Predict the onset of diabetes disease using Artificial Neural Network (ANN)

Manaswini Pradhan[1] , Dr. Ranjit Kumar Sahu[2]

[1]*Lecturer, P.G. Department of Information and Communication Technology, Fakir Mohan University, Odisha, India*
[2] *Consultant, Tata Memorial Cancer Hospital (TMH), Mumbai, India*

[1]ms.manaswini.pradhan@gmail.com, [2]drsahurk@yahoo.co.in

Abstract: Diabetes Mellitus is a chronic metabolic disorder, where the improper management of the blood glucose level in the diabetic patients will lead to the risk of heart attack, kidney disease and renal failure. Data Classification is a prime task in Data mining. Accurate and simple data classification task can help the clustering of large dataset appropriately. In this paper we have experimented and suggested an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients into two classes. For achieving better results, genetic algorithm (GA) is used for feature selection. The GA is used for optimally finding out the number of neurons in the single hidden layered model. Further, the model is trained with Back Propagation (BP) algorithm and GA (Genetic Algorithm) and classification accuracies are compared. The designed models are also compared with the Functional Link ANN (FLANN) and several classification systems like NN (nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) for Data classification accuracies. It is revealed from the simulation that our suggested model is performing better compared to NN(nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) and FLANN model and it can be a very good candidate for many real time domain applications as these are simple with good performances.

Key words: ANN, Genetic Algorithm, Data classification, FLANN

1. Introduction

Diabetes is one of the most deadly, disabling, and costly diseases observed in many of the nations at present , and the disease continues to be on the rise at an alarming rate .Women tend to be hardest hit by diabetes with 9.6 million women having diabetes. This represents 8.8% of the total women adult population of the 18 years of age and above in 2003 and this is nearly a two fold increase from 1995 (4.7%). Women of minority racial and ethnic groups have the highest prevalence rates with two to four times the rates of the white population. With the increased growth of minority populations, the number of women in these groups who are diagnosed will increase significantly in the coming years. By 2050, the projected number of all persons with diabetes will have increased from 17 million to 29 million. Diabetes is metabolic disorder where people suffering from it either have a shortage of insulin or have a decreased ability to utilize their insulin. Insulin is a hormone that is produced by the pancreas and allows glucose to be converted to energy at the cell level. Diabetes that is uncontrolled, that is consistently high levels of blood glucose (> 200mgl/dL) leads to micro and macro vascular disease complications, such as, blindness, lower extremity amputations, end stage renal disease, and coronary heart disease and stroke. Diabetes is found in approximately one in ten individuals, but the chances increases to one in five as the age group increases to 65 years of age and above.

DIABETES Mellitus (DM) is a chronic and progressive metabolic disorder and according to the World Health Organization there are approximately million people in this world suffering from diabetes. The number of diabetic patients is expected to increase by more than 100% by the year 2030 [1]. Common manifestations of diabetes are characterized by insufficient insulin production by pancreas, ineffective use of the insulin produced by the pancreas or hyperglycemia. Causes like obesity, hypertension, elevated cholesterol level, high fat diet and sedentary lifestyle are the common factors that contribute to the prevalence of diabetes. Development of renal failure, blindness, kidney disease and coronary artery disease are types of the severe damage which are resulted by improper management and late diagnosis of diabetes. Even though there is no established cure for diabetes, indeed, the blood glucose level of diabetic patients can be controlled by well-established treatments, proper nutrition and regular exercise [1]-[3].

Data classification is a classical problem extensively studied by statisticians and machine learning researchers. It is an important problem in variety of engineering and scientific disciplines such as biology, psychology, medicines, marketing, computer vision, and artificial intelligence [1]. The goal of the data classification is to classify objects into a number of categories or classes. Given a

