© 2014 IJIRT | Volume 1 Issue 10 | ISSN: 2349-6002

Prediction of Uncertainty in Clinical Database Using Clustering Technique

Dr.S.Sapna[1], M.Pravin Kumar[2]
[1]Assistant Professor, Department of Computer Science,
Bharathidasan College of Arts & Science, Erode – 638 116, Tamilnadu, India
[2]Assistant Professor, Department of Electronics and Communication Engineering,
K.S.R. College of Engineering, Tiruchengode – 637 215, Tamilnadu, India

Abstract - **Fuzzy logic system is used for solving a wide range of problems in different applications. When the inputs to the fuzzy logic system are increased, then the number of rules increases exponentially. It is difficult to proceed in the system. So the clustering technique is followed and the cluster centers are determined where each cluster center belongs to one rule in fuzzy logic. There have been many computerized methods proposed to generate fuzzy rules-base for diagnosis of diabetes. The vital idea is to generate the optimal rules needed to control the input without compromising the quality of control. The paper proposed the generation of fuzzy rule base by fuzzy C-means clustering method for the prediction of hidden knowledge from clinical diabetic database.**

Index Terms - **Fuzzy logic, fuzzy C-mean clustering (FCM), fuzzy rule base, clinical diabetic data.**

I. INTRODUCTION

Fuzzy logic system is used for solving a wide range of problems in different application. It also allows us to introduce the learning and adaptation capabilities. The fuzzy set framework has been used in several different process of diagnosis of disease. Fuzzy logic is a computational paradigm that provides a mathematical tool for dealing with the uncertainty and the imprecision typical of human reasoning.

The fuzzy inference system is a popular computing framework based on the concept of fuzzy set theory, fuzzy if-then rules and fuzzy reasoning [7]. The fuzzy inference system is designed based on the past known behavior of a target system. The fuzzy system is then expected to be able to reproduce the behavior of the target system [6]. For example if the target system is the medical doctor, then the fuzzy inference becomes a fuzzy expert system for medical diagnosis [4]. In the fuzzy logic method, any reasonable number of inputs can be processed and numerous outputs will be generated although defining the rule-base quickly becomes complex if too many inputs and outputs are chosen for a single implementation since rules defining their interrelations must also be defined. This will increase the number of fuzzy rules and complexity but may also increase the quality of the control. Many methods were proposed to generate fuzzy rules-base. The basic idea is to study and generate the optimum rules needed to control the input without compromising the quality of control.

Clustering is an exploratory data analysis method applied to data in order to discover certain groupings in a data set. Cluster analysis divides data into groups (clusters) such that similar data objects belong to the same cluster and dissimilar data objects to different clusters. In non-fuzzy or hard clustering, data is divided into crisp clusters, where each data point belongs to exactly one cluster. In fuzzy clustering, the data points can belong to more than one cluster and associated with each of the points are membership grades which indicate the degree to which the data points belong to the different clusters. In the clustering method the center is calculated and each center is taken as a single rule i.e. number of centers is equal to the number of rules.

Clinical database have enormous amount of information about doctor's guidance, patients and medical examination. Based on the outcome of laboratory test and results of physical examination, diagnosis is noted during time periods. The clinical process progress more time and all the clinical events are indescribable without considering time [5]. Therefore time is very important factor to many medical domain problems. By using clinical databases vital information are recovered and various diseases can be predicted. The clinical database is generally redundant, incomplete, vague and unpredictable. In this paper FCM technique is applied on clinical diabetic database.

Fuzzy C-Means Clustering (FCM), also known as fuzzy ISODATA, is a data clustering algorithm in which each data point belongs to a cluster to a degree specified by a membership grade [2]. The FCM takes a data set and a desired number of clusters and returns optimal cluster centers and membership grades for each data point. As an illustration, the fuzzy logic system with Multiple Input and Multiple Output (MIMO) process [9] is considered for which the fuzzy C-means clustering method is applied to generate optimum rules.

II. Diabetes Mellitus

Diabetes Mellitus often simply referred to as diabetes is the coercive effect of insulin on the glucose metabolism. Insulin is a hormone central to regulating carbohydrate and fat metabolism in the body. Insulin is produced from the islets of langerhans [1]. In Latin the word Insula means- "island". Its concentration has wide spread effect throughout the body. When control of insulin levels fails, diabetes mellitus will result. As a consequence, insulin is used medically to treat some forms of diabetes mellitus. Diabetes type 1 is lack of it whereas diabetes type 2 is the resistance towards it. Not only insulin regulates the glucose in the blood but it is also responsible for lipid metabolism. Insulin Secretion from beta-cells is principally regulated by plasma glucose levels [10]. Increased uptake of glucose by pancreatic beta-cells leads to a concomitant increase in metabolism. One must understand that insulin is offered as medicine only when the above criteria are broken. Physicians will become familiar with other aspects of managing the patient with diabetes, including the importance of postprandial glucose control, diabetes self-management training etc.

