Indian Journal of Science and Technology, Vol 9(35), DOI: 10.17485/ijst/2016/v9i35/99146, September 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Novel Hybrid Model for Diabetic Prediction using Hidden Markov Model, Fuzzy based Rule Approach and Neural Network

Nasib Singh Gill* **and Pooja Mittal**

Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak - 124001, Haryana, India; nasibsgill@gmail.com, mpoojamdu@gmail.com

Abstract

Objectives: Data mining approaches are used for developing the decision making systems. The current study proposes a novel hybrid model for diabetic prediction by using data mining techniques. The main objective of this study is to improve the accuracy rate by significantly reducing the size of the data under analysis at every stage. **Methods/Statistical Analysis**: To achieve the objectives, the PIMA Female Diabetic dataset, extracted from UCI repository, is used. The 10-fold cross validation method is used for extracting the testing and the training samples. Three rank based selection techniques are used for the attribute selection. The association between different attributes is identified and then clustering is performed under criticality using HMM and Fuzzy improved Neural Network. **Findings**: The data size reduces significantly when appropriate selection methods are applied in the respective sequence. For categorical data, the gain ratio attribute selection method out performs. Clustering is more effective when performed after identifying the exact associations among attributes. The proposed hybrid model achieved 92% of overall accuracy. The blend of supervised and un-supervised techniques achieved better results than the techniques when applied individually on the same data, as figured by the comparative analysis. The earlier prediction models worked either on classification or clustering. But in this present study, the classifiers and the clustering are performed. The Fuzzy improved Neural Networks are used for predicting the diabetes disease over the data. The result analysis proved that the prediction accuracy is poor (Naïve Bayes: 76.30%, Neural Networks: 75.13, Support Vector Machine: 77.47, K-Nearest neighbor: 69.79, Decision Tree (J48): 74.21), when the classifiers are implemented separately but when these are amalgamated with each other, produces better results. **Application/ Improvements:** The proposed hybrid model can be used as an expert system application, under the guidance of diabetic expert to assist the physicians for taking the decisions regarding the early diagnosis of the disease. In future, the proposed model can be applied on gender independent dataset. Further, the accuracy rate of the model can be improved by replacing the missing values of the dataset with the most appropriate value.

Keywords: Associative Clustering, Diabetes, Fuzzy Improved NN, Hidden Markov Model, Information Gain

1. Introduction

Data mining is a challenging and a versatile field of the data world. Medical mining is one of the most crucial and critical application of data mining. Medical mining primarily focuses on diagnostic prediction. It is always very hard-hitting to work with medical data sets, due to their high dimensional nature. From last decade, different data mining techniques are applied on medical data sets for extracting hidden and useful patterns for further decision making process. Past work shows that data mining techniques can effectively predict all kinds of diseases like heart disease, diabetes, skin ailments, lung disease, chronic cancer, etc.

In the present work, the Pima Indian diabetic dataset is considered for diagnostic prediction. Diabetes mellitus data is selected as diabetes is a chronic disease which may lead to a large range of health problems. It is the most common endocrine disease in all populations and among all age groups[1]. As at 2013, about 382 million people have

*Author for correspondence

diabetes worldwide[2]. It can also lead to a poor quality life as a person suffering from diabetes may face abundant complications like neurological, ophthalmic, peripheral circulatory, renal complications, optical disease, low immune body system, pregnancy problems for females and much more. Broadly, diabetes is of two types: Type1 and Type 2. Type 2 diabetes makes up about 90% of the cases[3]. In 2012 it was resulted in 1.5 million deaths worldwide, making it the 8th leading cause of death. Diabetes mellitus occurs throughout the world, but it is more likely to occur in developing countries[4]. Diabetes is a threat to mankind as it can't be cured completely but can be controlled, if predicted at an early stage, and even the post complications of diabetes can be avoided[1]. These reasons become the motivation for the present research work, which is intended to improve the diagnostic results that too at an early stage.

The paper is majorly organized in to six Sections. Section 1 introduces the research area. In Section 2, related is presented. Proposed model is described in Section 3 including the data description, feature selection approaches and three classification approaches: HMM, Fuzzy based Rule approach and Neural Network. Result analysis is presented in Section 4 including the comparative analysis. Finally the paper is concluded in Section 5 along with the future scope of the present research.