dataset, its classification may fall into two tasks. First, supervised classification in which given data object is identified as a member of predefined class Second, unsupervised classification (or also known as Clustering) in which the data object is assigned to an unknown class. Supervised classification (here onwards to be referred as classification) algorithms have been widely applied to speech, vision, robotics, diseases, and artificial intelligence applications etc where real time response with complex real world data is necessity. There have been wide ranges of machine learning and statistical methods for solving classification problems. Different parametric and non-parametric classification algorithms have been studied [4-11]. Some of the algorithms are well suited for linearly separable problems. Non-linear separable problems have been solved by neural networks [12], support vector machines [13] etc. However, in many cases it is desired to find a simple classifier with simple architecture. There has been wide spectrum of work on developing ANN based classification models consisting of many hidden layers and large number of neurons in the hidden layers. It is obviously understood from the literature of ANN that more number of hidden layers and large number of neurons may sometimes present a good solution for the problem but at the expense of computational cost. There have been also some attempts made to use FLANN [15] for classification purpose. In Functional Link Artificial Neural Networks (FLANNs) the hidden layer is removed without giving up non-linearity by providing the input layer with expanded inputs that are constructed as the functions of original attributes [15]. Removal of hidden layer makes these networks extremely simple and computationally cheap. Identification of nonlinear processes using FLANNs has been reported by researchers [16]-[19]. FLANNs have an inherent limitation, of not guarantying universal approximation, which has deterred interest in them. Only a few applications using FLANNs are available in literature. In this paper ANN (Artificial Neural Network) model is suggested in which we have proposed a ANN model having m-n-p as the model parameters wherein m is the number of inputs (based on the dataset under investigation) , n is the number of neurons in hidden layer (only one hidden layer is used to minimize the computational complexity) and p is the number of output neurons (based on the dataset under investigation). The optimal numbers of neurons in the hidden layer are chosen by Genetic Algorithm (GA) [14]. The weights of the novel ANN are tuned by BP algorithm and GA for different datasets and results are compared. Our results are also compared with FLANN based models, NN (nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset). It is found that the models suggested by us are less in complexity and better in performance. The paper is sequentially arranged in the following order. Section II comprises related survey works. In section III, Back Propagation (BP) algorithm is briefly described, which is used to train the SANN. Section IV describes the GA approach for optimally finding the values for the number of neurons in the hidden layer. FLANN basics are described in section V. In section VI the dataset is described. Section VII discusses the simulation and results. Conclusion and future research are given in Section VIII.

2. Related Works

Apart from the works mentioned above, a lot of research has been done specifically using ANN in diagnosing diabetes mellitus and some approaches are discussed below.

Siti Farhanah, Bt Jaffar and Dannawaty Mohd [20] proposed a method for diagnosing diabetes. The diagnosis is accomplished using back propagation neural network algorithm. The inputs to the system are plasma glucose concentration, blood pressure, triceps skin fold, serum insulin, Body Mass Index (BMI), diabetes pedigree function, number of times a person was pregnant and age. The biggest challenge to this method was the missing values in the data set. This system was later modified and presented by T.Jayalakshmi and Dr.A.Santhakumaran[21].They have proposed an idea to overcome the missing values that was not addressed by Siti Farhanah Bt Jaafar [20] and this included constructing the data sets with reconstructed missing values, thereby increasing the classification accuracy[21]. They have also proposed an alternate method to overcome missing value by performing data pre-processing, which also speeds up the training process by reducing the actual learning time. Various missing value techniques and pre-processing methods were analyzed. By adopting these modifications, the results improved and achieved a classification accuracy of 99% [20]. Rajeeb Dey [22] proposed a method to predict diabetes mellitus using back propagation algorithm of Artificial Neural Network (ANN). The problem of diagnosing diabetes has been treated as a binary classification, i.e., those predicted to be diabetic falls under category 1 and others falls under category 0. The basic architecture of ANN used for accomplishing this classification task is a supervised multilayer feed-forward network with back propagation learning algorithm. The parameters considered in this system to diagnose diabetes are Random Blood Sugar test result, Fasting Blood Sugar test result, Post Plasma Blood Sugar test, age, sex and their occupation. The performance has been measured in terms of absolute error calculated between network response and desired target. Classification performance achieved using the system is 92.5%. Eng Khaled Eskaf [23] proposed a method for managing diabetic patients by trying to predict their glucose levels in the near future on the basis of current

levels of glucose. The prediction is done using Artificial Neural Network (ANN). Feature extraction procedure was implemented on diabetic blood glucose time series. Blood glucose values of diabetic patients are recorded for 24 hours for about one week with a sampling frequency of 5 minutes. A dynamic model is used as an extraction procedure in order to extract different values from a blood glucose test. ANN was then trained using this knowledge to predict the blood glucose levels of a diabetic patient with reasonable accuracy. Gregory Hastings [24] proposed a hybrid model for diagnosis of diabetes mellitus by integrating three different data mining techniques using supervised and unsupervised learning algorithm. The inputs were processed using Support Vector Machine (SVM) and rules were extracted using electic approach. Real time data set is taken as input from which rules are formulated to describe relationships between input features and output class labels and thus diagnosis is done based on the rules extracted through electic approach. Of the two rules generated by the electic approach one was found to be inconsistent with generally acceptable medical knowledge in diagnosis of diabetes. The above survey clearly highlights that ANN technique for diagnosis of diabetes gives much better results than other existing techniques. Also, when considering the inputs to all these systems, there is at least one input value for which the patient should get the help of a doctor or a hospital staff. The proposed system aims to avoid the patients from undergoing blood tests, checking diastolic and systolic blood pressure etc, thus creating a user friendly environment without the need for a doctor or a hospital staff. The inputs designed are based on the symptoms which could appear during the early stages of diabetes and based on the physical conditions.

