Interpretable Classifier of Diabetes Disease

Nesma Settouti, Meryem Saidi, and Mohamed Amine Chikh

***Abstract*—Interpretability represents the most important driving force behind the implementation of fuzzy-based classifiers for medical application problems. The expert should be able to understand the classifier and to evaluate its results. The main purposes in this work is the application of a new method based on FCM and ANFIS to diagnose the diabetes diseases by using a reduced number of fuzzy rules with relatively small number of linguistic labels, removing the similarity of the membership functions, preserving the meaning of the linguistic labels (interpretability), and in same time improving the classification performances. Experimental results show that the proposed approach FCM-ANFIS can get high accuracy with fewer rules. On the contrary, by using ANFIS more rules are needed to get a lower accuracy. Moreover the features projected partition in ANFIS is ambiguous and cannot preserve the meaning of the linguistic labels. The best number of the rules is a trade-off between the accuracy and the rules number, also with a minimum of clusters (c=2) and just two fuzzy rules, FCM-ANFIS approach has given the best results with CC = 83.85%, Se = 82.05% and Sp = 84.62% comparing to the other cases.**

***Index Terms*—Interpretable classification; fuzzy rules; FCM; neuro-fuzzy ANFIS; UCI machine learning database.**

I. Introduction

Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. Hyperglycemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. Diabetes is the most rapidly growing chronic disease of our time. It has become an epidemic that affects a large number of people in the world [1].

Diagnose of diabetes for medical expert is a difficult task. For this reason a much research effort has been put till today in diagnosis of diabetes disease literature. Actually fuzzy logic and neural networks have provided attractive structures to complex systems. Adaptive network-based fuzzy inference system (ANFIS) is a specific approach in neuro-fuzzy modeling which utilizes the neural networks to tune the rule-based fuzzy systems [2]. Successful applications of ANFIS in biomedical engineering have been reported recently in data identification [3], [4], and pattern recognition [5].

The exponential increasing number of fuzzy rules in neuro-fuzzy classifier is one of the difficult problems to be overcome in this paper, such problem is solved by adopting the FCM clustering method [6] for the structure identification in the ANFIS, wish is one of the most widely used neuro-fuzzy models proposed by Jang [2], against diabetes diagnosis problem. The algorithm fuzzy c-means (FCM) is a typical clustering algorithm, which was used in a wide variety of engineering and scientific disciplines such as modeling [7] the decision [8], pattern recognition and classification [9], segmentation [10]. Recently the classification with the approach FCM-ANFIS [11] was applied on database of diabetes collected by the Faculty of Computer Science and Information, University of Technology Malaysia; the results have reached 72.66%.

The remainder of the paper is organized as follows, the Adaptive Neuro-Fuzzy Inference System (ANFIS) is proposed based on the model of Takagi-Sugeno. Pima Indian Database and neuro-fuzzy Classifier learning are presented in Section 3. In section 4, the results are presented and discussed. Finally, section 5 concludes the findings.

Manuscript received March 10, 2012; revised May 5, 2012.

Nesma Settouti is with the Biomedical Engineering Laboratory, Tlemcen University –Algeria (Tel: +213 055 610 476 39; fax: +213 043285686; e-mail address: nesma.settouti@gmail.com).

Meryem Saidi is with the Biomedical Engineering Laboratory, Tlemcen University – Algeria.

Mohamed Amine Chikh is with the Tlemcen University-Algeria. Actually he is the head of CREDOM research team at Biomedical Engineering Laboratory.

II. Theory

A. Fuzzy C-Means Algorithm

The fuzzy c-means (FCM) clustering algorithm was first introduced by Dunn [12] and later extended by Bezdek [6]. It is based on the concept of fuzzy c-partition, introduced by Ruspini [13]. FCM is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , 1 \le m < \infty \quad (1)$$

where

- m is any real number greater than 1, it was set to 2.00 by Bezdek.
- u_{ij} is the degree of membership of x_i in the cluster j;
- x_i is the ith of d-dimensional measured data;
- c_j is the d-dimension center of the cluster,
- ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the c_j cluster centers by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (3)$$

This iteration will stop when $\max_{ij} \{ | u_{ij}^{(k+1)} - u_{ij}^{(k)} | \} \prec \varepsilon$ where ε is a termination criterion between 0 and 1 and k the iteration steps are. This procedure converges to a local minimum or a saddle point of J_m.

B. Adaptive Neuro-Fuzzy Inference System

A hybrid system named ANFIS (Adaptive-Network-Based Fuzzy Inference System or Adaptive Neuro-Fuzzy Inference System) has been proposed by Jang in [8].

