International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 24323-24330

http://www.ripublication.com

Fuzzy Classifier for Diagnosis Pima Indians Diabetes

Juan Contreras[1], Laura Martínez[2], Yuliana Puerta[1,] Maria Claudia Bonfante[1]

[1]DepartmentofEngineering
CorporaciónUniversitariaRafael Núñez, Cartagena, Colombia.
epcontrerasj@ieee.org; puertacruz@gmail.com;
mariaclaudia.bonfante@curn.edu.co;
[2]DepartmentofEngineering
Fundación Tecnológica Antonio de Arévalo, Cartagena, Colombia
lauramargarcia@gmail.com

Abstract

Since 1998, there are countless research studies about the diagnosis of Diabetes Mellitus, based on information from the data set PIDD - Pima Indian Diabetic Database by the National Institutes of Diabetes and Digestive and Kidney Diseases. PIDD is a standard for analysis and accuracy in the diabetes diagnosis, applied to different algorithms with a hit rate from 66% to 82.6%. This article presents the application of a training algorithm that can improve accuracy and interpretability of data, through a fuzzy classifier to determine, if the patient shows signs of the disease according to the criteria of the World Health Organization (WHO).

Key Words: training algorithm, Diabetes mellitus, diabetes diagnosis, Pima Indians, fuzzy classifier.

1. Introduction

The diagnosis of a patient with Diabetes Mellitus (DM) is a process that requires a lot of knowledge by the person who apply it. The medical professional must interpret all the data collected, through the tests and analysis of the patient behavior. The handling of inaccurate information can show a high degree of uncertainty in the test. Therefore, the diagnosis requires a special study.

Many techniques are used for diagnosing diabetes. Zolfaghari proposed ensemble two machine learning method, Neural Network NN and support vector machine SVM, to predict the presence of diabetes. The approach was applied on PIDD and reported an accuracy of 88.04% [1]. Another approach by using adaptive neuro fuzzy inference system ANFIS to train the neural network adaptive group based

k nearest neighbor algorithm is proposed for diagnosis of diabetes. The experimental results of this approach applied on PIDD showed an accuracy of 80.0% [2].

An approach that uses Linear Discriminant Analysis and Support Vector Machine for classifying the samples is proposed. The approach was applied on PIDD and reported 76.6% for training accuracy and 75.65% for testing accuracy. 615 samples were used for training and 153 for testing [3].

The reasoning and fuzzy classification emerges as an alternative in the early recognition of the diabetes. This paper presents a fuzzy classifier to identify Diabetes Mellitus from experimental data input and output. The training of the fuzzy algorithm is conducted through the data set Pima Indians Diabetes, provided by the UCI Repository (Repository of Machine Learning Databases). The performance of the proposed algorithm is tested with a data set PIDD and compared with some well-known algorithms from the literature. The comparisons demonstrated that the proposed algorithm can be successfully applied on improving the degree of accuracy of the test.

2. Data Set Description

Considering the important role played by experts in the diagnosis of diabetes, this research proposes the design of a fuzzy classifier as an alternative diagnostic tool with a high degree of accuracy in the test.

For checking the performance of the fuzzy classifier algorithm a benchmark is used from the repository of machine learning UCI, through the data set PIDD, which has been under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases, with a high rate of incidence of diabetes [4]. The Pima Indian diabetes dataset, donated by Vincent Sigilli to, is a collection of medical diagnostic reports of female patients from the Pima Indian tribe of Arizona, equal or greater than 21 years old. The Pima Indians are known to be genetically predisposed to diabetes [5].

The data set contains 768 observations and 9 variables with no missing values reported. However, some researchers have pointed out that there are a number of impossible values, such as: five patients with a glucose of 0, eleven patients with a body mass index of 0, one hundred and ninety two with a skin fold thickness readings of 0, twenty eight with a diastolic blood pressure of 0 and one hundred and forty with a serum insulin levels of 0. Because of that there are only 392 cases with no missing values. [1],[6].

