Scientific Journal Impact Factor (SJIF): 3.518

International Journal of Modern Trends in Engineering and Research

IJMTER

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

Design and Development of Hybrid Genetic Classifier Model for Prediction of Diabetes

E.Sreedevi[1] and Prof.M.Padmavathamma[2]

[1]Research Scholar, Department of Computer Science, S.V.University,Tirupati,sreedevi

[2]Professor & Head, Department of Computer Science, S.V.University, Tirupati

Abstract-Diabetes is generally a high sugar problem that doesn't make enough insulin to our body and leads to polygenic disease characterized by abnormal high glucose in the blood. By the statistical survey 95% of the diabetic cases in the world suffering with type 2 diabetes. Genetic algorithm (GA) is considered to be an optimal search algorithm to find the optimal solution by cleaning out the worse gene strings based on a fitness function. GA had established efficiency in solving the problems of unsupervised data classification. This paper proposes a new Hybrid Genetic Classifier Model (HGCM) for the prediction of type 2 diabetes by integrating different distance methods as fitness function in GA Classifier & feature Selection method for getting classification accuracy. By HGCM two rules are generated for the prediction of type 2 diabetes.

Keywords-Genetic Algorithm, KNN Algorithm, Minkowski Distance Method, Feature Selection, Diabetes Mellitus

I. INTRODUCTION

Diabetes mellitus is a chronic, lifelong condition that influences body's capacity to use the energy found in food. There are three types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes mellitus. It can lead to serious long-term complications, and increases the risk of cardiovascular disease, heart disease, stroke, kidney disease, blindness, nervous problem and feet problem. etc., with diabetes mellitus, either our body doesn't make enough insulin, or it can't use the insulin it does produce, or a combination of both.

Type 2 diabetes was also called non-insulin-dependent diabetes. The most common form of diabetes is type 2 diabetes, accounting for 95% of diabetes cases in adults all over the world.

We combine Genetic Algorithm (GA) with K-Nearest Neighbor (KNN) Approach by using different distance metrics of KNN algorithm. In our proposed work we are using Type 2 diabetes dataset with 8 attributes for prediction using Hybrid Genetic Algorithm (HGA).

II. RELATED WORK

T. Santhanam a et.al proposed genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM) as classifier for classification and K-Means clustering is used for removing the noisy data[1], Cătălin Stoean et al. proposed an Elitist Generational Genetic Chromodynamics algorithm [EGGC] for diabetes diagnosis through the means of multi model evolutionary algorithm [2]. Based on the values of eight factors, patients should be tested positive or negative for diabetes. The algorithm uses some part of the data, the training set, in order to learn the most appropriate attributes values that made doctors decide whether a patient was ill or not; this way, IF-THEN rules are built, having as the condition part the values medically leading to the conclusion part, i.e. one of the two possible outcomes. These rules are built in present algorithm in an evolutionary manner, that is, they encode chromosomes that are evolved during this training step and then applied for the classification of the rest of the data, i.e. the test set. These obtained rules are themselves of high importance, as they can also provide the reasoning rules underlying the decision-making and not only the results.

They have taken PIMA Indian dataset from UCI repository with nine attributes and calculated the fitness of the chromosomes by means of distance method given below by taking a chromosome c = (c1, c2, …, c8, c9) and a patient from the training set p = (p1, p2, …, p8, p9), the distance between c and p is computed by

$$d(c,p)=\sum_{i=1}^{8}\frac{|c_i-p_i|}{b_i-a_i}$$

Where ai and bi represent the lower and upper bounds of the i-th attribute. By means of EGGC algorithm, they have generated two rules one for each class and these are applied to test data.

E.P.Ephzibah proposed a system that solves the feature subset selection problem and proposed a model that uses the diabetes dataset and generated the best feature subset using Genetic Algorithm and fuzzy logic for effective prediction of the disease[3].

Mohammad Khanbabaei et. al proposed a new hybrid classification model based on a combination of clustering, feature selection, decision trees, and genetic algorithm techniques to pre-process the input samples to construct the decision trees in the credit scoring model. The proposed hybrid model choices and combines the best decision trees based on the optimality criteria [4]. Omar S Soliman et. al proposed a hybrid algorithm of Modified-Particle Swarm Optimization and Least Squares Support Vector Machine for the classification of type II DM patients. LS-SVM algorithm is used for classification by finding optimal hyper-plane which separates various classes & Modified-PSO algorithm is used as an optimization technique for LS-SVM parameters [5].