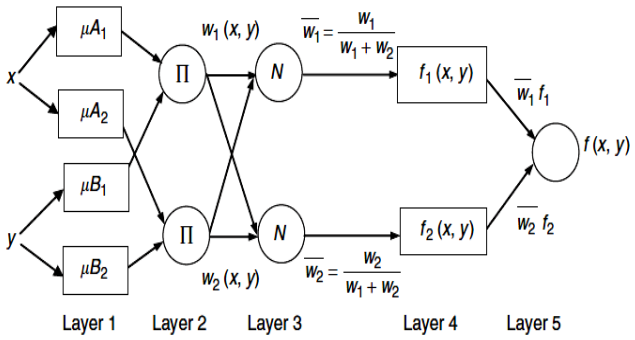

Fig. 1. ANFIS architecture

The ANFIS is a fuzzy inference system (FIS) based on the model of Takagi-Sugeno (TKS) [14]. It is the fuzzy-logic based paradigm that grasps the learning abilities of neural networks to enhance the intelligent system's performance using a priori knowledge. The ANFIS architecture that allows representing the basic rules is carried out by an adaptive network that contains fixed nodes (circular) and adaptive nodes (square) as illustrated in Fig.1.

Each node square or circular applies a function on its input signals and for a given layer nodes have the same type of function. The generalization of the system to a system with multiple inputs does not pose any problem. The number of nodes in the first layer is equal to the total number of linguistic terms defined.

III. Experimentation

A. Diabetes Disease Database

We have used Pima Indian Dataset taken from UCI machine learning repository [15] in our applications. This dataset is commonly used among researchers who use machine learning method for diabetes disease classification. The database contains 768 records and two classes. The class distribution is: Class 1: non- diabetics (500) (65.1%) Class 2: diabetics (268) (34.9%). Each record has eight attributs:

1) Npreg: Number of time pregnant
2) Glu: Plasma glucose concentration a 2 h in oral glucose tolerance test
3) BP: Diastolic blood pressure (mm Hg)
4) Skin: Triceps skin fold thickness (mm)
5) Insulin: 2-h serum insulin (μU mL−1)
6) BMI: Body mass index (weight in kg/(height in m)2)
7) Ped: Diabetes pedigree function
8) Age: (years)

B. Neuro-Fuzzy Classifier Learning

There are several ways that structure learning and parameter learning can be combined in a neuro-fuzzy model classifier. They can be performed sequentially: structure learning is used first to find the appropriate structure of a neuro-fuzzy system; and parameter learning is then used to fine-tune the parameters. In other cases, only parameter learning or structure learning is necessary when structure (fuzzy rules) or parameters (membership functions) are given by human experts, also the structure in some neuro-fuzzy model [2] is fixed a priori.

1) Structure learning phase

Structure learning is a more difficult task than parameter learning, one of the most known methods for structure initialization is uniform partitioning of each input variable range into fuzzy sets, resulting to a fuzzy grid. This approach is followed in ANFIS; a well-known TSK model algorithm in this study is accomplished by using the Silhouette validation technique [16] which calculates the silhouette width for each sample, average silhouette width for each cluster and overall average silhouette width for a total data set. Using this approach each cluster could be represented by so-called silhouette, which is based on the comparison of its tightness and separation. The average silhouette width could be applied for evaluation of clustering validity and also could be used to decide how good the number of selected clusters is. The largest overall average silhouette was obtained for c=2 with 0.44, then 0.38 for c=3. That indicates the best clustering is equal to 2 (number of cluster). Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters.

2) Parameter learning phase

Hybridization FCM-ANFIS

This work concentrates on optimization of diabetes disease ANFIS-based fuzzy classifier. We propose to use FCM clustering strategy. The data clustering is to identify structure based on the scatter partition. So, this approach can be adopted to reduce the dimension of the classifier as well as training time since the number of fuzzy rules equals the number of membership functions regardless of the dimension inputs. FCM method is used to predict the distribution of fuzzy membership functions by universe of discourse partition. FCM tries to partition numerical data into clusters. The belongingness of a data point to a specific cluster is given by the membership value of the data point to that cluster. The membership value is calculated by minimization of a FCM function which emerging research affiliation with

the least error.

The function needs approximate cluster centers, as well as a metric for membership evaluation as input, e.g., the Euclidean distance. The minimization is an iterative process where new cluster centers are computed as weighted averages of all data points, where the membership values are the weights. The stop criterion is when the membership values have changed very little in comparison with the previous iteration. It can be shown in Fig.2 the repartition of two cluster centers obtained for the height feature inputs of the data base Indians PIMA.