Most of the food we eat is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body. When you have diabetes, your body either doesn't make enough insulin or can not use its own insulin as well as it should [3]. This causes sugar to build up in your blood leading to complications like heart disease, stroke, neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage and death.

A. General Symptoms of Diabetes

Increased thirst, Increased urination - Weight loss, Increased appetite – Fatigue, Nausea and/or vomiting - Blurred vision, Slow-healing infections - Impotence in men.

B. Types of Diabetes

Type 1: Diabetes also called as Insulin Dependent Diabetes Mellitus **(IDDM)**, or Juvenile Onset Diabetes Mellitus commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. In type 1 diabetes, the pancreas undergoes an autoimmune attack by the body itself therefore; pancreas does not produce the hormone insulin. The body does not properly metabolize food resulting in high blood sugar (glucose) and the patient must rely on insulin shots. Type I disorder appears in people younger than 35, usually from the ages 10 to 16.

Type II: Diabetes is also called as Non-Insulin Dependent Diabetes Mellitus (NIDDM) or Adult Onset Diabetes Mellitus. Patients produce adequate insulin but the body cannot make use of it as there is a lack of sensitivity to insulin by the cells of the body. Type II disorder occurs mostly after the age 40.

Gestational Diabetes: Diabetes can occur temporarily during Pregnancy called as Gestational Diabetes which is due to the hormonal changes and usually begins in the fifth or sixth month of pregnancy (between the 24th and 28th weeks). It usually resolves once the baby is born. 25-50% of women with eventually develop diabetes later in life, especially in those who require insulin during pregnancy and those who are overweight after their delivery.

C. Diagnostic Tests

Urine Test: A urine analysis may be used to look for glucose and ketones from the breakdown of fat. However, a urine test alone does not diagnose diabetes. The following blood glucose tests are used to diagnose diabetes.

Fasting Plasma Glucose Level (FPG): The normal range of fasting blood glucose is <100 mg/dl. It is done after 8-12 hours of fasting. People with fasting glucose levels from 100-125 mg/dl are considered to have impaired fasting glucose. Patients with FPG >126 are consider to have diabetes mellitus.

Post Prandial Plasma Glucose Level (PPG): A blood sugar test taken after two hours of a meal is known as the post prandial glucose test or PPG. The normal range for PPG is <140 mg/dl. People with fasting glucose levels from 140-200 mg/dl are considered to have impaired glucose tolerance. Patients with PPG >200 mg/dl are consider to have diabetes mellitus.

D. Treatment

The major goal in treating diabetes is controlling elevated blood sugars. Type 1 diabetes is mainly treated with insulin, exercise, and a diabetic diet. Type 2 diabetes is treated with weight reduction, a diabetic diet, exercise and oral medications are used. Insulin medications are also considered. It can be overcome by a good nutritious diet, great physical activity, quitting smoking and excess alcohol and proper management of the condition.

III. Ease of Use Clustering Techniques

Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behavior. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever there are large amount of data are presented, it usually tend to summarize the huge number of data into a small number of groups or categories in order to further facilitate its analysis. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics.

A. Need of Clustering

In the fuzzy logic system if the number of inputs to the system is increased, then the number of rules increases exponentially. In predicting diabetes, five inputs and two outputs are considered. So it is difficult to proceed in the system and the concept of clustering is followed. In the clustering method the cluster centers are determined where each cluster center belongs to one rule in fuzzy logic.

B. Fuzzy C-Means Clustering

FCM employs fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership grades between 0 and 1. However, FCM still uses a cost function that is to be minimized while trying to partition the data set [8].

The membership matrix **U** is allowed to have elements with values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots\ldots, n. \qquad - (1)$$

The cost function for FCM is

$$J(U, c_1, \ldots., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}, \quad - (2)$$

where u_{ij} *is* between 0 and 1; c_i is the cluster center of fuzzy group i ; and $d_{ij} = \left\| c_i - x_j \right\|$ is the Euclidean distance between the ith cluster center and the jth data point; and $m \in [1, \infty)$ is a weighting exponent. The necessary conditions for equation (2) to reach its minimum are

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad - (3) \text{ and}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \qquad - (4)$$

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. In a batch mode operation, FCM determines the cluster centers c_i and the membership matrix **U** using the following steps:

Step 1: Initialize the membership matrix U with random values between 0 and 1 such that the constraints in Equation (1) are satisfied.

