Diabetes Diagnosis Expert System Using Fuzzy Inference Methods

Akshay Parekh
Roll No - 166101008

Jyoti Prakash Mohanta
Roll No - 166101012

Ajinkya Sanjay Mankar
Roll No - 164101059

Mangirish Kenkre
Roll No - 164101051

Abstract—**Diabetes is one of the most common and hazardous disease, which can affect almost every organ of the body. Diagnosis of diabetes requires determining all medical data related to the disease. However, the nature of the data is very uncertain, which can affect the disease diagnosis. This project presents three fuzzy based inference mechanisms, Template based, Neuro-Fuzzy and Fuzzy C-Means Clustering for efficient diagnosis of diabetes. Medical and clinical real data from PIMA Indian Diabetes Dataset is used to develop and test the system. And Finally, Efficiency of all 3 methods have been calculated.**

1. Introduction

Diabetes has affected millions of people in the world, and hundreds of thousands of people across the globe die every year. This rapid rise of disease is a cause of great concern and requires efficient diagnosis system. The manual diagnosis cannot be much relied on, hence there is a need of an efficient automated system.

In this project we have tried to study and implement *Fuzzy Inference Mechanisms* and achieved an accuracy of around 75%. The approaches that we have studied are : *1. Template Based Approach*, *2. Neuro-Fuzzy Approach* and *3. Fuzzy C-Means clustering*. The Dataset that we have used for implementation of all 3 approaches is, PIMA Indian Diabetes Dataset [9].

This report is divided into 6 sections including Introduction. Next section is Literature Survey, which discuss details and precious works in details. Then third section is Methodology, where implementation steps are explained. Fourth Section is related to the description of Dataset. The last 2 section discusses about results and conclusion. and Finally the references.

2. Literature Survey

2.1. Diabetes

Diabetes results from reduced production of insulin, resistance to the effect of insulin or both of these. This results in abnormally high glucose levels in the blood and also widespread disturbances to metabolism. So it is a chronic disease resulting from inability to produce or reduced sensitivity to insulin. The pancreas produce insulin to regulate the levels of glucose in the blood which acts to trigger the liver to store glucose as glycogen, cells are encouraged to take up the glucose and prevent them from releasing protein and fats as energy. Diabetes is a major cause of death and disability. On average, the life expectancy is reduced by 20 years in Type 1 DM, and up to 10 years in Type 2 DM. Some of the common complications of diabetes are Neuropathy, Retinopathy and Nephropathy.

Diabetes Mellitus(DM or Diabetes) can be mainly classified into 2 types. They are Type 1 DM and Type 2 DM. In Type 1 DM, the pancreas is unable to produce insulin. It is diagnosed mainly in children or young adults. In Type1 DM, the symptoms develop quickly. Symptoms of Type 1 DM are frequent and excessive urination, dehydration, thirst, feeling of tiredness, blurred vision, urinary infections, loss of weight. The most important risk factor in case of Type 1 DM is genetic factor. Type 2 DM is diagnosed generally in older adults. It can be seen in younger age groups as well. It is characterized by insulin resistance, but it may also have deficiency. In Type 2 the symptoms develop gradually. It has symptoms similar to Type 1. Risk factors of Type 2 can be classified as modifiable risk factors and population risk factors. Modifiable risk factors are obesity, lack of exercise, smoking.

The purpose of this project is to develop a fuzzy based diagnosis system for Diabetes. PIMA Indian Diabetes Dataset has been used for developing our model. The features that we are using in our project are Age, Plasma Glucose Concentration, Diastolic Blood Pressure, Insulin, BMI and Diabetes Pedigree Function.

2.2. Previous Works

Over the years, many researchers have came up with different approach to make the diagnosis of Diabetes better using different Artificial Intelligence techniques. Some of the related works in the literature are following :

1) *Diabetes Mellitus forecast using Artificial Neural Network(ANN)* [2]. Authors in this paper used back-propagation algorithm for learning and testing. This work helps to determine whether someone is suffering from diabetes or not without performing a blood test.

2) *Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree* [3]. This paper applies decision trees for analyzing the risk of diabetes. It

doesnt consider patients from younger age groups.