3. Back Propagation Training of ANN

An MLP (Multi-Layer Perceptron) network with 2-3-3 neurons (2, 3 and 3 denote the number of neurons in the input layer, the hidden layer and the output layer respectively) with the back-propagation (BP) learning algorithm, is depicted in Figure1.

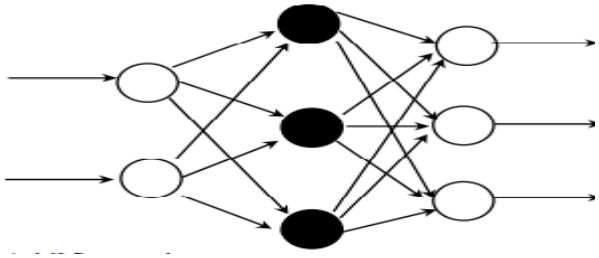


Figure 1: MLP network

Initialize the weights: In BP algorithm, initially the weights are initialized to very small random numbers. Each unit has a bias associated with it. The biases are similarly initialized to small random numbers.

Propagate the inputs forward: First, the training tuple is fed into the input layer of the network. The inputs pass through the input units, unchanged. Next, the net input and output of each unit in the hidden and output layers are computed. The net input to a unit in the hidden or output layers is computed as a linear combination of its inputs. To compute the net input to the unit, each input connected to the unit is multiplied by its corresponding weight, and is summed. Given a unit in a hidden layer or output layer, the net input, I_j, is

$$I_j = \sum_i W_{ij} O_i + \theta_j$$

(1)

where W_{ij} is the weight of the connection from the unit I in the previous layer to unit j .

O_i output of unit I from the previous layer

θ_j is the bias of the unit

The bias acts as a threshold in that it serves to vary the activity of the unit.

Each unit in the hidden and output layers takes its net input and then applies an activation function. The function symbolizes the activation of neuron represented by the unit. The logistic or sigmoid function is used and the output of unit j, is computed as

$$O_j = \frac{1}{1+e^{-I_j}}$$

(2)

This function is also referred to as squashing function, because it maps a large input domain onto the smaller range of 0 to 1.

Back propagate the error:The error is back propagated backward by updating the weights and biases to reflect the networks prediction. For a unit j in the output layer, the error is computed by

$$Error_j = O_j(1-O_j)(T_j - O_j)$$

(3)

where O_j is the actual output of unit j, and T_j is the known target value of the given training tuple. To compute the error of a hidden layer unit j, the weighted sum of the errors of the units connected to unit j in the next layer is considered. The error of a hidden layer unit j is,

$$Error_j = O_j(1-O_j)\sum_k Error_k W_{jk}$$

(4)

where W_{jk} is the weight of the connection from unit j to a unit k in the next higher layer, and $Error_k$ is the error of unit k.

The weights and biases are updated to reflect the propagated errors. Weights are updated by following equation,

$$W_{ij} = W_{ij} + (l)Error_j O_i$$

(5)

The variable *l* is the learning rate, a constant having a value between 0 to 1. Back propagation learns using a method of gradient descent to search for a set of weights that fits the training data so as to minimize the mean squared error. The learning rate helps avoid getting stuck at a local minimum and encourages finding global minimum.

Biases are updated by following equation,

$$\theta_j = \theta_j + (l)Error_j$$

(6)

Terminating condition:

Training stops when

i) All weights in the previous epoch were so small as to below some specific threshold

ii) The percentage of tuples misclassified in the previous epoch is below some threshold

iii) A prespecified number of epochs has expired.

A classifier which gives a higher accuracy value is considered as a good classifier.

Algorithm:

1. Initialize weights and biases in the network.
2. Propagates inputs forward in the usual way, i.e. All outputs are computed using sigmoid threshold of the inner product of the corresponding weight and input vectors. All outputs at stage n are connected to all the inputs at stage n+1
3. Propagates the errors backwards by apportioning them to each unit according to the amount of this error the unit is responsible for.
4. Terminating condition (Error is minimum or till the iterations are exhausted).

4. GA for optimally finding Neurons in hidden layer of ANN

Genetic algorithms (GA) are an evolutionary optimization approach which is an alternative to traditional optimization methods. GAs is most appropriate for complex non-linear models where location of the global optimum is a difficult task. It may be possible to use GA techniques to consider problems which may not be modeled as accurately using other approaches. Therefore, GA appears to be a potentially useful approach. GA follows the concept of solution evolution by stochastically developing generations of solution populations using a given fitness statistic (for example, the objective function in mathematical programs). They are particularly applicable to problems which are large, non-linear and possibly discrete in nature, features that traditionally add to the degree of complexity of solution. Due to the probabilistic development of the solution, GA does not guarantee optimality even when it may be reached. However, they are likely to be close to the global optimum. This probabilistic nature of the solution is also the reason they are not contained by local optima. The standard GA process consists of an initialization step and the iterative generations [14]. The Genetic Algorithm (GA) process is described below. First, a population of chromosomes is created. Second, the chromosomes are evaluated by a problem defined fitness function. Third, some of the chromosomes are selected for performing genetic operations. Fourth, genetic operations of crossover and mutations are performed. The offspring produced out of genetic operations replace their parents in their initial population. This GA process repeats until a user defined criterion is met.