Step 2: Calculate ***c*** fuzzy cluster centers, $c_{i\,,}$, i=1,…, *c* using Equation (3).

Step 3: Compute the cost function according to Equation (2). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Compute a new U using Equation (4). Go to step 2.

The performance of FCM depends on the initial membership matrix values; thereby it is advisable to run the algorithm for several times, each starting with different values of membership grades of data points.

IV. FCM – IMPLEMENTATION FOR PREDICTION OF DIABETES MELLITUS

The clinical diabetic data were collected from various diabetic centers at Erode, Tamilnadu and mainly from SRC Diabetic Centre, Erode. About 1050 diabetic patients data were considered for this prediction and some of which is shown in Table 1. The input-output of the clinical diabetic data is used in FCM clustering method to analyze their performance. The inputs considered are Age, FPG-Fasting Plasma Glucose, PPG-Post Prandial Plasma Glucose, G-Gender, P/NP-Pregnant or Non Pregnant and the outputs are D-Diabetic Status and T1/T2/GD (T1 – Type 1/T2 - Type 2/GD - Gestational Diabetes). As five inputs and two outputs are considered, the number of rules will increase exponentially, so the FCM technique is applied to get the optimal number of rules. By applying FCM the number of cluster centers is obtained, where each cluster center belongs to one rule in fuzzy logic and also the hidden knowledge from the clinical database is predicted which aids the physician in decision-making. This prediction system also helps the user to anticipate by himself whether he is affected with diabetes or not.

TABLE 1 PRACTICALLY OBSERVED CLINICAL DATABASE OF DIABETES

INPUTS						OUTPUT	
S.No.	**Age**	**FPG (mg/dl)**	**PPG (mg/dl)**	**G**	**P / NP**	**D**	**T1/ T2 / GD**
1	52	95	290	0	0	0.7	0.7
2	51	109	452	1	0	0.91	0.7
.	.	.	.	.	.	.	.
1050	43	178	99	0	1	0.68	0.8

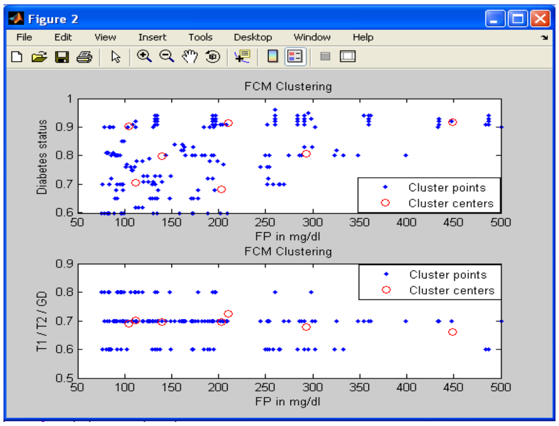
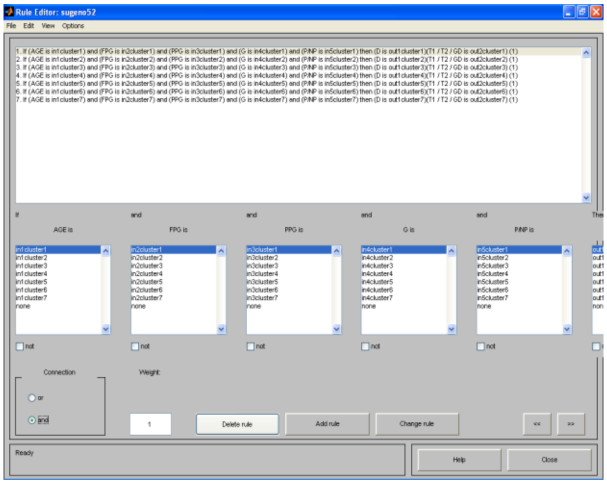


Fig 1. Performance of Fuzzy C-means Clustering

From Figure 1 it is shown that 7 cluster centers are obtained (indicated as red circle), where each cluster center belongs to one rule in fuzzy logic. The fuzzy if then rules is shown in Figure 2 and its inference system is shown in Figure 3 corresponding rule view is shown in Figure 4 which is obtained through fuzzy C-means clustering for the diagnosis of diabetes.