Classification is one of the most popular and functional approach of data mining having roots in statistics, machine learning, and pattern recognition and optimization techniques[5]. Several studies have been carried out on diabetes prediction using the classification techniques. Diabetes is mainly caused due to the metabolic disorder. Metabolic control rests heavily on self care of the patient and their family members[6]. Earlier, Logistic Regression (LR) was considered to be one of the most popularly used classifier for diabetes identification and classification. A study was conducted to assess the quality of diabetes care among 885 American Indian/ AlaskanNative[7] (AI/AN) adults in Canada with LR for Type 2 Diabetes Mellitus (T2DM). Another study performed by[8] determined the prevalence of Chronic Kidney Disease (CKD) among Chinese adults with diabetes and pre-diabetes using LR. However, later it was realized that the LR being a simple statistical approach cannot completely explain the complex relationship among features and diabetes[9]. The convoluted classifiers like Neural Networks (NN), Clustering, Association rules, Hybrid models, Fuzzy theory, proved to be more effectual by achieving better accuracy than simple classifiers, by revealing the hidden complex, non-linear relationship among different features more proficiently[10]. Neural networks are marginally better than Logistic Regression (LR) model, in terms of sensitivity and specificity, based on selection criteria[11]. NN was used by the authors[12] to evaluate an existing Health Risk Appraisal (HRA) for diabetes prediction based on a simulated learning technique. A review paper was presented[13] on early diabetes diagnosis with Support Vector Machine (SVM), NN, neuro fuzzy and attribute selection algorithms. A fuzzy classification system[14] was developed for diabetes diagnosis for small size dataset. A pragmatic work[15] at discovering associations between Diabetes Mellitus and many medically relevant attributes. Though, every classifier is equally good in predicting the disease but every classifier suffers from its own limitations. When these classifiers are applied in a hybrid approach, produces better results. A hybrid prediction model was proposed[16], using clustering approach for predicting the diabetes. Making clusters of the given data is also significant in improving the prediction results. It was concluded that the clustering algorithms require some metric[17] to evaluate the distance between different objects based on the object's features. A modified Gini index-Gaussian fuzzy decision tree algorithm with fuzzy decision boundaries was proposed in[1] and achieved 75.8% accuracy for diabetic dataset. The stepwise selection and genetic algorithms were used to identify the appropriate features to predict the diabetes. To accurately predict the onset of diabetic nephropathy[11], SVM and feature selection methods were applied, to visualize the risk factors for the diabetic data of 292 patients with high prediction performance. The Oracle Data Miner was employed as a software mining tool for predicting modes of treating diabetes[18]. This indicates that hybrid model is also a productive approach for deriving better accuracy in diabetes prediction. An experimental evaluation of Bayesian classifiers was conducted on intrusion detection[19]. Three techniques were compared using WEKA tool. The results proved that Bayes Net is better approach for intrusion data. They used true positive, true negative, confusion matrix for evaluation purpose. The values for different types of errors were also calculated to verify the results.

Based on all the findings from the study, we have framed the research problem for this current study. The objective of this study is to propose a hybrid model with different stages to achieve high recognition rate, by applying different mining approaches at different stages of the model. In the next section, the experimental design of the proposed work is presented.

2. Materials and Methods

Over the last few years, many researchers highlighted the strength of classification techniques in inferring the clinically accurate models for the patient data and to provide the decision support in this area. The objective of the majority of researches in this domain was to improve the accuracy rate. To attain the high recognition rate, in the present study we have tried to follow the guidelines of predictive data mining process. In this section, the proposed model is described with the feature selection as data preparatory phase, coined with different attribute selection methods. Next to this, the actual analysis stage is described, comprising of three versatile classification approaches. The proposed model is applied on Pima Dataset, described in the next section. The suggested model has combined the statistical analysis with three main analytical stages defined by the individual algorithms, described gradually in the paper.

2.1 Pima Diabetic Dataset

The reliability and accuracy of any prediction system, largely depends on the dataset used. In this work, an authenticated dataset called Pima[19] is used. This dataset contains the examined medical data for Pima Indians generated by the National Institute of Diabetes and Digestive and Kidney Disease. This institution has the name to predict various diseases at high recognition rates[20]. The institution has defined several experiments to reduce the risk factor and to identify the diabetes over the patients. The present work used the same dataset to improve its prediction ratio. The dataset description is shown in Table 1.

The basic statistical information regarding the dataset is defined in the Table 2. But this information is not sufficient to completely describe the attribute set. The description of individual attribute in terms of unit, format and valid range is defined in Table 2. The complete dataset is defined with 9 attributes out of which 8 attributes are the information attributes and one is the result attribute. The description of these dataset attributes is shown in Table 2.

Table 1. The description of pima diabetic dataset[20]

Property	Description
Dataset URL	http://archive.ics.uci.edu/ml/datasets/Diabatic+Disease
Number of Attributes	9
Number of Records	768
Attribute Names	Preg, Plas, Pres, Skin, Insu, Mass, Pedi, Age, Result
Class Attribute	1
Process Attributes	3
Numerical Attributes	8

Table 2. Attribute set description of pima diabetic dataset[20]

Attribute Name	Description	Unit	Format	Range
Preg	Defines the Number of Pregnancies	Number of Times	Numeric	0-17
Plas	Plasma Glucose Concentration in 2 Hour	mg/dl	Numeric	0-199
Pres	Blood Pressure	mmHg	Numeric	0-122
Skin	Skin Fold Thickness	Mm	Numeric	0-99
Insu	Serum Insulin (2Hours)	mu U/ml	Numeric	0-846
Mass	Body Mass	kg/m2	Numeric	0-67.1
Pedi	Diabetes Pedigree Function	-	Numeric	0.078-2.42
Age	Age of Patient	-	Numeric	21-81
Result	Class of Disease	-	Nominal	0-No, 1-Yes

Range of these attributes, mentioned in the Table 2 is determined by consulting the physician and by referring the previous researches. Many researchers have used the Pima data set for their studies. All eight attributes are very relevant and significant for predicting the disease. Though, Pima is very significant and authenticated dataset, yet it suffers from number of data problems like missing values, zero substitution, gender specific data and others. Thus, to minimize the adverse affects of these data problems on the results, in the present model, feature selection and data filtration is considered as a data preparation stage, a primary component of the present model, as described in the following section. The suggested methodology of the proposed model is discussed in the next section.