3) *Diagnosis of Diabetes using OLAP and Data Mining Integration* [4]. This paper combines OLAP and data mining methods to determine whether the probability of a person being diagnosed with diabetes is low, medium or high.

4) *Medical Diagnosis on Pima Indian Diabetes using General Regression Neural Networks*. [5] This paper examines the general regression neural network (GRNN) on the Pima diabetes dataset for classification purpose whether a person is diabetic or not and compares its performance with multi-layer perceptron (MLP) and radial basis function(RBF) feed forward neural nets.

3. Methodology

The Dataset that we are using does not provide values with any fixed probability or crisp boundary set, therefore we need to rely on some computation paradigm that allows flexibility, the way like our brain thinks. One such paradigm is Fuzzy Sets and Fuzzy logic, which considers the world in imprecise terms and responds with precise actions. Fuzzy Inference is the process that uses fuzzy logic to map given inputs to an output. Fuzzy Based Inference Method that we are considering for our Diabetes Diagnostics system are:

1) Template Based Approach
2) Neuro-Fuzzy Based Approach
3) Fuzzy C-Means Clustering Approach

3.1. Template Based Approach

Template based approach partitions antecedent parameters into certain number of membership functions [6]. Following Steps are used for Rule generation :

- Partition the input and output parameters into fuzzy sets.
    1) We define domain intervals for input output variable considering the minimum and maximum values of parameter.
    2) Each domain interval is divided into three equal partitions.
    3) We used trapezoidal membership function for each fuzzy sets assigned to each partition.
- Generate Primary rule set
    1) First we determine degree of each input and output data in all the partitions and corresponding fuzzy sets having maximum degree.
    2) For each input output pair one rule is generated.
- Assign degree to each rule.
    1) In previous state we generating rule for data, we may get conflicting rules.
    2) In such cases conflicting rules are removed by assigning degree to each generated rule.
    3) Degree of rule can be represented as product of its components and degree of training example that generated this rule
- Obtain the final set of rules from preliminary set
    1) Rule with highest degree are chosen for each combination of antecedents.
- Using this approach we generated all the rules from the dataset. An example of the rule is given bellow:

  If Age is M AND Glucose is M AND Blood-Pressure is L AND Insulin is L AND BMI is H AND DPF is L Then Diabetes-Status is 0.

3.2. Neuro-Fuzzy Based Approach

Neuro-Fuzzy approach uses neural network for learning fuzzy sets from the input data. ANFIS (Adaptive Network Based Fuzzy System) is a popular neuro-fuzzy system used for parameter learning in fuzzy inference model [7]. The ANFIS structure consists of TSK type of rules. In ANFIS, parameters of the nodes are adopted keeping the structure fixed.

ANFIS consists of five layers as follows:

- **Layer 1:** Nodes belongs to this layer defines fuzzy-set in terms of Gaussian membership function. The output of this layer is degree of membership of a given input.
- **Layer2:** Firing strength of a rule is output of this layer.
- **Layer 3:** In this layer the nodes normalizes the firing strength of a rule.
- **Layer 4:** The nodes in this layer calculate weighted output form each rule.
- **Layer 5:** Finally node in layer five determine the final output.

The detail structure of ANFIS is presented in figure 1. We classify the input parameter values in to three fuzzy sets (low, medium and high) and give input to the ANFIS model. We train the ANFIS model using back-propagation algorithm to obtain the model parameters.

3.3. Fuzzy C-Means Clustering

Fuzzy C-Means Clustering algorithm is a method of clustering which allows one piece of data to belong to two

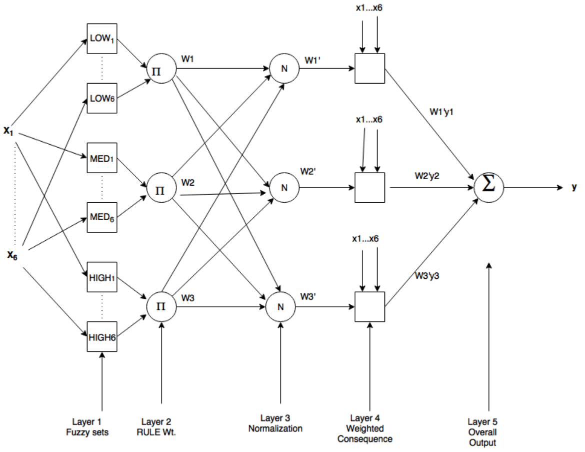

Figure 1. ANFIS Model used in this work.

or more clusters [8]. It is based on minimization of the following objective function:

$$J_m = \sum_{n=1}^{N} u_{ij}^m ||x_i - c_j||^2, 1 \leq m \leq \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith d-dimensional measured data, c_j is the d-dimension center of the cluster, and $|| * ||$ is any norm expressing the similarity between any measured data and the center.