Rule Editor: sugeno52

1. If (AGE is in1cluster1) and (FPG is in2cluster1) and (PPG is in3cluster1) and (G is in4cluster1) and (P/NP is in5cluster1) then (D is out1cluster1)(T1 / T2 / GD is out2cluster1) (1)
2. If (AGE is in1cluster2) and (FPG is in2cluster2) and (PPG is in3cluster2) and (G is in4cluster2) and (P/NP is in5cluster2) then (D is out1cluster2)(T1 / T2 / GD is out2cluster2) (1)
3. If (AGE is in1cluster3) and (FPG is in2cluster3) and (PPG is in3cluster3) and (G is in4cluster3) and (P/NP is in5cluster3) then (D is out1cluster3)(T1 / T2 / GD is out2cluster3) (1)
4. If (AGE is in1cluster4) and (FPG is in2cluster4) and (PPG is in3cluster4) and (G is in4cluster4) and (P/NP is in5cluster4) then (D is out1cluster4)(T1 / T2 / GD is out2cluster4) (1)
5. If (AGE is in1cluster5) and (FPG is in2cluster5) and (PPG is in3cluster5) and (G is in4cluster5) and (P/NP is in5cluster5) then (D is out1cluster5)(T1 / T2 / GD is out2cluster5) (1)
6. If (AGE is in1cluster6) and (FPG is in2cluster6) and (PPG is in3cluster6) and (G is in4cluster6) and (P/NP is in5cluster6) then (D is out1cluster6)(T1 / T2 / GD is out2cluster6) (1)
7. If (AGE is in1cluster7) and (FPG is in2cluster7) and (PPG is in3cluster7) and (G is in4cluster7) and (P/NP is in5cluster7) then (D is out1cluster7)(T1 / T2 / GD is out2cluster7) (1)

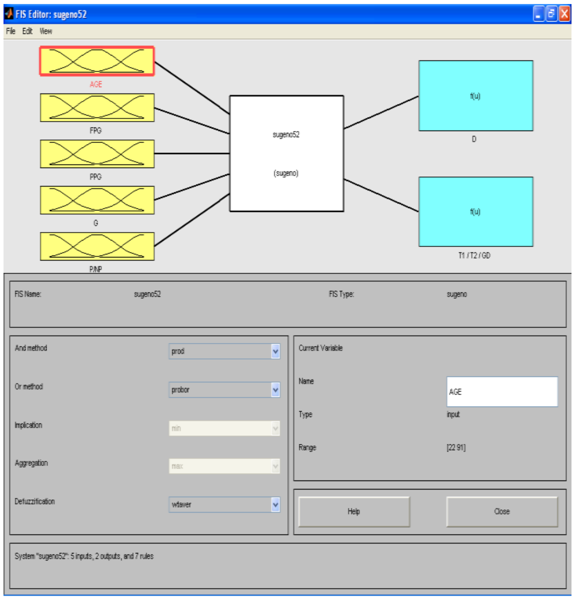
If	and	and	and	and
AGE is	FPG is	PPG is	G is	P/NP is
in1cluster1 in1cluster2 in1cluster3 in1cluster4 in1cluster5 in1cluster6 in1cluster7 none	in2cluster1 in2cluster2 in2cluster3 in2cluster4 in2cluster5 in2cluster6 in2cluster7 none	in3cluster1 in3cluster2 in3cluster3 in3cluster4 in3cluster5 in3cluster6 in3cluster7 none	in4cluster1 in4cluster2 in4cluster3 in4cluster4 in4cluster5 in4cluster6 in4cluster7 none	in5cluster1 in5cluster2 in5cluster3 in5cluster4 in5cluster5 in5cluster6 in5cluster7 none
□ not	□ not	□ not	□ not	□ not

Connection: ○ or ● and

Weight: 1

Ready

Fig 2. Fuzzy If Then Rules



FIS Editor: sugeno52

FIS Name: sugeno52

FIS Type: sugeno

And method: prod

Or method: probor

Implication: min

Aggregation: max

Defuzzification: wtaver

Current Variable

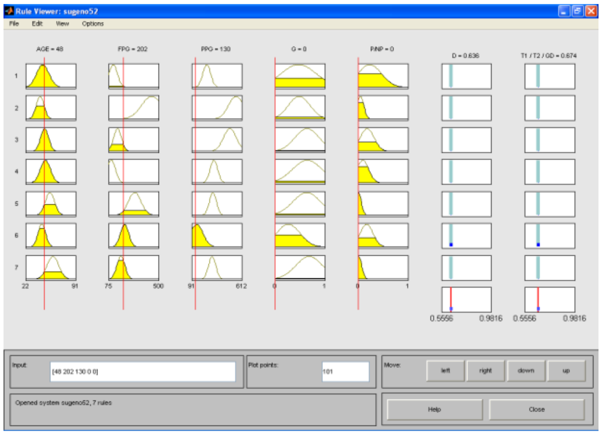
Name: AGE

Type: input

Range: [22 91]

