ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

A Threshold Genetic Algorithm for Diagnosis of Diabetes using Minkowski Distance Method

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

Research Scholar, Dept. of Computer Science, S.V.University, Tirupati, Andhra Pradesh, India[1]

Professor & Head, Dept. of Computer Science, S.V.University, Tirupati, Andhra Pradesh, India[2]

ABSTRACT: Diabetes is rising quickly across the globe. Early detection of diabetes helps to avoid its inception by taking preventive measures. Diagnosis plays vital role in Diabetes treatment otherwise it leads to long-term complications associated costs and risks. In the existing methods, many diabetes diagnosis algorithms relay on Genetic Algorithm with the means of Multimodal Evolutionary algorithm. In this paper, we proposed a Threshold Genetic Algorithm for diagnosis of diabetes using Minkowski Distance Method and concluded with results and analysis**.**

KEYWORDS: Diabetes mellitus, Diagnosis, Genetic Algorithm, PIMA Indian DataSet, Minkowski Distance Method

I. INTRODUCTION

Diabetes has become the fourth leading cause of death in developed countries especially in India and there is significant indication that it is reaching wave proportions in many developing and newly industrialized nations, with evidence pointing to preventable factors such as inactive lifestyle and poor diet.
According to the World Health Organization, it affects around 194 million people worldwide, and that number is expected to increase to at least 300 million by 2025. Prediction of this disease will help to prevent it in its early stage. In this paper we propose a threshold Genetic Algorithm to diagnose the diabetes using PIMA Indian dataset, which is used as an effective tool to improve the accuracy.

This paper is organized as follows: Section 2 deals with related work, Section 3 introduces about the Diabetes Mellitus, Section 4 gives brief description about Genetic Algorithms, Section 5 discussed about different distance methods, Section 6 deals with existing work, Section 7 explains the proposed Threshold Genetic Algorithm for diagnosis of diabetes using Minkowski Distance method, Section 8 gives brief description about results and analysis and finally conclusion reached to last section.

II. RELATED WORK

Cătălin Stoean et al. proposed a new fitness function for diabetes diagnosis through the means of multi model evolutionary algorithm [1]. Polatet et al. proposed two different approaches for diabetes data classification - principal component analysis and neuro-fuzzy inference and Generalized Discriminant Analysis (GDA) and least square support vector machine (LS-SVM).They achieved an accuracy of 89.47% and 79.16% respectively[8][10]. Karegowda et al. used neural networks and explained a hybrid model which uses Genetic Algorithms (GA) and Back Propagation Network (BPN) for classification of diabetes among PIMA Indians [11]. Hasan Temurtas et al.[9] proposed a neural approach for classification of diabetes data and achieved 82.37% accuracy .Muni, Pal, Das [15] discussed about a method for multiclass classifier and proposed a new concept of unfitness for improving genetic evolution. Pradhan et al. used Comparative Partner Selection (CPS) along with GP to design a 2-class classifier for detecting diabetes and discussed about GP approach for detection of diabetes [10]. Ephzibah [12] used a fuzzy rule based classification system for feature subset selection in diabetes diagnosis. This approach proves to be cost-effective. Arcanjo et al. proposed KNN-GP (KGP) algorithm, a semi-supervised transductive one, based on the three basic assumptions of semi-supervised learning. This system was implemented on 8 datasets of UCI repository but inferior results were obtained

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

for diabetes dataset [14]. Kishore et al. [7] explained an interesting method which considers a class problem as a set of two-class problems .Smart investigated the multi class approach using modified genetic operators and concluded that GP can be used to improve multi class problems [17]. Lim et al. presented an excellent comparison of 33 classification algorithms in [16]. Durga Prasad Muni et. al., proposed a large number of benchmark data sets for comparison [15].

III. DIABETES MELLITUS

Diabetes mellitus is a chronic, lifelong condition that affects your body's ability to use the energy found in food. There are three major types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes. Diabetes Mellitus occurs when the level of blood glucose is raised because the body cannot use it properly. With diabetes mellitus, either your body doesn't make enough insulin, it can't use the insulin it does produce, or a combination of both.[5]

Diabetes Mellitus usually cannot be cured but can be managed by controlling blood glucose. Poorly managed, it can lead to serious long-term complications, and increases the risk of cardiovascular disease etc., Generally high quality DM support and management is inherently a costly, complex, labor-intensive and time consuming task; involving the analysis of an array of PIMA Indian datasets collected through support from health professionals.[6]

IV. GENETIC ALGORITHMS

The term evolutionary computation (EC) has a variety of techniques and approaches based on natural selection. Genetic Algorithms (GA) is one of the important approaches that have been applied to problems based on classifier systems and evolutionary strategies. However, in recent years the field of GA has emerged as effective resources for evolving solutions to problems.