Table 1: Input variables with missing values

Variable	Criteria	Mean	Range
1	Number of times pregnant	3.8	[0,17]
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	120.9	[0,199]
3	Diastolic blood pressure (mm Hg)	69.1	[0,122]
4	Triceps skin fold thickness (mm)	20.5	[0,99]
5	2-Hour serum insulin (mu U/ml)	79.8	[0,846]
6	Body mass index (weight in kg/(height in m)^2)	32.0	[0,67.1]
7	Diabetes pedigree function	0.5	[0.078,2.42]
8	Age (years)	33.2	[21,81]

Table 2: Input variables with no missing values

Variable	Criteria	Mean	Range
1	Number of times pregnant	3.3	[0,17]
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	122.6	[56,198]
3	Diastolic blood pressure (mm Hg)	70.7	[24,110]
4	Triceps skin fold thickness (mm)	29.1	[7,63]
5	2-Hour serum insulin (mu U/ml)	156.1	[14,846]
6	Body mass index (weight in kg/(height in m)^2)	33.1	[18.2,67.1]
7	Diabetes pedigree function	0.5	[0.085,2.42]
8	Age (years)	30.9	[21,81]

Table 3: Set of classes

Class	Value	Number of cases
TESTED_NEGATIVE	0	262 (67%)
TESTED_POSITIVE	1	130 (33%)

The output variable is represented by the diagnosis with high levels of accuracy and interpretability in the data.

3. Fuzzy classifier

The fuzzy algorithm used in this study has been applied in identification and classification problems [7]-[9]. The universe of discourse of the numerical space of the inputs is partitioned using normalized triangular sets distributed symmetrically at each respective universe. An overlapping of 0.5 is required to guarantee that the supports of the fuzzy sets are different. Fuzzy singletons are used for the consequent membership functions.