2.2 Proposed Methodology

In the proposed hybrid model, the algorithmic approaches used are the HMM approach, the Fuzzy Based Rue Logic

and the Neural Network Approach. The model defined in this work is shown in Figure 1. The model begins with an adaptation of raw dataset as input. For preparing the data, the attribute selection is performed to identify the valid and correct data. To analyze the extracted attributes individually as well as collectively, a HMM (Hidden Markov Model) based clustering approach is applied on selected sub set of data. This HMM approach has generated the multiple clusters to classify the dataset under the criticality vectors. In this work, a two level HMM is applied to perform clustering and classification. Once the filtered grouped data is obtained, the fuzzy rules are applied on these attributes. These fuzzy rules are applied on individual attributes by performing the dynamic dataset analysis. This data analysis is performed under the specification of fuzzy ontology and the reconstructed dataset values. Now, at the final stage of the proposed hybrid model, the neural network approach is applied as the major predictive approach.

To perform the diabetes prediction, the complete dataset is divided in to the training and the testing data. The training dataset is here defined as the HMM filtered and classified dataset whereas the testing dataset is considered as the raw dataset that can be taken as input from the user. To generate the training and testing dataset, the 10-fold cross validation method is applied. According to this method, 90% instances are considered as the training data set and 10% are considered as testing data set. The proposed model is logically partitioned into two sections: Feature Selection and Algorithmic Analytical Section. The feature selection approaches are described in the next section.

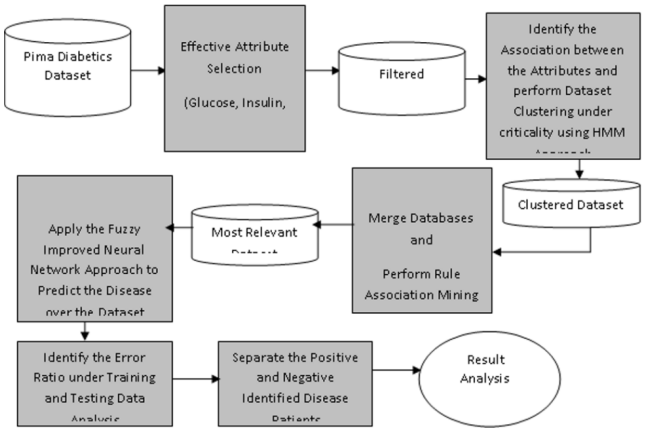

Figure 1. The proposed hybrid model for diabetes prediction.

2.2.1 Feature Selection

Feature selection is applied to identify the most relevant and reliable features over a large data set, to improve the efficiency and accuracy. There are number of rank based measures used to identify the effective dataset features. In the Pima dataset, only eight attributes are considered as the predictors for analysis and remaining one attribute is class label attribute used as a result attribute in the present study. In the present work, the Information Gain and Gain Ratio are used as the attribute selection methods, to achieve the optimum set of minimum number of characteristic attributes. Further, Fuzzy Based Feature Weight Evaluation technique is used to validate the impact of these selected attributes on the result.

2.2.1.1 Information Gain

Information Gain[21] being a versatile technique can be applied on single as well as multi-valued attributes. Higher gain value for an attribute indicates the higher relevancy of the attribute in the prediction. An effective work was carried to check the effects on diabetes mellitus based on distance based logistic regression and Information Gain[22]. The classification based association rule generation using MPSO-LSSVM is utilized first time with Information Gain as an outlier detection method. For the reason of eradicating the effect of unavoidable outliers in investigation sample on a scheme's performance, a new MPSO-LSSVM with the integration of outlier detection method was proposed[23]. The present evaluation is performed under the probabilistic estimation of an arbitrary tuple respective to the class, calculated by Class Estimation/Tuple Estimation. By estimating the different information gain values for every attribute of the given dataset, the following graph is generated, presented in Figure 2, representing the gain values for every attribute of the Pima dataset.

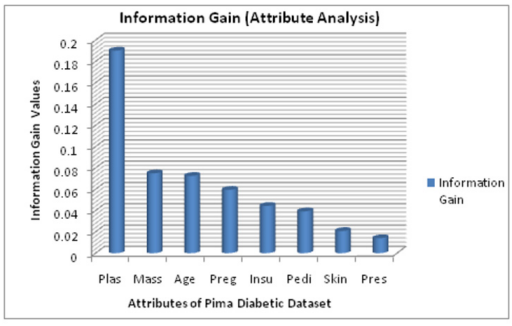

Figure 2. Information gain analysis of pima diabetic dataset.