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. In the selection stage, a temporary population is created in which the fittest individuals those having to the best fitness solutions contained in the population have a higher number of chances than those having less fitness. The reproductive operators are applied to the chromosomes in the population yielding a new population. At last, chromosomes of the original population are replaced by the new created chromosomes. This replacement usually tries to keep the best chromosomes deleting the worst ones. The whole process is repeated until certain termination criterion is reached.

V. DISTANCE METHODS OVERVIEW

To compute the similarity or reliability among the data attributes, distance methods play a vital role. The main reason of distance calculation in definite problem is to obtain an appropriate distance function. A metric function or distance function is a function which defines a distance between attributes of a set. In the present paper, the Genetic algorithm fitness is calculated using Minkowski distance method. There are several distance methods available. Some of them are:

Euclidean Distance

Euclidean distance computes the root of square difference between co-ordinates of pair of objects.

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}$$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Manhattan Distance

Manhattan distance computes the absolute differences between coordinates of pair of objects

$$Dist_{XY} = |X_{ik} - X_{jk}|$$

Chebychev Distance

Chebychev Distance is also known as maximum value distance and is computed as the absolute magnitude of the differences between coordinate of a pair of objects

$$D_{\text{Chebyshev}}(p,q) := \max_i(|p_i - q_i|).$$

Minkowski Distance

The Minkowski distance is a generalized metric that includes others as special cases of the generalized form.

$$Dist_{XY} = \left(\sum_{k=1}^{d} |X_{ik} - X_{jk}|^{\frac{1}{P}}\right)^{P}$$

When p=2, the distance becomes the Euclidean distance. When p=1 it becomes city block or Manhattan distance. Chebyshev distance is a variant of Minkowski distance where p=∞ (taking a limit). This distance can be used for both ordinal and quantitative variables.

In the equation DistXY is the Minkowski distance between the data record i and j, k the index of a variable, p the total number of variables y and λ the order of the Minkowski metric. Although it is defined for any λ > 0, it is rarely used for values other than 1, 2 and ∞. As infinity can not be displayed in computer arithmetics the Minkowski metric is transformed for λ = ∞.

The Minkowski metric of the order ∞ returns the distance along that axis on which the two objects show the greatest absolute difference.

VI. EXISTING WORK

Cătălin Stoean, Ruxandra Stoean, Mike Preuss and D. Dumitrescu proposed a multimodal evolutionary algorithm which has been applied to the problem of diabetes diagnosis. They have taken data from the UCI repository of machine learning databases and originally belonged to Johns Hopkins University. The distance of the attributes is calculated by the following fitness function in Multi Evolutionary Algorithm

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

They have obtained two rules from their proposed algorithm and therefore they concluded that a sort of 'incomplete' rule (leaving out some attributes) may achieve even better than 75 or 80%.

VII. PROPOSED WORK

In our proposed work, the dataset used for diagnosis purpose is PIMA Indian diabetes dataset from UCI repository, University of California. There are 8 attributes in this dataset they are 1) No. of times pregnant, 2) Plasma glucose concentration in oral glucose tolerance test 3) Diastolic BP, 4) Triceps skin fold thickness, 5) Serum insulin, 6) Body mass index, 7)Diabetes pre degree function and 8) Age. The diagnosis is done using Genetic Algorithm as shown in following figure 1.

In this paper we are using Minkowski Distance method for calculating fitness of the individuals and Tournament Selection method for selection process.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Threshold Genetic Algorithm for Diagnosis of Diabetes using Minkowski Distance Method