Procedure Genetic Algorithm

```
begin
i=0 /* i: number of iteration*/
initialize P(i) /* P(i): population for iteration I */
compute f(P(i)) /* f: fitness function */
  perform until (non termination condition)
          begin
          i=i+1;
          choose two parents P1 and P2 from P(i-1)
          perform genetic operations
          {
          crossover;
           mutation;
          }
        reproduce a new P(i)
        compute f(P(i))
  end-perform
end
```

In our work chromosomes are nothing but string of integers within a random range depicting the possible values for the number of neurons in the hidden layer. Depending on the dataset complexity i.e the dimension and number of data objects in the dataset this range has been chosen for simulation. The fitness value is nothing but the classification accuracy of the model. More the percentage of correct classifications better the model. The population size is chosen based on the dataset. However, a population size up to 30 is a good choice for our simulations. In our work as we are aiming for ANN having only one hidden layer, the binary GA is implemented for the simulation. In Binary GA the chromosomes initialized by integer is converted to binary values for applying the GA operators. In entire of our work the one point cross over is adopted with 0.8 crossover probabilities and with 0.01 mutation probability.

5. Basics of FLANN

FLANN architecture for predicting a diabetic patient is a single layer feed forward neural network consisting of one input layer and an output layer.

The FLANN generates output (positive '1' or negative '0') by expanding the initial inputs (drivers) and then processing to the final output layer. Each input neuron corresponds to a component of an input vector. The output layer consists of one output neuron that computes diabetic positive/negative as a linear weighted sum of the outputs of the input layer. Pao[15] originally proposed the FLANN architecture. They have shown that, their proposed network may be conveniently used for function approximation and pattern classification with faster convergence rate and lesser computational load than an MLP structure. The FLANN is basically a flat net and the need of the hidden layer is removed and hence, the learning algorithm used in this network becomes very simple. The functional expansion effectively increases the dimensionality of the input vector and hence the hyper planes generated by the FLANN provide greater discrimination capability in the input pattern space.

To bridge the gap between the linearity in the single layer neural network and the highly complex and computation intensive multi layer neural network, the FLANN architecture is suggested. The FLANN architecture uses a single layer feed forward neural network and to overcome the linear mapping, functionally expands the input vector. Let each element of the input pattern before expansion be represented as z(i),1 < i < d where each element z(i) is functionally expanded as $z_n(i)$,1 < n< N , where N=number of expanded points for each input element. Expansion of each input pattern is done as follows.

$$x_1(i) = z(i),\ x_2(i) = f1(z(i)),....,\ x_N(i) = fN(z(i))$$

where, z(i),1 < i < d , d is the set of features in the dataset.

These expanded input patterns (shown in Figure 2) are then fed to the single layer neural network and the network is trained to obtain the desired output. The set of functions considered for function expansion may not be always suitable for mapping the nonlinearity of the complex task. In such cases few more functions may be incorporated to the set of functions considered for expansion of the input dataset. However dimensionality of many problems itself are very high and further increasing the dimensionality by to a very large extent may not be an appropriate choice. So, it is advisable to choose a small set of alternate functions, which can map the function to the desired extent.

Architecture of FLANN:

The FLANN network can be used not only for functional approximation but also for decreasing the computational complexity. This method is mainly focused on functional approximation. In the aspect of learning, the FLANN network is much faster than other network. The primary reason for this is that the learning process in FLANN network has two stages and both stages can be made efficient by appropriate learning algorithms. In this study the Functional Link Artificial Neural Network (FLANN) model for the task of pattern classification in data mining is evaluated. The FLANN model functionally expands the given set of inputs. These inputs are fed to the single layer feed forward ANN. A single layer model based on trigonometric expansion is presented. Let each element of the input pattern before expansion be represented as z(i),1 < i < I where each element z(i) is functionally expanded as $z_n(i)$,1 < n < N , where N = number of expanded points for each input element. In this, N = 5 and I = total number of features in the dataset as been taken. Expansion of each input pattern is done as follows:

x (i) = z(i), x (i)= sin π (z(i)),
x (i)=sin 2π (z(i)), x (i)= cos π (z(i)),
x (i)= cos 2π (z(i))

where, z(i),1 < i < d , d is the set of features in the dataset.