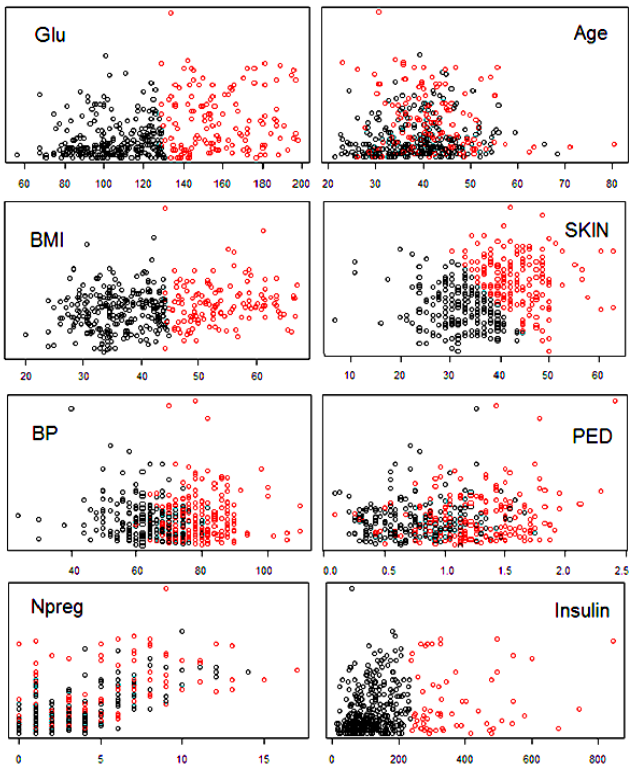

Fig. 2. Clustering results with 2 clusters

3) Knowledge extracted from the FCM-ANFIS model

The FCM-ANFIS classifier will thus generate a knowledge base of 2 rules for classification.

*Rule 1: If (Npreg is **High**) and (Glu is **High**) and (BP is **High**) and (Skin is **Medium**) and (Insulin is **High**) and (BMI is **High**) and (PED is **High**) and (Age is **High**) then (Class is Diabetic)*

*Rule 2: If (Npreg is **Medium**) and (Glu is **Normal**) and (BP is **Medium**) and (Skin is **High**) and (Insulin is **Low**) and (BMI is **Medium**) and (PED is **Low**) and (Age is **Medium**) then (Class is **Non Diabetic**)*

In this work, each input attribute has two membership functions, so 256 fuzzy rules have been extracted logically, the overall complexity of the knowledge base of the fuzzy classifier increases substantially. To reduce this complexity, application of fuzzy c-means algorithm is proposed. By applying this technique, the number of rules has been reduced from 256 to 2, thereby reducing the complexity of the knowledge base significantly.

C. Results and Discussion

The performances of the implemented classifier were evaluated by computing the percentages of sensitivity (SE), specificity (SP) and correct classification (CC), and TP (classifies Diabetic as Diabetic), TN (classifies No Diabetic as No Diabetic), FP (classifies No Diabetic as Diabetic) and FN (classifies Diabetic as No Diabetic). After learning phase, 392 testing data (262 non-diabetics, 130 diabetics) with 3-fold cross validation were used to evaluate the classifier (see TABLE.I).

TABLE I: PERFORMANCES OF FCM-ANFIS CLASSIFIER WITH DIFFERENT CLUSTERS

Cluster number	8 attributs (ANFIS)	C=2	C=3	C=4
CC %	73.08	83.85	83.08	81.54
Se %	66.67	82.05	69.23	82.05
Sp %	75.82	84.62	89.01	81.32
FP	62	42	31	45
TN	200	220	231	217
TP	94	109	97	105
FN	36	21	33	25
Number of rules	256	2	3	4

TABLE I indicates that by using FCM-ANFIS we can get high accuracy with fewer rules. On the contrary, by using ANFIS more rules are needed to get a lower accuracy. Moreover the features projected partition in ANFIS is ambiguous and cannot preserve the meaning of the linguistic labels. The best number of the rules is a trade-off between the accuracy and the rules number, moreover with a minimum of clusters (c=2) and just two fuzzy rules, FCM-ANFIS approach has given the best results with correct rate = 83.85%, Se = 82.05% and Sp = 84.62% comparing to the other case, giving a small number of misclassified cases (FP = 42 and FN = 21).

1) Patient number 107 is correctly classified as diabetic case (TP)

The TABLE II presents the characteristics of a diabetic patient; this example is correctly classified as diabetic with an apparent readiness to heredity (PED) and high values of glucose and insulin respectively.