The algorithm is composed of the following steps:

1) Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$
2) At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$
   $c_j = \sum_{n=1}^{N} u_{ij}^m * x_i / \sum_{n=1}^{N} u_{ij}^m$
3) Update $U^{(k)}, U^{(k+1)}$
   $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_i - c_j||}{||x_i - c_j||} \frac{2}{m-1} \right)}$
4) If $||U^{(k+1)} - U^{(k)}|| < \epsilon$ then STOP; otherwise return to step 2.

By using Fuzzy C-Means Clustering algorithm we obtain two clusters. One cluster corresponds to diabetes and other is for no diabetes. Then we compare the model output with the original data.

4. Dataset Description

We have considered PIMA Indian Diabetes data set ¡reference to data set for our experiments. The original dataset contains 8 values, such as: Plasma Glucose Concentration a 2 hrs in OGTT, Diastolic blood pressure (in mm Hg), Two hours serum insulin (in mu U/ml), Body mass index (kg per meter square), Diabetes Pedigree Function, Number of times Pregnant, Triceps Skin Fold Thickness(in mm) and Age. And out of these 8 we have considered following 6 after referring to domain expert :

1) Age of Person

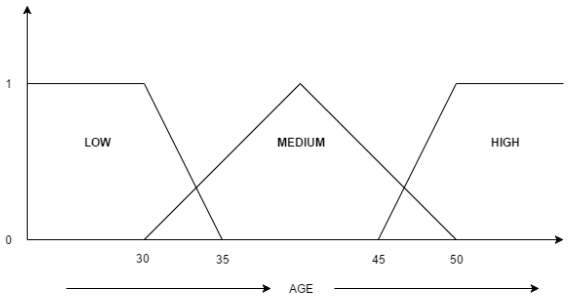

Figure 2. AGE

2) Glucose - Plasma glucose concentration (a two hours in OGTT)

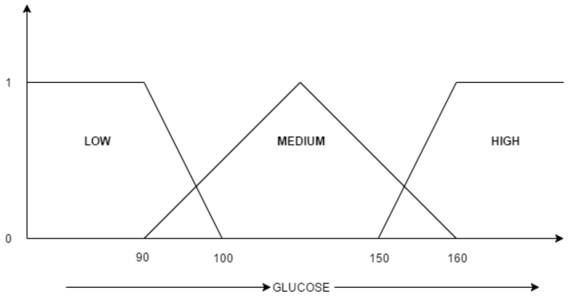

Figure 3. PLASMA GLUCOSE CONCENTRATION

3) Diastolic blood pressure (in mm Hg)

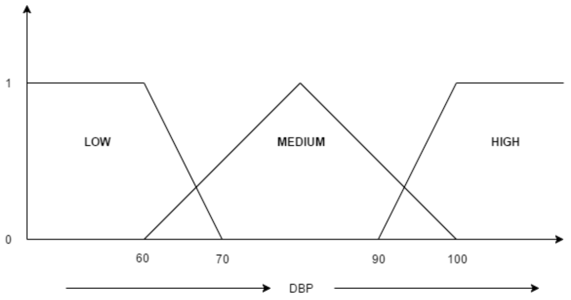

Figure 4. DIASTOLIC BLOOD PRESSURE

4) Insulin Two hours serum insulin (in mu U/ml)
5) BMI Body mass index
6) Diabetes pedigree function - Hereditary/Relationship Inheritence parameter

The original dataset had noise in terms of 0 for the values that are not available to database author. So, in database preprocessing, we removed all those entries of data that has noisy values. Then, using numerical values from dataset, we generated fuzzy classes for each entry. We give these fuzzy sets input to the three models and obtained the model output