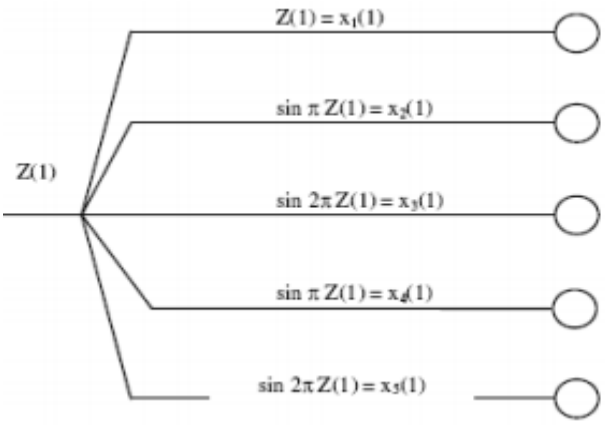


Figure 2: Functional expansion of each unit

These nonlinear outputs are multiplied by a set of random initialized weights from the range [-0.5, 0.5] and then summed to produce the estimated output. This output is compared with the corresponding desired output and the resultant error for the given pattern is used to compute the change in weight in each signal path P, given by

$$\Delta W_j(k) = \mu * xf_j(k) * e(k) \quad (7)$$

where, $xf_j(k)$ is the functionally expanded input at kth iteration.

Then the equation, which is used for weight update, is given by

$$W_j(k+1) = W_j(k) + DW_j(k) \quad (8)$$

where, $W_j(k)$ is the jth weight at the kth iteration, m is the convergence coefficient, its value lies between 0 to 1 and 1<j<J, J = M´d. M is defined as

the number of functional expansion unit for one element.

$$e(k) = y(k) - \hat{y}(k) \quad (9)$$

where, y(k) is the target output and ŷ (k) is the estimated output for the respective pattern an and is defined as:

$$\hat{y}(k) = \sum xf_j(k).w_k \quad (10)$$

where, xf_j is the functionally expanded input at kth iteration and $W_j(k)$ is the jth weight at the kth iteration and $W_j(0)$ is initialized with some random value from the range [-0.5, 0.5]. The FLANN for classification shown in Figure 3. FLANN is a single layer nonlinear network. Let *k* be number of input-output pattern pairs to be learned by the FLANN. Let the input pattern vector $\boldsymbol{X_k}$ be of dimension *n*., and the output y*k* be a scalar. The training patterns are denoted by $\{ X_k, y_k\}$. A set of *N* basis functions

$$\phi(X_k) = [\phi(X_k)\phi(X_k)\phi(X_k)...\phi(X_k)]^T \quad (11)$$

are adopted to expand functionally the input signal

$$X_k = [x_1(k)x_2(k)...x_n(k)]^T . \quad (12)$$

These *N* linearly independent functions map the n-dimensional space into an *N*-dimensional space, that is $R^n \rightarrow R^N, n < N$ .The linear combination of these function values can be presented in its matrix form, that is $S = W\phi$. Here $[S_k = S_1(k)S_2(k)....Sm(k)]^T$, W is the *m X N* dimensional weight matrix. The matrix S_k is input into a set of nonlinear function $\rho(.) = \tanh(.)$ to generate the equalized output $\hat{Y}=[\hat{y}_1, \hat{y}_2 \ldots\ldots \hat{y}_m]$, $\hat{y}_j = \rho(S_j), j = 1,2,...,m.$ The major difference between the hardware structures of ANN and FLANN is that FLANN has only input and output layers and the hidden layers are completely replaced by nonlinear mappings. In fact, the task performed by the hidden layers in an ANN is carried out by functional expansions in FLANN. Being similar to ANN, the FLANN also uses BP algorithm to train the neural networks.

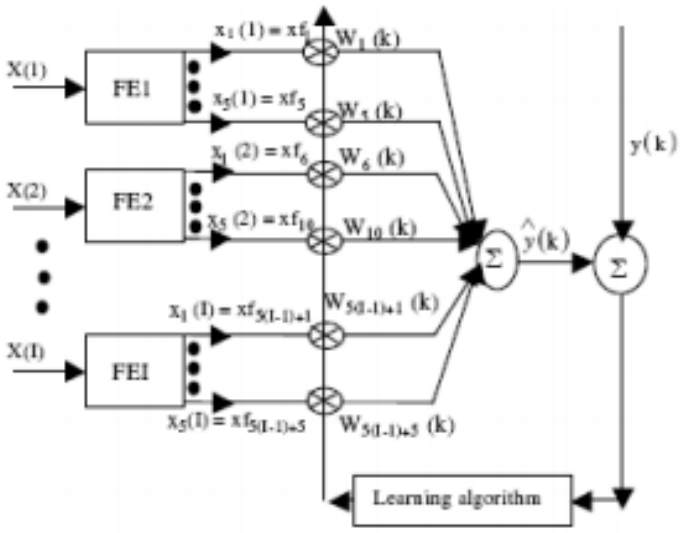


Figure 3: FLANN nonlinear model for classification

6. Dataset Description

In our work we have used Pima Indian Diabetes data sets [25] for training and testing the neural network model.