TABLE II: PARAMETERS OF PATIENT NUMBER 107

Patient Number	**107**
Npreg	3
Glu	173
BP	78
SKIN	39
Insulin	185
BMI	33.8
PED	0.97
Age	31
Class	1

2) Patient number 28 is correctly classified as non-diabetic case (TN)

The TABLE III presents the characteristics of a non-diabetic patient; this example is correctly classified as non-diabetic with a normal value of glucose and low quantity of insulin. The fuzzy rule R2 is activated with a degree of fulfilment equals to 71.25%.

TABLE III: PARAMETERS OF PATIENT NUMBER 28

Patient Number	**28**
Npreg	1
Glu	87

BP	68
SKIN	34
Insulin	77
BMI	37
PED	0.401
Age	24
Class	0

3) Patient number 102 (diabetic) recognized as non-diabetic (FN)

The TABLE VI presents the characteristics of a diabetic patient; this example is misclassified, i.e. recognized as non-diabetic with a normal value of glucose and low quantity of insulin. The fuzzy rule R2 is activated with a degree of fulfilment around 74%.

TABLE IV: PARAMETERS OF PATIENT NUMBER 102

Patient Number	**102**
Npreg	2
Glu	93
BP	78
SKIN	64
Insulin	32
BMI	38
PED	0.674
Age	23
Class	1

4) Patient number 68 (no. diabetic) recognized as diabetic (FP)

The TABLE V presents the characteristics of a non-diabetic patient; this example is misclassified, i.e. recognized as diabetic (false alarm) with high value of plasma glucose concentration. The fuzzy rule R_1 is activated with a degree of fulfilment equals to 60.71%.

TABLE V: PARAMETERS OF PATIENT NUMBER 68

Patient Number	**68**
Npreg	2
Glu	157
BP	74
SKIN	35
Insulin	440
BMI	39.4
PED	0.134
Age	30
Class	0

As it could be noticed from these examples, fuzzy rules are solicited, according to input vector parameters, and then the fuzzy logic inference engine evaluates these inputs using the knowledge base and then diagnoses the patient class. These solicited fuzzy rules are closer to the medical expert reasoning. The neuro-fuzzy classifier justifies its results by the most solicited fuzzy rules. So the expert can easily check the fuzzy model classifier for plausibility, and can verify why a certain classification result was obtained for a certain patient, by checking the degree of fulfilment of the individual fuzzy rules. Instead of a simple classification the physician also gets a description of the patients in terms of the fuzzy rules that are active for this case.

Experimental results have shown that the proposed approach with Fuzzy c-means (FCM) is simple and effective in explaining the interpretability of neuro-fuzzy classifier by reducing the 256 rules to 2 efficient rules while preserving the model accuracy at a satisfactory level. Many studies using ANFIS structure have been performed in diagnosis of diabetes disease. Polat and Gunes [17] have reported 89.47% classification accuracy using principle components analysis (PCA) and ANFIS. But when principle components are applied for fuzzy rule-based approaches, interpretation of produced rules naturally becomes more difficult or even impossible. Vosoulipour et al. [18] have applied four features witch selected by a genetic algorithm (GA), to a neural network and second to an ANFIS structure, they have reported respectively 77.60% and 81.30% classification accuracy. But the obtained results are not interpretable. The detail accuracy of these studies can be seen in [17].

The results obtained in this paper are very interesting than the cited studies in literature, since our neuro-fuzzy classifier built in this work aims at finding with Fuzzy c-means a reduced set of fuzzy rules that can be interpreted linguistically. The clinician obtains information about why the diagnosis was selected.

IV. CONCLUSION

This study presents a fuzzy classification model of diabetes. Here, two criteria are used to evaluate the proposed method. The first criterion is the good accuracy performances of the classifier and the other is comprehensibility of obtained results. This study shows that the contribution of the Fuzzy C-Means (FCM) algorithm on ANFIS has reduced the size of models as well as learning time, since the number of centre's is equal to the number of membership functions regardless of the size of incoming. With the hybridization of these methods extracting of knowledge has been made to provide rules that are reliable, accurate and sufficiently simple to be understood; all in improving performance with a rate reaching 83.85%.

REFERENCES

[1] Medical Dictionary. O. Diabetes. [Online] Available: http://medicaldictionary.thefreedictionary.com/diabetes.

[2] J. S. R Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, pp. 665–685. 1993.