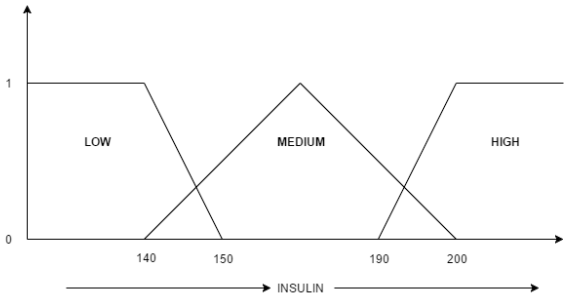

Figure 5. INSULIN

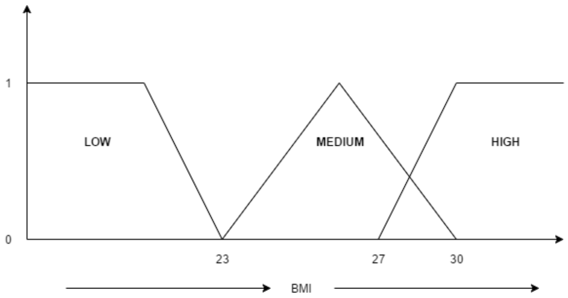

Figure 6. BODY MASS INDEX

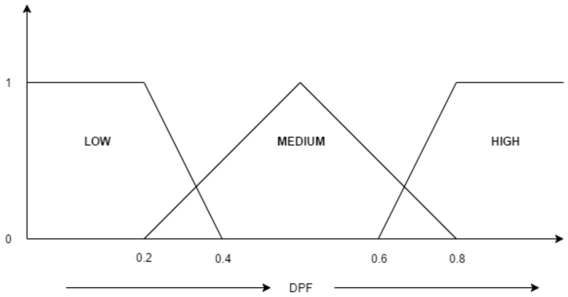

Figure 7. DIABETES PEDIGREE FUNCTION

Variable name	Variable Type	Input Value	Fuzzy Set
A1	AGE	≤ 35 30-50 ≥ 45	LOW MEDIUM HIGH
A2	GLUCOSE	≤ 100 90-160 ≥ 150	LOW MEDIUM HIGH
A3	DBP	≤ 70 60-100 ≥ 90	LOW MEDIUM HIGH
A4	INSULIN	≤ 150 140-200 ≥ 190	LOW MEDIUM HIGH
A5	BMI	≤ 23 23-30 ≥ 27	LOW MEDIUM HIGH
A6	DPF	≤ 0.4 0.2-0.8 ≥ 0.6	LOW MEDIUM HIGH

TABLE 1. FUZZY SETS

and compared with the original desired value for presence and absence of diabetes.

5. Experiment Results

5.1. Template Based Approach

The algorithm is implemented in Python. Based on the dataset and algorithm, fuzzy sets have defined, which is later used to determine fuzzy set of each feature value. Once the values are categorized according to fuzzy set, rules are generated and finally redundant rules are eliminated.

To test the accuracy of the system, Naive Bayes classifier is trained based on dataset and generated rules are passed as test cases. With the above procedure, accuracy achieved is : *74.074%*. Figure 8 shows the difference between original value and predicted value, horizontal axis is the data point and vertical axis tells the class it belongs.

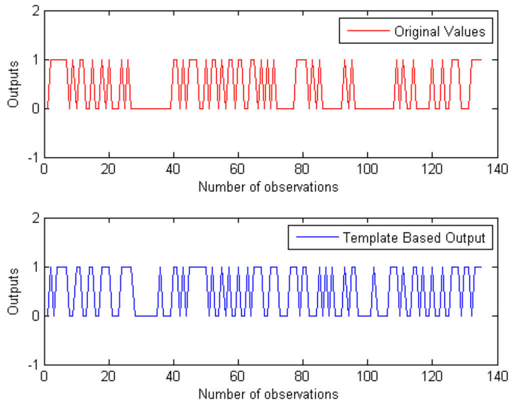

Figure 8. Template Based Approach

5.2. Neuro-Fuzzy Based Approach

The algorithm is implemented in Python. Some libraries are being used for critical mathematical analysis. Five layer Neural Network is defined. Membership Function used is Gaussian Function. Once the forward pass reaches final layer, error is calculated using Least Square Error Calculation for each instance. In Back-propagation parameters are updated. Repeating the above process for 50 epoch.