As depicted in the Figure 2, five top most attributes with high information gain values are Glucose Plasma Ratio, Body Mass Index Value, Age, Pregnancy Count and Insulin. Out of these attributes, age is not considered as an analysis parameter for further process, as age is a generic attribute in disease estimation and can be used as suggestive attribute, as consulted by the physician. Pregnancy count is also not considered as it is specific to gender as well as to age group. To make the proposed model more generic, these two attributes are not considered as the analysis parameters, in the present study. Rest from the remaining attributes, the top attributes: glucose plasma, body mass index and insulin are considered as the most effective domain specific attributes.

2.2.1.2 Gain Ratio

Gain Ratio[21], another measure developed to overcome the realistic limitation of Information Gain. In the proposed study, Gain Ratio is formulated as the second feature selection technique to minimize the effects of the biasness due to Information Gain, on the results. In the present model, the gain values for all the attributes are obtained by training the dataset under WEKA environment. The Figure 3 depicts the results derived from gain ratio formulation for all eight numeric attributes of Pima dataset.

From the above graph, it is clear that top five attributes identified for analysis are same as selected from Information Gain analysis. Out of these five attributes: age and pregnancy count are eliminated because of the same above mentioned reasons. As an outcome of second feature selector, glucose plasma value, mass value and insulin value are considered as the most effective attributes for further algorithmic stages.

In order to ensure the relevancy of extracted attributes from these two extractors, a fuzzy rule based quantified analysis is performed and the discussions with diabetic doctors as well as the references from existing researches are also considered.

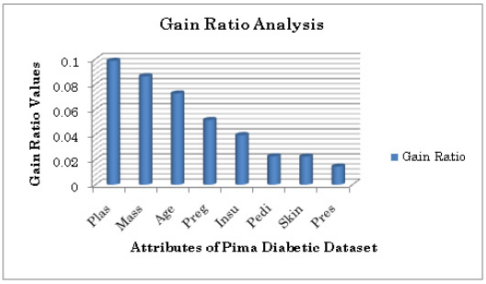

Figure 3. Gain ratio analysis of pima diabetic dataset.

2.2.1.3 Fuzzy based Feature Weight Identification

To achieve the data range based relevancy of selected attributes, a fuzzy based quantified analysis is performed on selected attributes (predictors for further analysis), in this section, to classify the given predictor values under Low, Medium or High range. From this fuzzy based analysis, the role of attribute in the further predicting process can be analyzed. Before applying these fuzzy rules on the predictors, the primary work is to analyze the dataset for positive and negative disease occurrence under frequency analysis. At first, the fuzzy analysis is performed to classify the plasma glucose predictor level in Low, Medium and High classes. The analysis shown in Figure 4 is performed under the following fuzzy range specification: 0 to 66.3 represents glucose_low, 66.4 to 132.67 represent the glucose_medium and 132.68 to 199 represent the glucose_high. The dataset analysis under these range specification, has proved that if the glucose quantity is high, there are more chances for the patient to be diabetic, whereas if the Glucose quantity is low or medium, the diabetic chances are comparatively lesser. In Figure 4, the pictorial representation of positive and negative occurrence of diabetes according to Low, Medium and High range of glucose is presented.

The second predictor selected for disease prediction is insulin level, in the present study. It actually analyzes whether the body is able to produce the adequate insulin or not. The degree of insulin in the body is the major factor to detect the diabetes as it is the main reason that affects the insulin origin cells. The high level of insulin in the blood is considered as the other important reason for disease occurrence. In this work, the fuzzy rule is applied to insulin data values to predict the disease probability. This analysis is also performed under high, medium and low range. The fuzzy rule defined the low range between 0

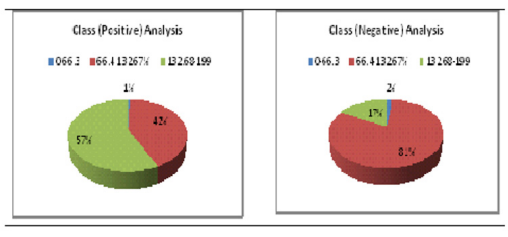

Figure 4. Plasma glucose feature analysis under positive and negative classes for low, medium and high range.

and 100 whereas medium rule defined the range between 101 and 300 and high insulin range is defined between 301 and the MAX_Insulin value in the dataset. The frequency analysis of the values over the dataset is shown in Figure 5. As shown in the Figure, the low insulin level is another main reason for disease occurrence.

Third and the last predictor considered in this work is the body mass value of the patient, which is measured in kg/m^2. In this data set, the body mass lies between 0 and 67. To analyze the diabetic disease occurrence probability, the fuzzy logic is applied to this predictor again under Low, Medium and High range. The low range for body mass is defined between 0 and 22. The fuzzy analysis clearly depicts that the patients falling in this range have the lesser probability of diabetes. Another fuzzy range applied here is the Medium_Mass that defined the mass range between 23 and 45. As shown in Figure 6, maximum number of patients with positive occurrence comes under this range andabout 93% patients having the chances of being diabetic. Third fuzzy vector under this predictor is defined as High_Mass range between 46 and 67. About 6% patients having body mass in the High Mass range are susceptible of being diabetic. The analysis of dataset values is shown in Figure 6.

