is no significant difference between two groups. So, the zero value is used to code missing observations in NPG attribute.

Table 4 displays the patterns of missing value for eight clinical attributes in diabetic patients. Each pattern corresponds to a group of elements with the same pattern of incomplete and complete data. For example, pattern 1 represents 336 cases with no missing values while pattern 10 represents 170 cases with missing values on TSF and INS. Pattern 15 represents eight cases with missing values on DIA, NPG, TSF, and INS. Pima Indians diabetes dataset contains potentially $2^8 = 256$ patterns for eight attributes. However, only 16 patterns of 768 cases in the dataset are represented. Table 4 shows that the dataset has a monotone missing data pattern. In Table 4, patterns 13, 15 and 16 have missing values in more than 4 attributes of 8 available attributes. The ratio of corresponded patients to these patterns is less than 50 percent information in the case of diabetic disease. Therefore, these patients will be removed and 753 patients will be considered for modeling.

The 753 patients have 702 missing value in eight clinical attributes. The multiple imputation techniques are used for substituting the missing values with predicted values. In step I, $m = 5$ complete datasets are generated. The accuracy values of $m = 5$ complete datasets in step I are measured by MLP Neural Network model in step II. The accuracy values are gained between 75% and 80%. All $m = 5$ datasets are integrated into a final dataset. As mentioned in Section 2.3, the weighted mean of predicted values in $m = 5$ completed datasets is considered as imputed missing values. The weights are the accuracy values in step II.

4.3. The result of dimension reduction

In Table 5, there are three components that their eigenvalues are greater than one and it represents 65.55 percent of total data variance. In the rotated space, these three components are orthogonal. Each component is a linear transformation of eight attributes in Pima Indians diabetes dataset. The coefficients of three linear OT transformations are shown in Table 6.

The important degree (sum of square coefficients) of 8 attributes in three components is computed in the last column. The five attributes that have the most important degree are PGL, TSF, INS, BMI and AGE. Therefore, these five attributes will

Table 3 Missing values and their effects on the descriptive statistics in 8 clinical attributes.

	No. of valid data	Missing data		Mean		Std. deviation		*t*-test for comparing mean	
		Count	Percent	Valid data	Total data	Valid data	Total data	Value of *t*-statistic	Sig.
NPG	657	111	14.5	4.49	3.85	3.22	3.37	−3.66	0
PGL	763	5	0.7	121.69	120.89	30.54	31.97	−0.50	0.62
DIA	733	35	4.6	72.41	69.11	12.38	19.36	−3.95	0
TSF	541	227	29.6	29.15	20.54	10.45	15.95	−11.79	0
INS	394	374	48.7	155.55	79.80	118.78	115.24	−10.40	0
BMI	757	11	1.4	32.46	31.99	6.93	7.88	−1.24	0.22
DPF	768	0	0.0	0.47	0.47	0.33	0.33	–	–
AGE	768	0	0.0	33.24	33.24	11.76	11.67	–	–

Please cite this article in press as: R. Ramezani et al., A novel hybrid intelligent system with missing value imputation for diabetes diagnosis, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2017.03.043

Table 4 Missing values patterns of diabetic patients.

No. of pattern	No. of patient	Attributes							
		DPF	AGE	PGL	BMI	DIA	NPG	TSF	INS
1	336								
2	1			🟥					
3	56						🟥		
4	1				🟥		🟥		
5	119								🟥
6	4			🟥					🟥
7	2					🟥			🟥
8	21						🟥		🟥
9	1				🟥		🟥		🟥
10	170							🟥	🟥
11	2				🟥			🟥	🟥
12	17					🟥		🟥	🟥
13	6				🟥	🟥		🟥	🟥
14	23						🟥	🟥	🟥
15	8					🟥	🟥	🟥	🟥
16	1				🟥	🟥	🟥	🟥	🟥

Type
□ Nonmissing
🟥 Missing

Table 5 Total variance explained by different components.

Component	Eigenvalue	Percent of variance	Cumulative %
1	2.60	32.56	32.56
2	1.43	17.84	50.40
3	1.21	15.15	65.55
4	.94	11.74	77.29
5	0.71	8.82	86.11
6	0.46	5.71	91.82
7	0.37	4.64	96.46
8	0.28	3.54	100.00

be used for making classifier model instead of eight attributes and the dimension is reduced from eight to five.

4.4. Discussion on experiment results of LANFIS intelligent system

The best classification accuracies of some different methods and the proposed method for Pima diabetes disease dataset are given in Table 7. The LANFIS intelligent system, LDA-ANFIS and ARTMAP-IC that are based on fuzzy inference systems, have the higher accuracy in comparison with other methods except for [24] and [25]. Therefore, the fuzzy inference systems are appropriate for modeling this dataset. Although these systems present significant results, there are two issues that have not been studied in existing systems: The first one is related to the output of ANFIS which is continuous while the response attribute in Pima diabetes disease dataset is binary. Hence, our novel hybrid classifier integrated Logistic regression and ANFIS to classify Pima Indians diabetes dataset. Therefore, the output of LANFIS model becomes binary.