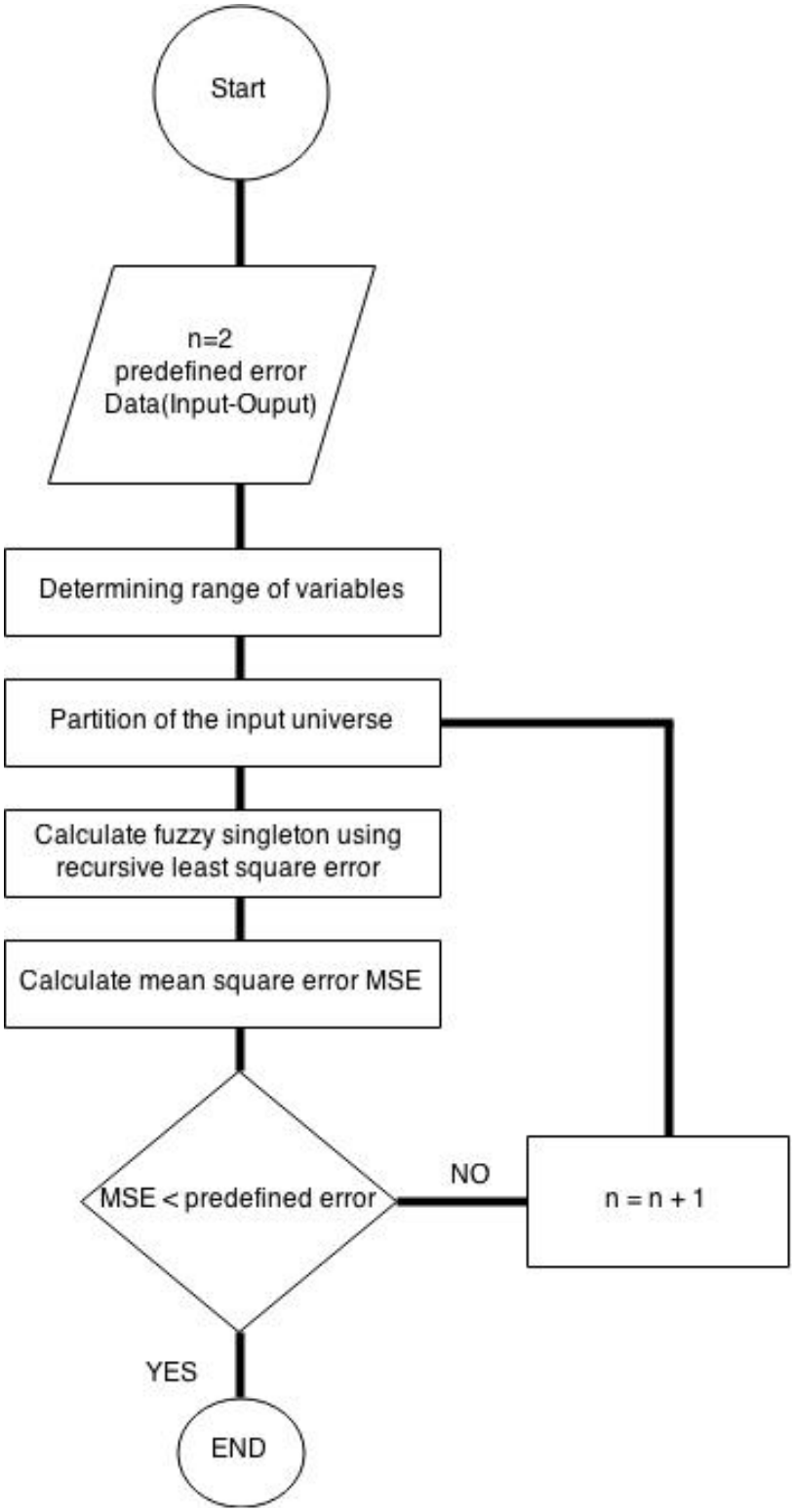

Figure 1. Describing the steps of fuzzy algorithm

4. Results

Initially, the algorithm is used to approximate the function to the real diagnostic in the data set, which are the subject of analysis. In each diagnosis is assigned a numeric value or class, considered as an array of data. The fuzzy model obtained should approximate as accurately as possible. The function can take the values 0 or 1.

Figure 2, shows the variation of the Mean Squared Error (MSE)between the output of fuzzy model and the real diagnosis. For training process 2 to 6 triangular membership functions for each input has been used.