In this section, the individual predictor analysis is performed under fuzzy rules. The fuzzy process is applied on various ranges (Low, Medium and High) to observe the data weight significance. From this fuzzy analysis, we obtained the critical and the safe data ranges for all three predictors: Glucose, insulin and body mass. After

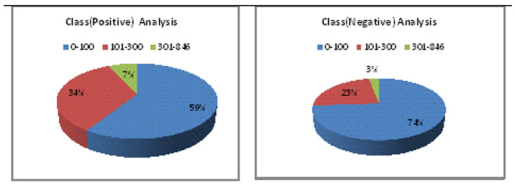

Figure 5. Insulin level analysis under positive and negative classes for low, medium and high range.

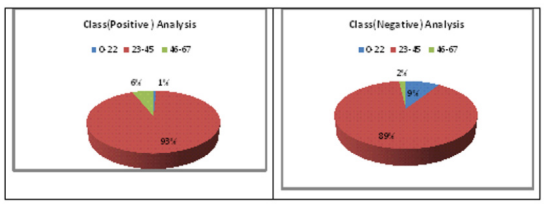

Figure 6. Body mass attribute analysis under positive and negative classes for low, medium and high range.

performing this fuzzy based feature weight analysis, the selection of the three predictors is validated as the decision vectors for the further diabetic prediction process. In the present study, the primary predictor considered for diabetic analysis is the blood glucose level (highest gain value). After this first level classification, HMM model is applied as discussed in the next section.

1.1.2 Algorithmic Analytical Section

In this section, the Hidden Markov Model, Fuzzy based Rule approach and Neural network are applied on the filtered data, derived from the feature selection step. All the three approaches are described in the following sections:

2.2.2.1 Hidden Markov Model (HMM)

To analyze the extracted attributes individually as well as collectively, the Hidden Markov Model[21] based clustering approach is applied. This HMM approach has generated the multiple clusters to classify the dataset under the critical vectors. In this work, a two level HMM is applied to perform clustering and classification. The first level HMM is applied on individual attributes to analyze the critical nature of all the attributes individually. At the second level, all these attributes are combined and a collective analysis is performed to generate the clusters for association over the dataset and to classify the dataset values. The HMM technique used as a clustering approach in present work is presented in Table 3.

The HMM algorithm discussed in Table 3, at first stage, divides the dataset in clusters at individual levels, on the basis of distance analysis. At the second stage, the associated attribute analysis is performed. The associated mean based distance analysis is carried out here according to sensitivity and strength analysis. A threshold limit is applied for effective associated data selection. The strong data pairs are analyzed to gain the maximum weight value as the filtered rules. These selected rules or data values are considered as the optimized predicted dataset. The clusters with extremely smaller data values are ignored. Based on these clusters, the state space is filtered, to take more accurate decisions for training dataset. This generated rule is considered as an input for the next stage classification.

2.2.2.2 Fuzzy based Rule Approach

Once the filtered data groups are obtained, the fuzzy rules are applied at the second stage of the proposed model, on the paired attributes gained from HMM modeling, by

Table 3. HMM based clustering at individual and collective levels of attributes

```
Step 1. Accept the Filtered Diabetic Dataset as the Input Dataset.
Step 2. Specify the Expected Number of Clusters
Step 3. For a=1 to No. Of Attributes
        {
Step 4.  Centroid (a)=Mean(Data Set(a,:)) /*Calculate attribute level centroid */
Step 5.  For r=1 to Data set. Size /*Apply row wise analysis on dataset*/
        {
Step 6.  Dist=Dataset(a,r)-Centroid (a) /*Estimate distance from centroid for each dataset value*/
Step 7.  Cindex=Diff(Clusters,Dist,AttribLimits) /*Identify the eligible cluster based on distance analysis*/
Step 8.  Clusters(Cindex).Add(Dataset(a, r))/*Include the dataset value in relative cluster*/
        }
        }
Step 9.  For r =1:Dataset.Size          /*Apply Predictive Analysis For each row*/
        {
Step 10. For a=1 to No. Of Attributes /* Attribute level analysis*/
        {
Step 11. For b=1 to No. Of Attributes
        {
        {
Step 12.  Ass=GetAssociativity(Dataset(a,r),Dataset(b,r))/*To perform the data value based Associativity Analysis*/
Step 13.  Cent=Associativity Mean(Dataset(a,:),Dataset(b,:))/*Generate Associated Mean as updated centroid*/
Step 14.  D1=DistDiff(Cent,Ass,Strength,Sensitivity)/*Obtain the dataset prediction with associative centroid, strength and
sensitivity ranges*/
Step 15. If(D1<Range) /*Check for Valid Data values*/
        {
Step 16.  Strength Data(Dataset(a,r),(b,r))/*Select as predicted data rule value*/
        }
        }
Step 17.  Filtered Rules. Add(Max(Strength Data)) /*Take the maximum accepted value as considered Filtered Dataset
Rules*/
        }
        }
Return Filtered Rules
```