[3] T. Takagi and M. Sugeno, "Fuzzy Identification of Systems and Its Applications to Modeling and Control," *IEEE Transactions on Systems, Man, and Cybernetics*, pp.116–132. 1985

[4] Z. Heydari, F. Farahmand, H. Arabalibeik, and M. Parnianpour, "Adaptive neuro-fuzzy inference system for classification of acl-ruptured knees using arthrometric data," *Annals of Biomedical Engineering* vol. 36, no. 9, pp.1449-1457, 2008.

[5] S. Y. Belal, A. F. G. Taktak, S. A. S. Andrew John Nevill, D. Roden, and S. Bevan, "Automatic detection of distorted plethysmogram pulses in neonates and pediatric patients using an adaptive-network-based fuzzy inference system," *Artificial Intelligence in Medicine*, vol. 24, pp.149–165, 2002.

[6] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers Norwell, MA, USA .1981.

[7] Y. Miao and Z.Q Liu, "On causal inference in fuzzy cognitive maps," *IEEE Trans. Fuzzy Syst.* vol. 8, no.1, pp.107–119, Feb. 2000.

[8] Z. Q Liu and R. Satur, "Contextual fuzzy cognitive map for decision support in geographic information systems," *Fuzzy Systems, IEEE Transactions*, vol. 7, no. 5, pp. 495 – 507. 1999
[9] X. Chang, W. Li, and J. Farrell, "A c-means clustering based fuzzy modeling method," in *Proc the Ninth IEEE International Conference. Fuzzy Systems, 2000. FUZZ IEEE 2000 in San Antonio, TX 7-10* May 2000, vol. 2, pp. 937 - 940.
[10] S. R. Kannan, "A new segmentation system for brain MR Images based on fuzzy techniques," *Appl. Soft Comput*, vol. 8, pp. 1599–1606. 2008.
[11] M. F. B Othman and M.S.Y Thomas, "Neuro fuzzy classification and detection technique for bioinformatics problems," in *Proceedings of the First Asia International Conference on Modelling and Simulation*, 2007, pp. 375–380.
[12] J. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters," *Journal of Cybernetics*, vol. 3, pp.32–57, 1974.
[13] E. H. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, pp. 22–32. 1969
[14] M. Sugeno and G. T. Kang, "Structure identification of fuzzy model," *Fuzzy Sets Syst*, vol. 28, pp. 15–33. 1988.
[15] A. Frank and A. Asuncion, "UCI Machine Learning Repository Irvine," CA: University of California, School of Information and Computer Science.
[16] P. J. Rousseeuw, E. Trauwaert, and L. Kaufman, "Fuzzy clustering with high contrast," *Journal. Comput. Appl. Math*. vol. 64, pp. 81–90, Nov 1995.
[17] K. Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digit. Signal Process.* vol. 17, pp. 702–710, 2007.
[18] M. Vosoulipour, A. Teshnehlab, and H. A. Moghadam, "Classification on diabetes mellitus data-set based-on artificial neural networks and anfis," *4th KUALA LAMPUR international conference on biomedical engineering 2008. IFMBE proceedings*, 2008, vol. 21, part 3, part 1, pp. 27-30.

Nesma Settouti received her Engineer degree in Electrical Biomedical from the Tlemcen University, Algeria in 2009. In 2011 she obtains a magister degree in the same option. And is currently pursuing her PhD Thesis in the Biomedical Engineering Laboratory, her research interests have been in computer assisted medical decision support systems, neural networks, clustering methods, optimization, classification and artificial intelligence.

Meryem Saidi graduated from the Department of Computer Science at Tlemcen University, Algeria in 2009 with an Engineer degree and obtained a Magister degree in the same department in 2011. Her research interests are in the areas of artificial intelligence and Automatic diagnosis of disease. She actually purchases her Ph.D in artificial intelligence.

Mohamed Amine Chikh is graduated from The Electrical Engineering Institut (INELEC) of Boumerdes –Algeria in 1985 with an Engineering degree in Computer science and in 1992 with a Magister of Electronic from Tlemcen University. He also received a Ph.D in electrical engineering from the University of Tlemcen (Algeria) and INSA of Rennes (France) in 2005. And is currently Professor at Tlemcen University-Algeria. Actually he is the head of CREDOM research team at Biomedical Engineering Laboratory. He conducted post-doctoral teaching and research at the University of Tlemcen. Dr Chikh has published over 90 journal and conference papers to date and is involved in a variety of funded research projects related to biomedical engineering. His is a member of several scientific conferences. His research interests have been in artificial intelligence, machine learning, medical data classification, computer assisted medical decision support systems.