PIMA INDIAN DIABETES Dataset:

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Attribute Information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

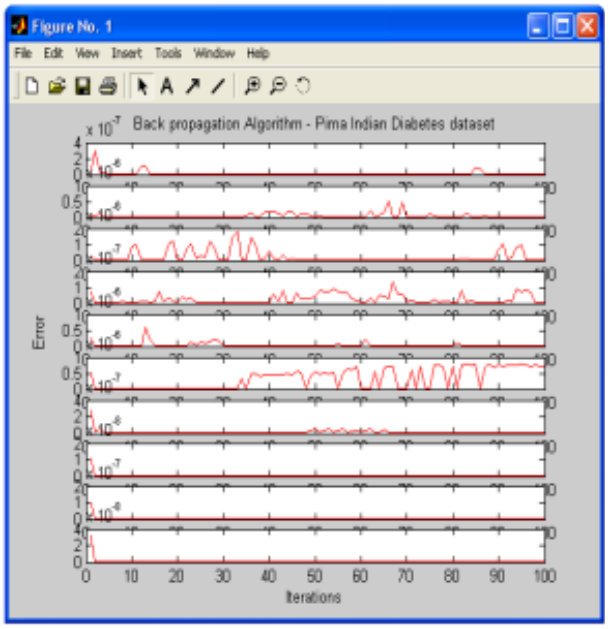
7. Simulation and Results

The neural network, which we are using in back propagation algorithm, is m-n-1 type network. It represents that input layer would contain 'n' nodes, which will be equal to the number of attributes in the dataset we are using. Say in Pima Indian Diabetes dataset we have 8 attributes so of construction neural network for this we require 8 nodes in its input layer. In the above notation m represents nodes in hidden layer (only one hidden layer we are considering). We can have any number of nodes in hidden layer and in output layer we are considering only one node.

For training and testing we have adopted 10-fold cross validation for Pima Indian Diabetes, in which we divide tuples in the dataset into 10 equal

divisions. We apply back propagation algorithm for first 9 divisions and train the network for certain number of iterations. After training the network we then apply same algorithm without propagation of errors back and find the accuracy of the 10th division. We repeat this process for all the remaining divisions by placing the last division on the top moving down the remaining tuples Such that each division will take part in training the network.

GA is used for optimally selecting the n values. For Pima dataset the ANN gives the best accuracy with 5 neurons in the hidden layer. Best accuracy being 72% with average accuracy of 72.2%. The MSE is at 1.6838e- 004. The error plot is shown below.



Back propagation algorithm X: Y –>Iteration versus Error

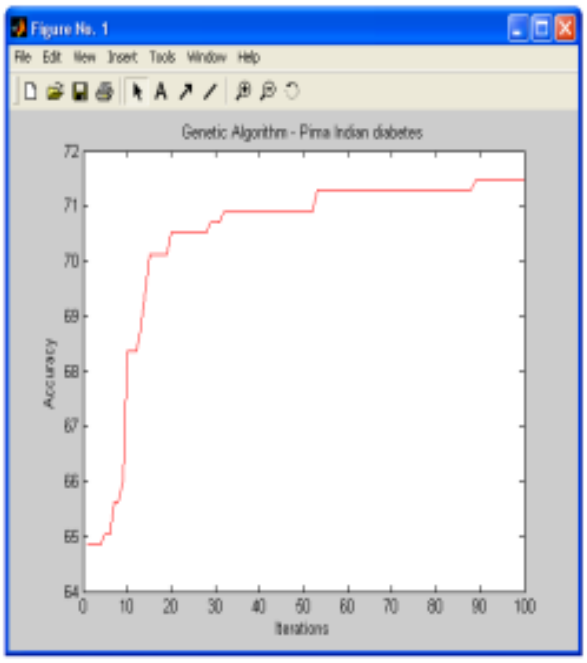
The results obtained using BP algorithm for all the investigated dataset reveal that the networks fall into the local minima. The results can be improvised if some randomized optimization techniques is used for training the ANN i.e optimizing weights of the ANN. This has motivated us to explore the use of GA for optimizing the weights of ANN. In our experiments we have used the same ANN which we have obtained with GA and trained with BP. For Pima the ANN is 8-5-1. For example, in the Pima Indian diabetes dataset ANN model total weights to optimized are (8*5+5*1) =45. In our simulation the population size is taken to be 30. The performance of the model is shown in the table 1 given below.

Table 1: Performance of Pima Indian Diabetes model

Model/Dataset	Best Accuracy	Average accuracy(standard)
8-5-1/Pima	73.438	71.212(0.324)

The weights of these networks are optimized using GA. The chromosomes of GA are the strings of real numbers randomly chosen in a range. The length of chromosome is determined by the number of weights to be optimized. The number of weights to be optimized is (m*n+n*p)
The performances of these models are shown in the table 1 given below.