performing the dynamic dataset analysis. This knowledge construction process is defined by specifying the fuzzy operators on the predicted dataset. This fuzzy based rule approach is applied under the structural analysis of domain information and considered as one of the primary stages applied to the dataset. This structural analysis has categorized the individual attribute values under high, medium and low constraints. After performing the attribute level structural analysis, the group relation analysis is performed on HMM predicted dataset to generate the fuzzy grouping on the dataset values. This structural analysis mechanism is applied on all instances collectively to derive the knowledge and to generate the repository under the validation specification. This specification is performed under the specification of fuzzy ontology and the reconstructed dataset values. The AND group operator is used here to combine all these attributes collectively to predict the disease and the rule based data values are generated. Based on this predictive data based fuzzy formation, more selective and weighted data pairs are collected. The paired data is here weighted in terms of fuzzy observations, shown in Figure 7, which are considered as an input to the neural network stage.

Here Figure 7 is showing the results of second level fuzzy analysis applied on group pairs. This analysis is applied with attribute relevance specification and identifies the potency of the selected attributes in the prediction process. The figures also conclude that the weighted predictor value of plasma glucose has maximum contribution in prediction process as compared to insulin and body mass attributes. Finally, at the third stage of the proposed model, this weighted dataset is processed under neural network modeling, as explained in the next section.

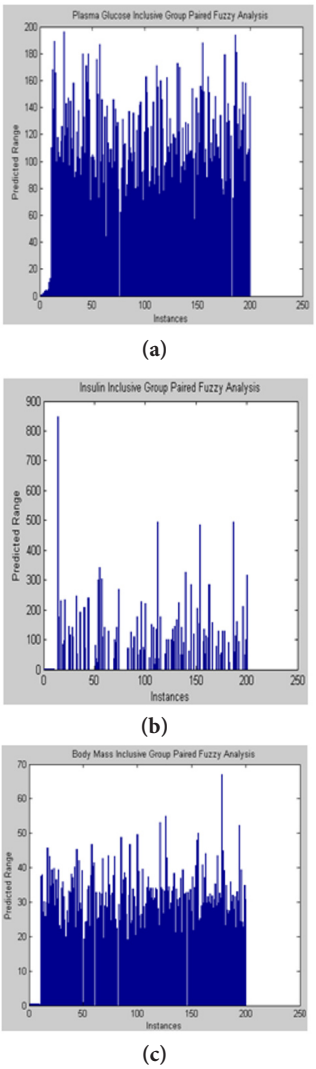

Figure 7. **(a)** Plasma glucose inclusive group paired fuzzy analysis. **(b)** Insulin inclusive group fuzzy analysis. **(c)** Body mass inclusive group fuzzy analysis.

2.2.2.3 Neural Network

Predictor values derived from fuzzy rules are more capable of deriving effective outcomes from the dataset. Now, at the final stage of the proposed work, the neural network approach is applied as the major predictive approach. To perform this prediction, the dataset is divided in to training and testing data. To generate the training and testing dataset, the 10-fold method is applied. According to this method, 90% instances are considered as the training data set and 10% are considered as testing data set. After this dataset partition, the fuzzy rules are applied on the training set, to filter the data values for more accurate prediction. As described above, the training dataset is processed under HMM and fuzzy filtration classification for gaining the rule set. To perform the result validation, the raw input is considered as the testing dataset, which is processed under training set adaptive rules for generating the data class for the testing data, as shown in the Figure 8. Based on these rules mapping, the results are compared at the final stage to analyze the prediction accuracy. The rules are applied on the dataset attributes individually as well as collectively for testing the data.

Later on the neural network parameters are defined for these training and testing datasets. Neural Network (NN) used here is a multilevel perception. Here, the fuzzy weighted diabetic dataset values are represented as the processing neurons and these neurons are defined in the form of NN layers of the model shown in Figure 8. The fuzzy rule based values are used as the weighted links and are applied as the middle layer of the system of the model defined in Figure 8. The input testing dataset is defined to obtain the target values with the specification of MAE (Mean Absolute Error). Lesser the error in the prediction process, higher the accuracy obtained over the system. As the network is activated, the input dataset is propagated to the neural system with weighted interconnection under fuzzy specifications. The activation function is applied at different layers of neuron process to perform the conversion from dataset values to the target values.