K.Rajeswari et.al proposed a new model for prediction of complications developing due to Diabetes Mellitus. Artificial Neural Network technique is used to predict the complications developing & focuses on modeling an effective Diagnosis of a special complication called neuropathy [6]. N.M.Lavanya et.al proposed fuzzy optimization to construct large scale knowledge based system for diagnosis of diabetes and takes PIMA Indian diabetes dataset as input [7].

Keshavamurthy B. N[19] et.al proposed the rule based genetic algorithm classifier by improve upon the fitness function parameter modification. Also, it compares the results with the probabilistic approach such as Naïve Bayes which is always gives better results and very efficient in case there is no attribute dependency in the problem, which is not true in most of the real world problem including nursery dataset we have considered for our work.

Keshavamurthy B. N[19] et.al proposed the rule based genetic algorithm classifier by improve upon the fitness function parameter modification. Also, it compares the results with the probabilistic approach such as Naïve Bayes which is always gives better results and very efficient in case there is no attribute dependency in the problem, which is not true in most of the real world problem including nursery dataset we have considered for our work.

III. FEATURE SELECTION

Feature selection is the process of selecting a subset of related features for use in model building. The central idea of using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features. The feature Selection process removes redundant features and selects best features for classification.

Feature selection is a data pre-processing technique applied to diabetes dataset. It selects subset of features from entire feature set and removes redundant features. There are mainly three approaches in feature selection they are filter method, wrapper method, and embedded methods [20].

To improve classification accuracy we apply feature selection to diabetes dataset. Feature selection method selects the optimal feature subsets of the diabetes dataset that enhances the performance of the Genetic Algorithm classifier.

IV. GENETIC ALGORITHM

Genetic Algorithms (GAs) were first suggested by John H. Holland in his book Adaption in Natural and Artificial Systems (David E. Glodberg, 2000), first published in 1975. He had noticed that simple representations in bit strings could encode complicated structures and that simple transformations could improve these structures. The genetic algorithms are adaptive techniques that can be successfully used to solve complex search and optimization problems (Maulik et.al 2000). GA has been popularly applied as a tool in order to construct solutions to problems using optimization techniques. It incorporates search techniques where it continuously tries various probable solutions with Genetic Operators like selection, Crossover and Mutation.

Selection: Selection Operator is used for selecting individuals for reproduction according to their fitness value

Crossover: Crossover Operator is for combining the genetic information of two individuals. In many respects the effectiveness of crossover is depended on coding

Mutation: Mutation takes place after Crossover to prevent declining all solutions in population into a local optimum of solved problem. Mutation changes randomly and generates new individuals.

The algorithm involves three major stages: selection, reproduction and replacement.

After an initial population is randomly generated, the algorithm evolves through three major stages: selection, crossover & Mutation. Selection which is used to find the probability of the individuals, crossover which represents mating between individuals and mutation which represents random modifications.

V. KNN ALGORITHM

KNN is a non parametric lazy learning algorithm. KNN assumes the data in a feature space. The data can be scalars or multidimensional vectors. The data points are in feature space, so there is a need of distance methods to calculate the distance of all training vectors to test vectors.

Each of the training data consists of a set of vectors and class label associated with each vector. The class label can be either Positive or Negative. KNN will work evenly well with random number of classes.

There are various distance metrics to determine the distance between query instance and training samples

a. Euclidean Distance
b. Manhattan Distance
c. Minkowski Distance
d. Chebyshev Distance

a. **Euclidean Distance:**The Euclidean Distance between points X(x1,x2,…,xn) and Y(y1,y2,…..,yn) can be defined as

$$d(x,y) = \sqrt{(x1-y1)+(x2-y2)} + ... + (xn-yn)$$
$$= \sqrt{\sum_{i=1}^{n}(xi-yi)^2}$$