Table 6 The coefficients of three linear transformations in OT.

	Coefficients of linear transformation			Important degree of attributes
	1	2	3	
NPG	0.517	−0.544	−0.339	0.678
PGL	0.671	−0.026	0.568	0.774
DIA	0.577	−0.161	−0.332	0.469
TSF	0.633	0.524	−0.326	0.782
INS	0.625	−0.030	0.627	0.785
BMI	0.625	0.597	−0.299	0.836
DPF	0.211	0.293	0.249	0.192
AGE	0.572	−0.622	−0.115	0.727

The second problem is the missing values in Pima diabetes disease dataset. This problem is very important and has not been considered by other researches such as MLNN with LM [15], ARTMAP-IC [12] and LDA-ANFIS [13]. We had done the missing value analysis in data preprocessing step and 3–5% increase in accuracy is obtained. According to the analysis of missing value patterns, 15 patients have a missing value in four or more attributes. These patients had little information in diabetic disease and they were eliminated. Therefore, the number of patients is reduced to 753 patients. Moreover, 417 patients have missing value in less of 4 attributes where the missing values imputed to predicted values.

However, in the fuzzy modeling, the number of fuzzy if-then rules exponentially increases by the number of attributes. Therefore, high dimensionality of dataset caused the "Out of memory" computational error in our programming computer. To manage this problem, we decrease the dimension of dataset in data preprocessing stage by OT method.

Table 7 Accuracies of LANFIS classifier system and previous methods.

Study	Method	Accuracy (%)
Deng and Kasabov [26]	ESOM (10× FC)	78.40
Polat et al. [2]	LS-SVM (10× FC)	78.21
	GDA–LS-SVM (10× FC)	79.16
Temurtas et al. [16]	MLNN with LM (10× FC)	79.62
	PNN (10× FC)	78.05
Kayaer and Yıldırım [20]	GRNN (conventional valid)	80.21
	MLNN with LM (conventional valid)	77.08
Carpenter and Markuzon [12]	ARTMAP-IC (conventional valid)	81.00
Temurtas et al. [16]	MLNN with LM (conventional valid)	82.37
	PNN (conventional valid)	78.13
Dogantekin et al. [13]	LDA-ANFIS (conventional valid)	84.61
Bozkurt et al. [27]	DTDN ((3× FC))	76.00
Yilmaz et al. [24]	Modified K-Means Clustering + SVM (10× FC)	96.71
Zhu et al. [25]	Multiple Factors Weighted Combination (5× FC)	93.00
Other studies reported. in Dogantekin et al. [13]	Various methods (3× FC, 10× FC, conventional valid)	Between 59.5 and 77.7
Our study	LANFIS (3× FC)	88.05

Table 8 The architecture and training parameters of LANFIS model used for diabetes in this study.

The number of layers	Input:	Rules number	Output	Range of influence	Squash factor	Accept ratio	Reject ratio
5	5	187	1	0.1–0.2	0.9–1.1	0.1–0.9	0.15

Table 9 Confusion matrix of LANFIS classifier.

	Health predicted	Diabetic diseases predicted
Health	141	18
Diabetic diseases	12	80

Table 10 Compare the obtained values of sensitivity and specificity using LANFIS intelligent diagnosis system for diabetes.

	Sensitivity (%)	Specificity (%)
LS-SVM (Polat et al., 2008) [2]	73.91	80
GDA–LS-SVM (Polat et al., 2008) [2]	79.16	83.33
LDA-ANFIS (Dogantekin et al., 2010) [13]	83.33	85.18
DTDN (Bozkurt et al., 2014) [25]	53.33	88.75
LANFIS (our study)	92.15	81.63

Finally, the parameters of the LANFIS classifier are given in Table 8. The fuzzy rules are generated by a subtractive clustering method. We obtained 88.05 percent accuracy by using

4.4.1. LANFIS

This accuracy confirms that this classifier is better than fuzzy classifiers in the available literature by deleting all samples to missing values. The confusion matrix using LANFIS intelligent system for 3-fold cross validation is given in Table 9. In Table 10, the obtained sensitivity and specificity values of proposed method for diagnosing diabetes disease are compared with other methods.

5. Conclusion

In this paper, we proposed a novel intelligent system for diagnosing diabetic diseases with missing values in clinical attributes. Our studies investigated the multiple imputation techniques for predicting missing values, orthogonal transformation techniques for reducing the dimension of input data and applying a novel hybrid classifier. The proposed classifier is combination of logistic model and adaptive network based on fuzzy inference system.