After this fuzzy improved neural model, the target values are obtained along with the effective error rate. The obtained target values are compared with the input dataset values to analyze the prediction rate. The difference analysis between the actual disease, class and the obtained results is considered as the result estimation. To perform this matching the quantitative measures are performed on target values and the actual results. The input layer, output layer and biasness defined for the neural network, can

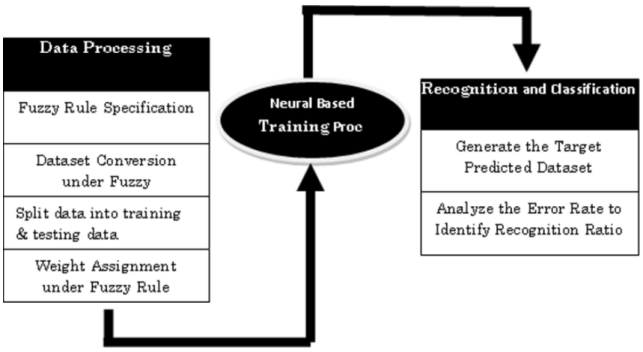

Figure 8. Neural network model applied, the last stage of the proposed model for classification.

be noticed from Figure 9. The neural process performed on the dataset can be seen in Figure 10, defined with maximum number of epoch's specification, time and performance limits.

The error is identified in terms of difference between the predicted testing dataset and the relative actual results. The prediction result of the present study is represented in terms of error occurrence. The optimality is achieved after about 38 iterations as shown in Figure 11. It is clear from the snapshot that, any further iteration would not change the system performance.

Here Figure 11 is showing the gradient graph of the neural process used in the proposed model. Back propagation is a gradient based algorithm. It is showing the data value and the validation variation up to the defined iterations.

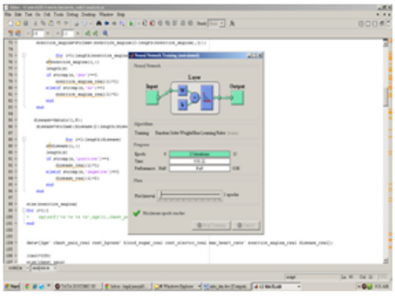

Figure 9. Neural process showing the training of the network.

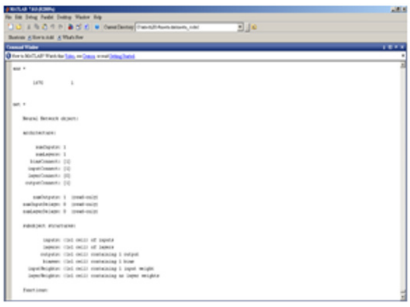

Figure 10. Neural network parameters.

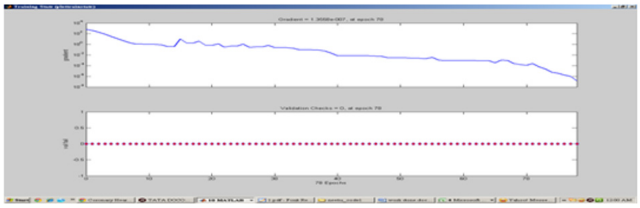

Figure 11. Validation analysis by gradient graph of the neural process.

3. Results and Discussions

In this paper, a novel hybrid model is proposed for the diabetic prediction which is verified under 10-fold cross validation and achieved 92% accuracy. To ensure the reliability of the proposed model, it is analyzed under different system conditions, by varying the sample size and by changing the learning rate, while keeping the epoch size fixed, equal to 10, for calculating the error occurrence rate. Following Table 4 presents the different error values occurred in the network, with different system parameters:

As evident from the results presented in the Table 4, error rate majorly depends on the learning rate of the neural network. Lesser the learning rate value, higher the error rate of the system. The obtained results of the proposed model, is verified under different system parameters, as presented in the Table 4, taken for different sample sizes and the different learning rates. After the complete analysis, the model performance is calculated as 92% accuracy rate. In next section, a comparative analysis is presented to validate the model performance.

3.1 Comparative Analysis

To validate the proposed hybrid model, a comparative analysis is performed with existing popular data mining approaches. The WEKA tool is used to apply different classification approaches on Pima data set. The approaches considered under WEKA are Bayesian Network, Neural Network Approach, SVM, KNN and Decision Tree approach. From the comparative analysis performed under WEKA tool, presented in Table 5, it is clearly evident that proposed model has outperformed all other mentioned classification approaches.

Here Figure 12 is depicting the analytical results graphically, obtained from this study. These comparative results are based on recognition ratio (accuracy) obtained from the proposed model as well as from existing classification approaches. Here x-axis represents the

Table 4. Error occurrence by varying the data sample size and the learning rate, with fixed epoch size = 10

Learning Rate	Error Rate
0.001	2% to 8%
0.01	2% to 15%
0.2	1% to 7%
0.1	0.05% to 0.13%

Table 5. Comparative analysis of proposed model with other classification techniques

Approach used	Accuracy (%)
Naïve Bayes	76.30
Neural Networks	75.13
Support Vector Machine (SVM)	77.47
K-Nearest Neighbor (KNN)	69.79
Decision Tree (J48 Classifier)	74.21
Proposed Hybrid Model	**92**

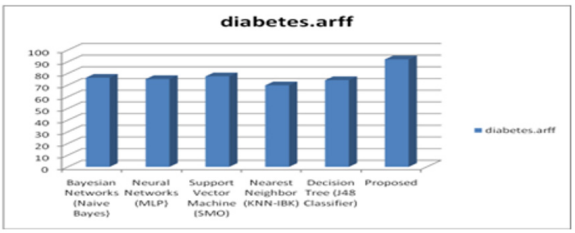

Figure 12. Comparative analysis of the proposed model with existing approaches for pima dataset.

classification methods and y-axis represents the different accuracy ratio achieved. As shown in the Figure 12, the proposed hybrid model has achieved 92% of recognition ratio which is better than existing approaches such as naive bayes, neural network, SVM, KNN and Decision tree based approach.