Where ‘n’ is number of attributes of Xi and Yi respectively

b. **Manhattan Distance:** The Manhattan distance function calculates the distance that is to be traveled from one point to the other. The Manhattan distance between two ponts is the sum of the differences of their corresponding components. The Manhattan Distance between points X(x1,x2,…,xn) and Y(y1,y2,…..,yn) can be defined as

$$d(x,y) = (x1-y1)+(x2-y2)+....(xn-yn)$$
$$= \sum_{i=1}^{n}|x_i - y_i|$$

Where ‘n’ is number of attributes of Xi and Yi respectively

c. **Chebyshev Distance:** The Chebyshev distance function calculates the distance that is to be traveled from one point to the other. The chebyshev distance between two points X(x1,x2,…,xn) and Y(y1,y2,…..,yn) can be defined as

$$d(x,y) = \lim_{x\to\infty}\left(\sum_{i=1}^{n}|x_i - y_i|^r\right)^{1/r}$$

Where r is a parameter, n is the number of attributes Xi and Yi respectively

d. **Minkowski Distance:** Minkowski Distance is a generalization of Euclidean Distance that calculates the distance from one point to another point.The Minkowski distance between two points X(x1,x2,…,xn) and Y(y1,y2,…..,yn) can be defined as

$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^r\right)^{1/r}$$

Where r is a parameter, n is the number of attribute and Xi and Yi respectively.
Different names for the Minkowski distance takes place to form the order

- When r=1 then it is Manhattan or City Block Distance method
- When r=2 then it is Euclidean Distance Method
- When r=∞ then it is Chebyshev Distance Method

VI. HYBRID GENETIC CLASSIFIER MODEL (HGCM)

We are proposing Genetic Algorithm (GA) by using different distance metrics of KNN algorithm as fitness function and Feature Selection for selecting best features. HGCM uses GA learning and evolution to find a best transformation through the evaluation of a KNN approach. This approach is based on the common organization of a classifier system, and focuses on the classification of feature patterns.

In the proposed Hybrid Genetic Classifier Model (HGCM), first the Data set is divided into training data and Testing data using 10 –fold cross validation method. Then dataset is reduced by removing redundant features by means of feature selection. Then the selected features are then given as input to GA Classifier. GA Classifier randomly generates train data and test data. Finally GA Classifier will generate two rules for prediction of diabetes.

Feature selection is a data pre-processing step applied to diabetes dataset. It selects subset of features from whole feature set based on some statistical score and removes redundant features that do not contribute to performance. There are three main approaches in feature selection: filter, wrapper, and embedded methods. Filter methods select high ranked features based on a statistical score as a preprocessing step. Wrapper and embedded approach require considering the design of a classifier to select subset of features

To calculate the similarity or reliability among the data attributes, distance methods play a crucial role. The main reason of distance calculation in definite problem is to obtain an appropriate distance function. A distance function is a function which defines a distance between attributes of a set. In the proposed Genetic algorithm fitness of individual is calculated using Minkowski distance method.

The fitness of each individual is evaluated using KNN distance methods. The analysis has been made by taking different distance metrics like Manhattan Distance Method, Euclidean Distance Method, Chebyshev Distance Method, and Minkowski Distance Method along with existing distance method proposed by Cătălin Stoean[2] as fitness function in Genetic Algorithm. Finally, Minkowski is getting more accuracy when compared to other distance methods. So, the distance between two individuals is calculated by means of Minkowski distance method

The probability of each individual is calculated by means of Tournament Selection Method. Selection of best individuals is done through cumulative probability by generating random numbers. Subsequently perform Crossover & Mutation to get optimal solution.

Finally, two rules are obtained, one for each result. The rules are then applied to the test data. For each patient from the test set, the distance between it and each of the resulted rules is computed with minkowski distance method. The result for the test set is concluded to be the same with the result of the rule that has the nearest value for the distance between rule and the test data.