The results reveal GA based optimization for ANN is far more accurate with comparison to BP based training. The fitness curves for the Pima dataset is given below.



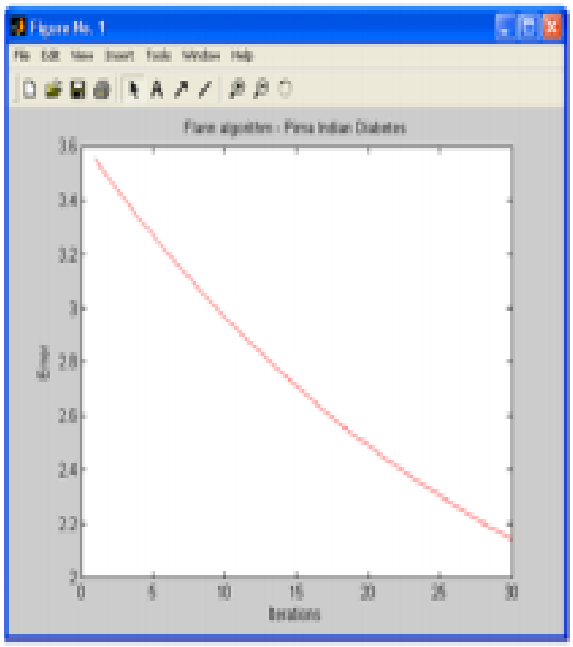
Genetic Algorithm X: Y –>Iteration versus Accuracy

The simulated result of FLANN model is shown below in a tabular format in Table 2. The experiment is done for 30 epochs and 10 such simulations are considered for finding average correct classifications.

Table 2: Simulation result of FLANN Pima Model

Dataset	Convergence coefficient	Best accuracy	Average accuracy
Pima Indian Diabetes	8e-006	71.845	59.76

The error curves using FLANN for the Pima model shown below is x: y is iterations verses error



FLANN algorithm X: Y –>Iteration versus Error

Comparison with other models:

The results obtained for the Pima Indian Diabetes Database dataset were compared with the results described in (Aha, D.W. and R.L. Bankert, 1994. Feature selection for case-based classification of cloud types: An empirical comparison. Proc. Am. Assn. for Artificial Intelligence (AAAI-94)-Workshop Case-Based Reasonings, pp: 106-112.) where the performance of several models is presented: NN(nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset).Table 3 presents the results obtained by the various different models.

Table 3: Comparison of the average performance of several other classification systems

Models	**Pima Indian diabetes dataset**
NN	65.1%
kNN	69.7%
BSS	67.7%
MFS1	68.5%
MFS2	70.5%
Novel ANN	73.4%
FLANN	59.8%

8. Conclusions

In recent years several researches have been conducted to classify and show that who is diabetic or not. Shanker used neural networks (NN) to predict the diabetic person and also showed that neural networks obtained a better accuracy which was higher than other methods like logistic regression, feature selection, decision tree etc. This paper has explored the design of a novel ANN for data classifications. We have evaluated the ANN model for the task of pattern classification in data mining. The novel ANN is envisioned by using GA for optimally deciding the number of neurons in single hidden layer architecture. The weights of such novel ANN is trained using BP algorithm and GA algorithm respectively. The experimental studies demonstrated that the ANN model performs the pattern classification task quite well. Again, this gives a very clear impression of the simplicity of the model without sacrificing at the cost of accuracy. This model proved to be better than other models and algorithms with which it was compared. The performance of this model is remarkable in terms of processing time, which is treated as one of the crucial aspect in data mining. Contrary to the views of many researchers it is felt that we can have novel ANN model with only one hidden layer and having very few neurons in the hidden layer. Even our designed novel ANN outperforms NN(nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) [26] and FLANN classification models in the dataset .Due to less computational cost at the hidden layer novel ANN can have applications in the real time domain.
However, this study is nascent to claim the universalness of the model. It is open for further study to examine how the other models like Bayesian classifier, Decision tree, etc, behave with comparison to this suggested models. It can also be further investigated to improvise the classification accuracies using some other randomized optimization techniques. Performance comparisons with some other well known approach for data classification can also be a better direction for future work.