With above algorithm, we have achieved *Accuracy : 78.62%*. Figure 9 shows the difference between original value and predicted value, horizontal axis is the data point and vertical axis tells the class it belongs.

5.3. Fuzzy C-Means Clustering

For implementation of Fuzzy C-Means Clustering we used Matlab Toolbox. From the input data, two cluster centers are computed. These two clusters represents two class of output. By using this cluster centers we compute the

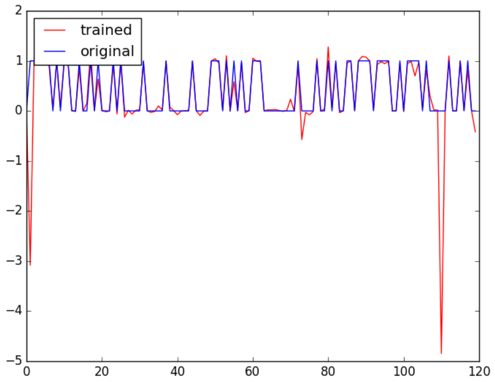

Figure 9. Neuro-Fuzzy Based Approach

class to which each data point belongs. Then we compare the output with original output. With this algorithm we have achieved *Accuracy : 76.84%*.

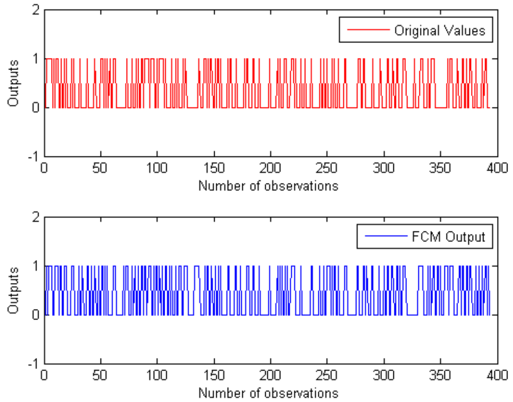

Figure 10. Fuzzy C-Means Clustering results.

6. Conclusion

In this project we have implemented three different type of fuzzy models for diabetes diagnosis system and obtained performances and accuracy of those methods. Based on our experiments we conclude the following order for performance: Neuro-Fuzzy system followed by fuzzy C-Means followed by Template based method.

References

[1] R. Dutta Baruah and D. Baruah, ”Modelling Fuzzy Rule-based Systems,” Handbook of Computational Intelligence,World Scientific Publishers, in press.

[2] Farhanah, Siti, Bt Jafan, and Darmawaty Mohd Ali. ”Diabetes Mellitus Forecast using Artificial Neural Networks (ANN).” Asian Conference on sensors and the international conference on new techniques in pharamaceutical and medical research proceedings (IEEE). 2005.

[3] Lowanichchai, Sudajai, Saisunee Jabjone, and Tidanut Puthasimma. ”Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree.” Faculty of Science and Technology Nakhon Ratchsima Raj abh at University, Nakhonratchasima 30000 (2006).

[4] Bagdi, Rupa, and Pramod Patil. ”Diagnosis of diabetes using OLAP and data mining integration.” International Journal of Computer Science Communication Networks 2.3 (2012).

[5] Kayaer, Kamer, and Tulay Yldrm. ”Medical diagnosis on Pima Indian diabetes using general regression neural networks.” Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP). 2003.

[6] Wang, L-X., and Jerry M. Mendel. ”Generating fuzzy rules by learning from examples.” IEEE Transactions on systems, man, and cybernetics 22.6 (1992): 1414-1427.

[7] Jang, J-SR. ”ANFIS: adaptive-network-based fuzzy inference system.” IEEE transactions on systems, man, and cybernetics 23.3 (1993): 665-685.

[8] Bezdek, James C., Robert Ehrlich, and William Full. ”FCM: The fuzzy c-means clustering algorithm.” Computers Geosciences 10.2-3 (1984): 191-203.

[9] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.