The high accuracy was obtained by LANFIS classifier and applying the subtractive clustering method to generating fuzzy rules. The best classification accuracy rate of the proposed intelligent diagnosis system is about 88.05% for diabetic data. The results of our examinations show that the proposed method significantly enhances the performance in comparison with related methods. Another advantage of our novel intelligent system is that it does not necessitate doctors to examine more tests.

References

[1] http://www.diabetes.org/about-diabetes.jsp (accessed 11.04.07).

[2] K. Polat, S. Gunes, A. Arslan, A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine, Expert Syst. Appl. 34 (2008) 482.

[3] http://www.fi.edu/biosci/healthy/diabetes.html (accessed 21.03.07).

[4] J.P. Medhi, S. Dandapat, An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images, Comput. Biol. Med. 74 (1) (2016) 30.

Please cite this article in press as: R. Ramezani et al., A novel hybrid intelligent system with missing value imputation for diabetes diagnosis, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2017.03.043

[5] P. Chowriappa, S. Dua, U.R. Acharya, M.M.R. Krishnan, Ensemble selection for feature-based classification of diabetic maculopathy images, Comput. Biol. Med. 43 (12) (2013) 2156.

[6] M.H.A. Fadzil, L.I. Izhar, H.A. Nugroho, Determination of foveal a vascular zone in diabetic retinopathy digital fundus images, Comput. Biol. Med. 40 (7) (2010) 657.

[7] D. Sidibé, I. Sadek, F. Mériaudeau, Discrimination of retinal images containing bright lesions using sparse coded features and SVM, Comput. Biol. Med. 62 (2015) 175.

[8] H.A. Elkayal, N.E. Ismail, M. Lotfy, Microwaves for breast cancer treatments, Alexandria Eng. J. 54 (4) (2015) 1105.

[9] E.E. Nithila, S.S. Kumar, Segmentation of lung nodule in CT data using active contour model and Fuzzy C-mean clustering, Alexandria Eng. J. 55 (3) (2016) 2583.

[10] M.U. Akram, S. Khalid, A. Tariq, S.A. Khan, F. Azam, Detection and classification of retinal lesions for grading of diabetic retinopathy, Comput. Biol. Med. 45 (2014) 161.

[11] J. Abawajy, A. Kelarev, M. Chowdhury, A. Stranieri, H.F. Jelinek, Predicting cardiac autonomic neuropathy category for diabetic data with missing values, Comput. Biol. Med. 43 (2013) 1328.

[12] G.A. Carpenter, N. Markuzon, ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases, Neural Networks 11 (1998) 323.

[13] E. Dogantekin, A. Dogantekin, D. Avci, L. Avci, An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS, Dig. Sig. Process. 20 (2010) 1248.

[14] E.D. Übeylı, İ. Güler, Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems, Comput. Biol. Med. 35 (5) (2005) 421.

[15] H.R. Marateba, M. Mansourianb, E. Faghihimanic, M. Aminic, D. Farinad, A hybrid intelligent system for diagnosing micro albuminuria in type 2 diabetes patients without having to measure urinary albumin, Comput. Biol. Med. 45 (2014) 34.

[16] H. Temurtas, N. Yumusak, F. Temurtas, A comparative study on diabetes disease diagnosis using neural networks, Expert Syst. Appl. 36 (2009) 8610.

[17] http://mlearn.ics.uci.edu/databases/pima-indians-diabetes/pima-indians-diabetes.names (accessed 11.04.07).

[18] J. Breault, Data mining diabetic databases: Are rough sets a useful addition?, in: Proc 33rd Symposium on the Interface, Computing Science and Statistics, Fairfax, VA, 2001.

[19] R. Pearson, The problem of disguised missing data, SIGKDD Explorations 8 (1) (2006) 83.

[20] K. Kayaer, T. Yıldırım, Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), 2003, p. 181.

[21] D.B. Rubin, Multiple Imputation for Non response in Surveys, Wiley, New York, 1987.

[22] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 2002.

[23] I. Jolliffe, Principal Component Analysis, Springer, Berlin, 1986.

[24] N. Yilmaz, O. Inan, M.S. Uzer, A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases, J. Med. Syst. 38 (2014) 48.

[25] J. Zhu, J.Q. Xie, K. Zheng, An improved early detection method of type-2 diabetes mellitus using multiple classifier system, Inf. Sci. 292 (2015) 1.

[26] D. Deng, N. Kasabov, On-line pattern analysis by evolving self-organizing maps, Neurocomputing 51 (2003) 87.

[27] M.R. Bozkurt, N. Yurtay, Z. Yilmaz, C. Sertkaya, Comparison of different methods for determining diabetes, Turkish J. Electr. Eng. Comput. Sci. 22 (2014) 1044.

Please cite this article in press as: R. Ramezani et al., A novel hybrid intelligent system with missing value imputation for diabetes diagnosis, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2017.03.043