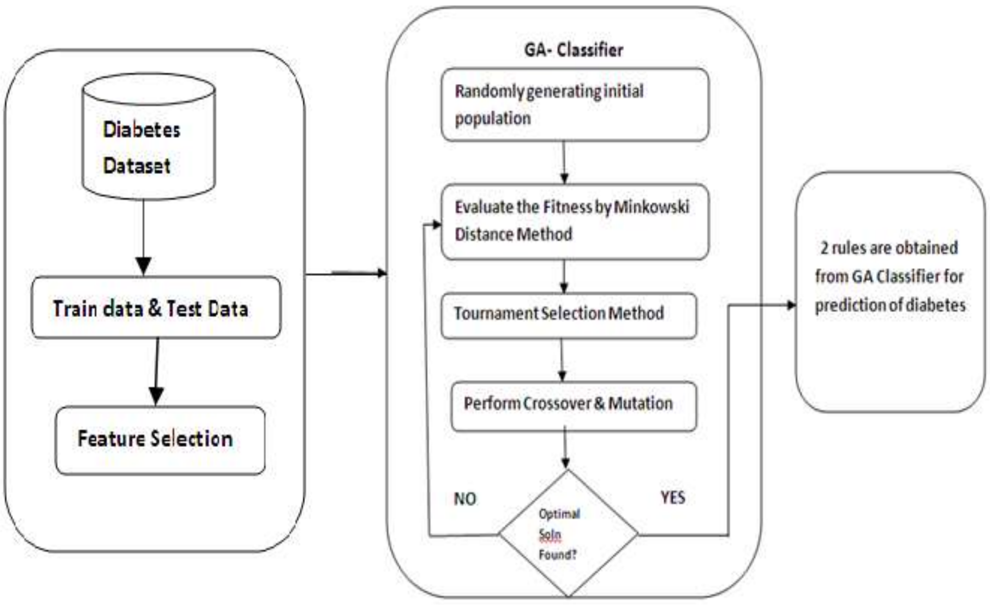
Diabetes Dataset
Train data & Test Data
Feature Selection
GA- Classifier
Randomly generating initial population
Evaluate the Fitness by Minkowski Distance Method
Tournament Selection Method
Perform Crossover & Mutation
NO
YES
Optimal Soln Found?
2 rules are obtained from GA Classifier for prediction of diabetes

Fig 1. Block diagram of Proposed Hybrid Genetic Classifier Model

The working principle of proposed system shown in Figure 1, comprises of 3 steps:

1. Data set is divided into training data and Testing data using 10 –fold cross validation method
2. Selection of attributes is done by using feature selection for dimensionality reduction
3. Dataset with reduced set of attributes given as input to the GA classifier
4. GA Classifier randomly generates train data and test data and evaluates the fitness of each individual by means of Minkowski distance method. Then calculate the probability of each chromosome by means of Tournament Selection Method. Subsequently Perform Crossover & Mutation and continue the procedure until optimal solution found.
5. Two rules will obtain from GA classifier for prediction of diabetes.

VII. PROPOSED HYBRID GENETIC ALGORITHM

In this section we describe the proposed Hybrid Genetic Algorithm for prediction of diabetes.

1. Start
2. Set t=0
3. Randomly generating initial population P(t)
4. Split the initial population into Training and testing data
5. Evaluate the fitness f(x) of each chromosome x in the population. The distance of order between two points X and Y where X Є n Chromosomes and Y Є Training set of diabetic patients. The Minkowski distance between two points X(x1,x2,…,xn) and Y(y1,y2,…..,yn) can be defined as

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}$$

6. Calculate the probability of each chromosome by means of Tournament Selection method as follows:

   6.1 Select n individuals form the population randomly

```
        6.2 Select the best individual from tournament with probability P
        6.3 Select the second best individual with probability P*(1-P)
        6.4 Select third best individual with probability P*((1-P)^2) and so on…
7.  Selection of best chromosomes using cumulative probability by generating random numbers.
8. Do crossover operation by assuming crossover rate (ρc)
        8.1 Assume crossover rate (ρc)
        8.2 Randomly generate the values 0 to 1 for all 8 attributes of population
        8.3   If random number <  ρc
        8.3.1. Select n chromosomes for crossover
        8.4 Again generate random numbers for cut point between 1 to length of
            Chromosome-1
9. Do mutation operation by assuming mutation rate (ρm)
        9.1 Calculate total length of gene in the population
        9.2 Total length of gene = No. of gene in chromosome * No. of Population
        9.3 Generate a random number between 1 and total gene length then we get mutation bit
              Points
        9.4 Random generation of positions and values to change between the range
10. t=t+1
11. Replace the current population with new population and go to step 6 until condition met
12. Formulate the rules from the obtained Genetic Algorithm
13. Apply testing data on the rules
14. Calculate the classification accuracy for all train and test datasets.
15. Stop
```

VIII. EVALUATION OF PROPOSED ALGORITHM

The diabetes dataset has been divided into training and testing datasets and were arranged by normalizing the instances of the data. By taking different distance metrics like Manhattan Distance Method, Euclidean Distance Method, Chebychev Distance Method, and Minkowski Distance Method along with existing distance method proposed by Cătălin Stoean as fitness function in Genetic Algorithm the analysis has been made. From the above said distance metrics Minkowski distance is getting more accuracy when compared to others. Hence for the proposed algorithm, we are using Minkowski distance method as fitness function.