4. Conclusions and Future Scope

In this paper, a novel hybrid model is developed using MATLAB, to predict the diabetic disease based on the symptom analysis. The proposed model is defined with comprehensible definition of its each intermediate stage. At first, the Pima diabetic dataset is analyzed and three attributes: Blood glucose, insulin and body mass, are selected as predictors for further analysis, based on the feature selection methods used. Then, at second stage, the three classifiers (HMM clustering, fuzzy based rule approach and the Neural Network approach) are applied on the selected data. The proposed model has achieved 92% accuracy rate when feature selection, three classifiers are implemented. The comparative analysis proved that the proposed model had outperformed the earlier models and approaches. The suggested hybrid model can be used as an expert system application, under the guidance of diabetic expert to assist the physicians for taking decisions regarding the early diagnosis of the disease. In future, the proposed model can be applied on gender independent dataset.

5. References

1. Varma KVSRP, Rao AA, Lakshmi TSM, Rao PVN. A computational intelligence approach for a better diagnosis of diabetic patients. Computers and Electrical Engineering. 2014; 40(5):1758–65.
2. Shi S, Yuankai Y, Hu FB. The global implications of diabetes and cancer. The Lancet. 2014; 383(9933):1947–8.
3. Melmed S, Kenneth S, Polonsky PMDP, Larsen RMD, Kronenberg HMMD. Williams book of endocrinology. 12th ed. Philadelphia: Elsevier/Saunders; 2015. p. 1371435.
4. Alex J S A, Mukhedkar A S, Venkatesan N. Performance analysis of SOFM based reduced complexity feature extraction methods with back propagation neural network for multilingual digit recognition. Indian Journal of Science and Technology. 2015 Aug; 8(18):1–8.
5. Obenshain KM. MAT application of data mining techniques to healthcare data. Statistics for Hospital Epidemiology. 2004; 25(8):690–5.
6. Sigurdardottir AK, Jonsdottir H, Benediktsson R. Outcomes of educational interventions in Type 2 diabetes: WEKA data mining analysis. Patient Education and Counseling. 2007; 67(1-2):21–31.
7. Dyck RF, Hayward MN, Harris SB. Prevalence, determinants and co morbidities of chronic kidney disease among First Nations adults with diabetes: Results from the circle study. BMC Nephrology. 2012; 13(1):57.
8. Jia W, Gao X, Pang C, Hou X, BaoY, Liu W. Prevalence and risk factors of albuminuria and chronic kidney disease in Chinese population with T2DM and impaired glucose regulation: Changhai diabetic complications study (SHDCS). Nephrology Dialysis Transplantation. 2009; 24(12):3724–31.
9. Su CT, Yang CH, Hsu KH, Chiu WK. Data mining for the diagnosis of type II diabetes from three dimensional body surface anthropometrical scanning data. Computers and Mathematics with Applications. 2006; 51(6–7):1075–92.
10. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics. 2008; 77(2):81–97.
11. Cho BH, Yu H, Kim KW , Kim TH , Kim IY, Kim SI. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial Intelligence in Medicine. 2008; 42(1):37–53.
12. Park J, Edington WD. A sequential neural network model for diabetes prediction. Artificial Intelligence in Medicine. 2001; 23(3):277–93.

13. Shankaracharya S. Computational intelligence in early diabetes diagnosis: A review. Rev Diab Stud. 2010; 7:252–62.
14. Mostafa M, Fathi G, Mohammad M, Saniee A. A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. Expert System with Applications. 2011; 38(12):14650–9.
15. Kasemthaweesab P, Kurutach W. Association analysis of Diabetes Mellitus (DM) with complication states based on association rules. 7th Conference on Industrial Electronics and Applications; Singapore. 2011. p. 1453–7.
16. Patil BM, Joshi RC, Durga T. Hybrid prediction model for type-2 diabetic patients. Expert Systems with Applications. 2010; 37(12):8102–8.
17. Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining. USA: Addison-Wesley; 2006.
18. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences. 2013; 25(2):127–36.
19. Koklu M. Pima Diabetic data set. AIM. 1994; 7(8):1–3.
20. Singh GN, Pooja M. A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. Journal of Theoretical and Applied Information Technology. 2016; 87(1):1–10.
21. Han J, Micheline K. Data mining: Concepts and techniques. 2nd ed. USA: Morgan Kaufmann Publishers; 2006.
22. Devi MN, Balamurugan AA, Kris MR. Developing a modified logistic regression model for diabetes mellitus and identifying the important factors of type II DM. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–8.
23. Karthikeyan T, Vembandasamy K. A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method. Indian Journal of Science and Technology. 2015 Apr; 8(S8):1–11.