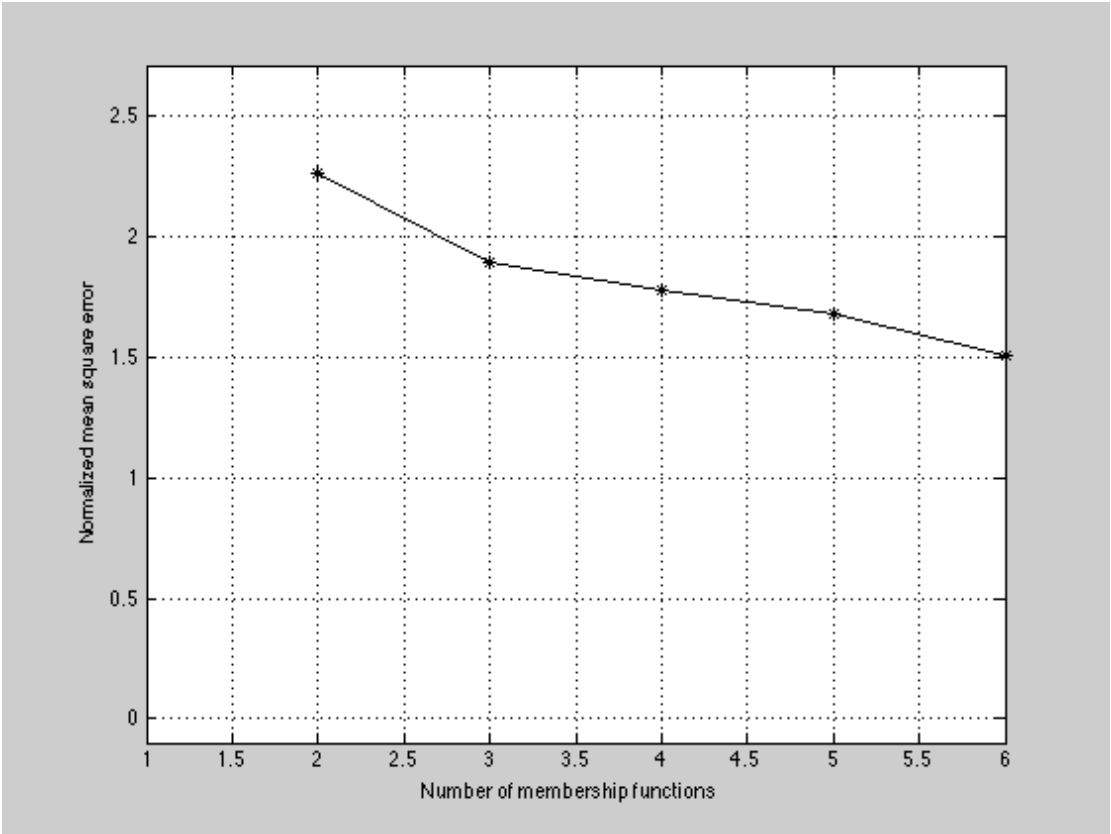

Figure 2. Mean Square Error between the output of fuzzy model to the real diagnosis and variations of the membership functions

The figure 3, shows a comparison between the diagnosis of the fuzzy model (o) and the real diagnosis (*) for both class.

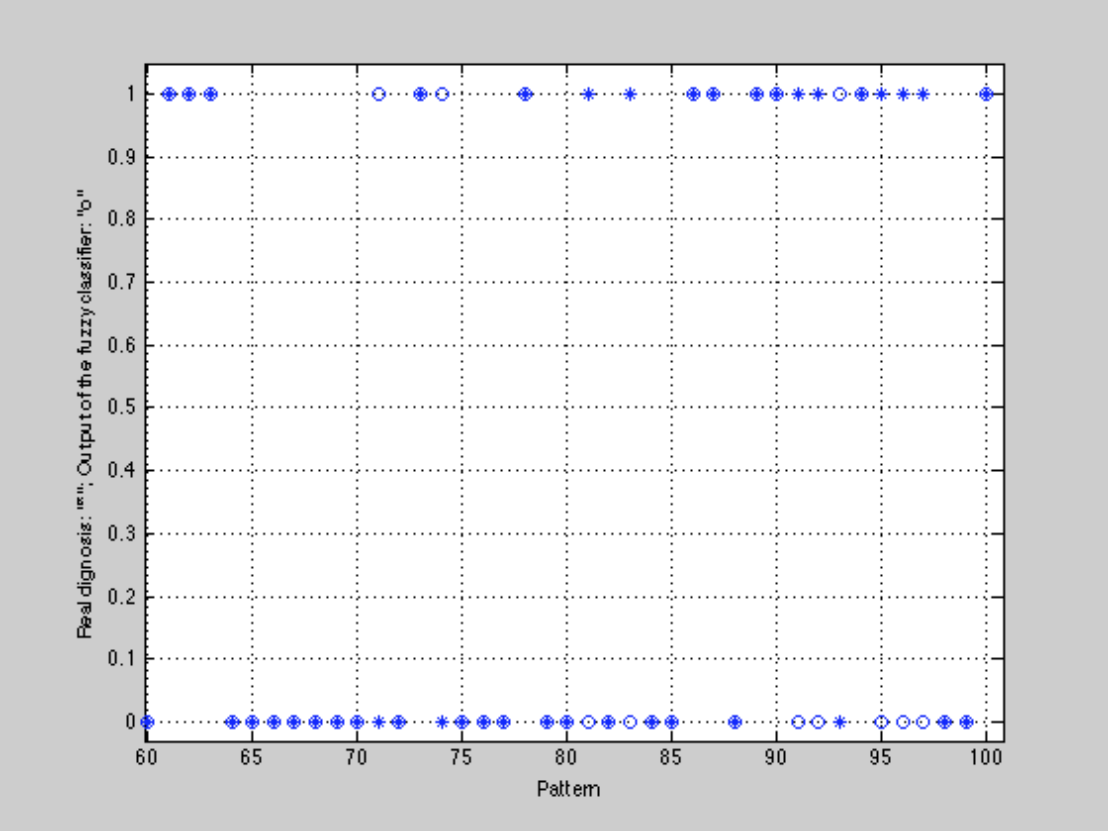

Figure 3. Comparison between diagnosis of fuzzy model (o) and the real diagnosis from the expert(*)