IX. CONCLUSION

In this paper, a Hybrid Genetic Classifier Model has been proposed to the problem of diabetes diagnosis. Feature Selection method is applied to the dataset for selection of best features removing redundant features. This study has implemented Hybrid Genetic Algorithm by using Minkowski distance method as fitness function to classify the diabetes dataset. The proposed model runs iteratively by the HGCM and generates two rules for the prediction of diabetes.

REFERENCES

[1] T. Santhanam a, M.S Padmavathi b "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis"

[2] Cătălin Stoean, Ruxandra Stoean, Mike Preuss and D. Dumitrescu," Diabetes Diagnosis through the Means of a Multimodal Evolutionary Algorithm", Elsevier, pp.76-82, doi: 10.1016/j.procs.2015.03.185

[3] E.P.Ephzibah, " Cost Effective Approach On Feature Selection Using Genetic Algorithms and Fuzzy Logic For Diabetes Diagnosis", International Journal on Soft Computing (IJSC), Vol.2, No.1, February 2011

[4] Mohammad Khanbabaei and Mahmood Alborzi ,"The Use Of Genetic Algorithm,Clustering And Feature Selection Techniques In Construction Of Decision Tree Models For Credit Scoring", International Journal of

[5] Managing Information Technology (IJMIT) Vol.5, No.4, November 2013

[6] Omar S Soliman et al "Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine", International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 1– Feb 2014

[7] Rajeswari, K.; Vaithiyanathan, V.; Gurumoorthy, T.," Modeling Effective Diagnosis of Risk Complications in Type 2 Diabetes -- A Predictive model for Indian Situation", European Journal of Scientific Research;6/ 1/2011, Vol. 54 Issue 1, p147, june 2011

[8] N.M.Lavanya et.al "An expert System based on Particle Swarm Optimization and Fuzzy Technique for Medical diagnosis", International Journal of Advanced Research in Technology (IJART), vol.2,Issue 4, 2012

[9] S.Sapna, Dr.A.Tamilarasi and M.Pravin Kumar, "Implementation of Genetic Algorithm in Predicting Diabetes", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012, ISSN (Online):1694-0814

[10] Denny Hermawanto, "Genetic Algorithm for Solving Simple Mathematical Equality problem", Indonesian Institute of Sciences (LIPI), INDONESIA

[11] R. P. Ambilwade , R. R. Manza , Bharatratna P. Gaikwad ,"Medical Expert Systems for Diabetes Diagnosis: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering

[12] K. Polat, S. Gunes, A. Aslan, "A cascade learning system forclassification of diabetes disease: Generalized discriminant analysis andleast square support vector machine", Expert systems with applications, vol.34 (1),, pp. 214-221, 2008.

[13] Laetitia Jourdan, Clarisse Dhaenens, El-Ghazali Talb ,"A Genetic Algorithm for Feature Selection in Data- Mining for Genetics", MIC'2001 - 4th Metaheuristics International Conference, July 16-20, 2001.

[14] M. A. Pradhan, Abdul Rahman, Pushkar Acharya, Ravindra Gawade, Ashish Pateria , "Design of Classifier for Detection of Diabetes using Genetic Programming", International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya, pp.125-130, Dec. 2011.

[15] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[16] E.Sreedevi et. al "A Threshold Genetic Algorithm for Diagnosis of Diabetes using Minkowski Distance Method ", International Journal of Innovative Research in Science,Engineering and Technology, Vol. 4, Issue 7, July 2015

[17] http://emedicine.medscape.com/article/117853-overview

[18] http://www.webmd.com/diabetes/guide/types-of-diabetes-mellitus?page=2#2

[19] https://en.wikipedia.org/wiki/Tournament_selection

[20] Keshavamurthy B. N , Asad Mohammed Khan & Durga Toshniwal, "Improved Genetic Algorithm Based Classification" , International Journal of Computer Science and Informatics (IJCSI) ISSN (PRINT): 2231 –5292, Volume-1, Issue-3

[21] Khyati K. Gandhi, Prof. Nilesh B. Prajapati "Study of Diabetes Prediction using Feature Selection and Classification", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 2, February - 2014 , ISSN: 2278-0181