System "sugeno52": 5 inputs, 2 outputs, and 7 rules

Fig 3. Fuzzy Inference System



Rule Viewer: sugeno52

Input: [48 202 130 0 0]

Plot points: 101

Opened system sugeno52, 7 rules

Fig 4. Rule View

The clustering process stops when the maximum number of iterations is reached. For the implemented practical diabetic data the clustering process stops at 65th iteration. From Figure 2, 3 and 4 the hidden knowledge from the clinical database can be predicted which aids the physician in decision-making. This prediction system also helps the user to anticipate by himself whether he is affected with diabetes or not.

V. Conclusion and Future Scope

Clinical processes progress over time and all the clinical actions are indescribable without considering time. Therefore time is vital to many medical domain problems. For the selected MIMO process the fuzzy C-means clustering method is applied and its performance is analyzed. FCM reduces the number of rules in the rule base by eliminating the redundant rules thereby reducing the computational time. In future, other fuzzy clustering algorithms based on volume criteria may be considered for rule optimization of various applications with more number of inputs and outputs.

References

[1] Aarogyam Preventive Health Care Magazine: Thyrocare’s, Basics of Diabetes, Vol.10, No.6, Nov 2006.

[2] Bezdek (1973), “Fuzzy Mathematics in Pattern Classification”, Ph.D., thesis, Applied Math. Center, Cornell University, Ithaca.

[3] Diabetes and YOU your guide to living well with diabetes, Novo Nordisk, LEAD GROUP.

[4] Gregory W. Ramsey, (2008),’Improving Chronic Disease Care Using Predictive Modeling and Data Mining’, morlab.mie.utoronto.ca/ DMHI2008 papers/C1_1.pdf, pp-31-36.

[5] Khanna Nehemiah, A.Kannan, K.Vijaya, Y.Nancy Jane, J.Brindha Merin,(2007), ‘Employing Clinical Data Sets for Intelligent Temporal Rule Mining and Decision Making, A Comparative Study’, ICGST-BIME Journal, Vol 7, Issue 1, Dec 2007, pp-37-45.

[6] Hossein Fazel Zarandi M., Mahammad Esmaeilian M. and Mehdi Fazel Zarandi, (2007), ‘A Systematic Fuzzy System

Modeling for Scheduling of Textile Manufacturing System', International Journal of Management Science and Engineering Management, Vol.2, No.4, pp.297-308.

[7] Jang J.S.R, Sun .C.T, and Mizutani.E (2004), "Neuro-Fuzzy and Soft Computing", Pearson Education, London.

[8] Mohanad Alata, Mohammad Molhim and Abdullah Ramini (2008), 'Optimizing of Fuzzy C-Means Clustering Algorithm Using GA', World Academy of Science, Engineering and Technology, Vol39, 2008, pp.224-229.

[9] Vijayachitra.S, Tamilarasi.A and Pravin Kumar.M, (2009), "Multiple Input Multiple Output (MIMO) Process Optimization Using Fuzzy GA Clustering", International Journal of Recent Trends in Engineering, Vol 2, Issue. 2, pp16-18.

[10] The New Indian Express, Health Article Tue, Aug 14, 2007, pp 1, by Dr. K.Bhujang Shetty.

BIOGRAPHY

Dr.S.Sapna received her B.Sc., Degree, M.C.A., M.Phil., Degree from Bharathiar University and Ph.D., from Mother Teresa Women's University. She has published several papers in reputed International and National Journals and Conferences. She is also the reviewer of various journals in the area of Data Mining, Soft Computing and Networking. Her research interest includes Soft Computing, Data Mining, Big Data, Mathematical Computations and Networks. She is a life member of ISTE and CSI.

M.Pravin Kumar received his B.E., & M.E., Degree, from Anna University, Chennai. He has published several papers in reputed International and National Journals and Conferences. His research interest includes Soft Computing, Networks and Big Data. He is a life member of ISTE.