The resulting fuzzy classifier has the following specifications: six normalized triangular fuzzy sets distributed symmetrically at each respective universe of input (for each one of eight different inputs) and 48 fuzzy singleton consequents.The results show81.4% for training accuracy and 80.5% for testing accuracy.

Table 4: Studiesconductedwiththe data set PIDD in the diagnosis of diabetes

Author(s)	Techniques	Accuracy
Zolfaghari [1]	NN-SVM	88.04%
Karthikeyani and Begum [10]	PLS-LDA	74.4%
Seeraand Lim [11]	FMM FMM-CART FMM-CART-RF	69.28% 71.35% 78.39%
Kalaiselvi and Nassira [2]	ANFIS-NNA	80.0%
Parashar et al [3]	LDA-SVM	75.65%
Our approach	**Fuzzy Logic**	**80.5%**

5. Conclusions

The present paper proposes asingle and effective fuzzy classifier algorithm which has been tested on one of the widely studied problems in bioinformatics: Diabetes diagnosis. The experimental results of the proposed approach show that classification accuracy is one of the better existing approaches: 81.4% for training accuracy and 80.5% for testing accuracy.

References

[1] RahmatZolfaghari (2012). "Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM".International Journal of Computational Engineering & Management, Vol. 15, No 4, p.p.: 115-121

[2] C. Kalaiselvi and G.M. Nasira (2014). "A Novel Approach for the Diagnosis of Diabetes and Liver Cancer using ANFIS and Improved KNN". Research Journal of Applied Sciences, Engineering and Technology, Vol. 8, No. 2, p.p.: 243-250

[3] A. Parashar, K. Burse, K. Rawat (2014)"A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network". International Journal of Advanced Research in Computer Science and Software Engineering.Vol 4, No.11. p.p.: 378-383

[4] W. C. Knowler, P. H. Bennett, R. F. Hamman, and M. Miller (1978). "Diabetes incidence and prevalence in Pima Indians: A 19fold greater incidence than in Rochester, Minnesota". American Journal of Epidemiology, 108(6), 497–505.

[5] Hanson, R. L., M. G. Ehm, 1998. "An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians." Am J Hum Genet 63(4): 1130-1138.

[6] R. M. Rahman, F.Afroz(2013). "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis".Journal of Software Engineering and Applications.Vol. 6, p.p.: 85-97

[7] J. Contreras, "Generating Fuzzy Controllers for Ship Steering". 30th North American Fuzzy Information Processing Society Annual Conference, El Paso, Texas, USA. March 18-20, 2011.

[8] J. Contreras. "Generating Singleton Fuzzy Models from Data with Interpretable Partition". Advanced Material Research. Vol 629, p.p. 784-791, 2012

[9] J. Contreras, M.C. Bonfante, A. Quintana, V. Castro ."Fuzzy Classifier for the Diagnosis of Pathology on the Vertebral Column".IEEE Latin America Transactions. Vol. 12, No. 6, pp. 1149 - 1154. 2014.

[10] V.Karthikeyani, I. P. Begum (2013). "Comparison a Performance of Data Mining Algorithms (CPDMA) inPrediction of Diabetes Disease",

"International Journal on Computer Science and Engineering", Vol.5, No.3 pp.205-210.

[11] M.Seera, Ch. P. Lim (2014)."A hybrid intelligent system for medical data classification", Expert Systems with Applications. Vol. 41, No. 5, p.p.: 2239-2249.