References

[1] World Health Organization. Available: http://www.who.int
[2] American Diabetes Asscociation. Available: http://www.diabetes.org
[3] Gan, D. editor. Diabetes atlas, 2nd ed. Brussels: International Diabetes Federation, 2003. Available at http://www.eatlas.idf.org/webdata/docs/Atlas%202003-Summary.pdf2]
[4] R.O.Duda and P.E.Hard, Pattern classification and Scene Analysis, John wiley & Sons, NY, USA, 1973.
[5] Breitman,L.,Friedman,J.H.,Olshen,R.A.,C. J.,Classification and Regression tress,Wadsworth,Belmont,CA, 1984.
[6] Buntine,W.L., Learning classification trees, Statistics and Computing, 1992,pp. 63-73.
[7] Cover,T.M.,Hart,P.E., Nearest neighbors pattern classification, IEEE Trans on Information Theory, vol. 13, ,1967,pp. 21-27.
[8] Hanson R.,Stutz,J.,Cheeseman,P., Bayesian classification with correlation and

inheritance, Proceedings of the 12 th International Joint Conference on Artificial Intelligence 2, Sydney,Australia,Morgan KaufSANN, 1992,pp. 692-698.

[9] Michie,D. et al , Machine Learning, Neural and Statistical Classification, Ellis Horwood,1994.

[10]Richard,M.D, LippSANN,R.P., Neural network classifiers estimate Bayesian a-posterior probabilities, Neural Computation ,vol.3, ,1991,pp. 461-483.

[11] Tsoi, A.C et al, Comparison of three classification Techniques, CART,C4.5 and multilayer perceptrons ,Advances in Neural Information Processing Systems, vol. 3, 1991 pp.963-969.

[12]C.Bishop,Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

[13] V.N.Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.10, October 2009 115 their probabilities, Theory of Probability and its Applications, 1971, pp.264-280.

[14] D.E.Goldberg, "Genetic Algorithms in Search, Optimization and machine Learning", AddisionWesley, New York, 1989.

[15] Y. H. Pao, Adaptive Pattern Recognition and Neural Networks, Reading, MA: Addison-Wesley, 1989.

[16] S. Chen and S. A. Billings, "Neural networks for non-linear dynamic system modeling and identification," *Int. J. of Control*, vol. 56, no. 2, pp. 319-346, 1992.

[17] J. C Patra., R. N. Pal and B. N.Chatterji, "Identification of non-linear dynamic systems using functional link artificial neural networks," *IEEE Trans. on Neural Networks*, vol. 29, no. 2, pp. 254-262, 1999.

[18] A Ugena., F.de Arriaga and M. El Alami, "Speaker-independent speech recognition by means of functional-link neural networks," *Int. Conf. on* Pattern Recognition (ICPR'00)-Vol. 2, Barcelona, Spain (2000)

[19] L.H.P Harada.; A.C Da Costa.; R.M Filho., "Hybrid neural modeling of bioprocesses using functional link networks," *Applied Biochemistry and Biotechnology*, vol. 98, no. 1-3, pp. 1009-1024, 2002.]

[20] Siti Farhanah Bt Jaafar and Dannawaty Mohd Ali, "Diabetes mellitus forecast using artificial neural networks", Asian conference of paramedical research proceedings, 5-7, September, 2005, Kuala Lumpur, MALAYSIA

[21] T.Jayalakshmi and Dr.A.Santhakumaran, "A novel classification method for classification of diabetes mellitus using artificial neural networks". 2010 International Conference on Data Storage and Data Engineering.

[22] Rajeeb Dey and Vaibhav Bajpai and Gagan Gandhi and Barnali Dey, "Application of artificial neural network technique for diagnosing diabetes mellitus", 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10.

[23] Eng Khaled Eskaf, Prof. Dr. Osama Badawi and Prof. Dr. Tim Ritchings," Predicting blood glucose levels in diabetes using feature extraction and artificial neural networks".

[24]Gregory Hastings, Nejhdeh Ghevondian, "A selforganizing estimator for hypoglycemia monitoring in diabetic patients", 20 th annual international conference of IEEE engineering in medicine and biology society, Vol. 20, No 3, 1998.

[25]UCI machine learning repository and archive.ics.uci.edu/ml/datasets.html

[26] Bay, S.D., 1999. Nearest neighbor classification from multiple feature subset. Intell. Data Anal. , 3: 191-209.

Authors Profile

Manaswini Pradhan received the B.E. in Computer Science and Engineering, M.Tech in Computer Science from Utkal University, Orissa, India.She is into teaching field from 1998 to till date. Currently she is working as a Lecturer in P.G. Department of Information and Communication Technology, Orissa, India. She is currently persuing the Ph.D. degree in the P.G. Department of Information and communication Technology, Fakir Mohan University, Orissa, India. Her research interest areas are neural networks, soft computing techniques, data mining, bioinformatics and computational biology.

Dr Ranjit Kumar Sahu,, M.B.B.S, M.S. (General Surgery), M. Ch. (Plastic Surgery). Worked as an Assistant Surgeon in post doctoral department of Plastic and reconstructive surgery, S.C.B. Medical College, Cuttack, Orissa, India. Presently working as a Consultant, Tata Memorial Cancer Hospital (TMH), Mumbai, India, He has five years of research experience in the field of surgery and published many national and international papers in Medical field.