1. Generate random population of P (n) chromosomes of diabetic patients to be diagnosed.
2. Evaluate the fitness f(x) of each chromosome x in the population. The distance of order between two points P and Q Where PЄ n Chromosomes and Q Є Training set of diabetic patients.
$$P = (x_1, x_2, \ldots, x_n) \text{ and } Q = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$$
is defined as:
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$
3. The probability of each chromosome is formulated by Roulette Wheel Selection Method as follows:
$$\mathbf{P(x_i) = f(x_i)/\sum_{i=1}^{n} f(x_i)}$$
4. Selection of best chromosomes for diagnosis of diabetes using cumulative probability by generating random numbers.
5. Do crossover operation by assuming crossover rate **(ρc)**
   5.1 Assume crossover rate (ρc)
   5.2 Randomly generate the values 0 to 1 for all 8 attributes of patients population
   5.3 If random number < ρc
   5.3.1. Select n chromosomes for crossover
   5.4 Again generate random numbers for cut point between 1 to length of chromosome-1
6. Do mutation operation by assuming mutation rate **(ρm)**
   6.1 Calculate total length of gene in the patient population
   6.2 Total length of gene = No. of gene in chromosome * No. of Population
   6.3 Generate a random number between 1 and total gene length then we get mutation bit points
   6.4 Random generation of positions and values to change between the range
7. Replace the current diabetic patient population with new diabetic patient population and go to step 2 until condition met.

We are proposing cycle of threshold Genetic Algorithm to optimize the diagnosis of diabetes as shown in below Fig1.

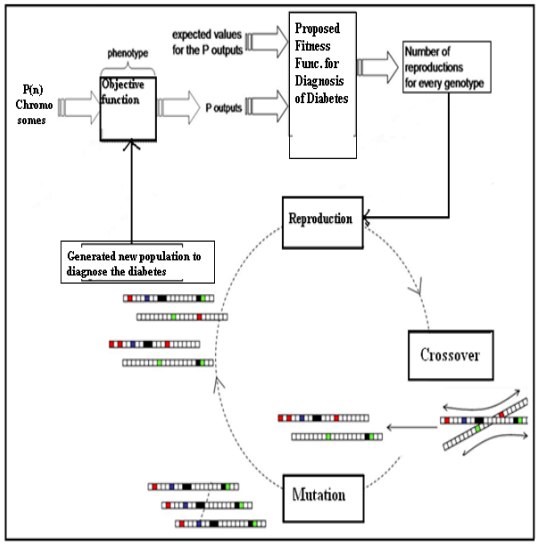

Fig 1: Cycle of Proposed Diabetes Diagnosis Algorithm

In our proposed work, we are considering P (n) chromosomes of diabetic patients to be diagnosed and evaluate the fitness of each individual using Minkowski Distance Method. Performing Reproduction, Crossover and mutation iteratively until we reach the condition and consider the new generated population to diagnose the diabetes.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

VIII. RESULTS & ANALYSIS

The PIMA Indian diabetes dataset collected from UCI repository, University of California. The training and testing datasets were arranged by normalizing the instances of the data. The analysis has been made 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. From the above said distance metrics Minkowski distance is getting more accuracy when compared to others. The Accuracy of each distance method is interpreted in tabular form as well as column chart as shown below.

Distance Method	Accuracy
Stoeans fitness function	70.305
Manhattan Function	60.031
Euclidean Function	70.076
Minkowski Function	72.214
Chebychev Function	70.763

Table1: Accuracy of distance methods

From above, table 1 show the accuracy of different distance methods and it shows that Minkowski Distance method is getting more accuracy when compared to other methods

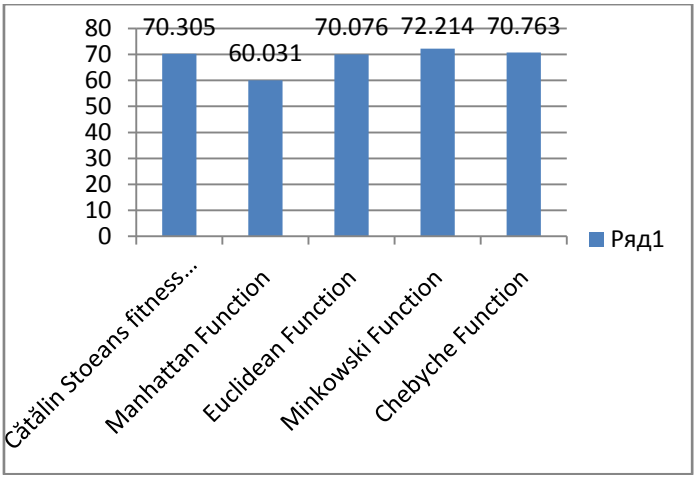

Fig 2: Column chart showing accuracy of distance methods

Figure 2 shows the Accuracy of each distance method in column chart and it shows that Minkowski distance method is getting more accurate results than existing Stoeans method.

IX. CONCLUSION

Early diagnosis of any disease with less cost is preferable. Diabetes diagnosis was done using Genetic Algorithm profitably by considering machine learning problems in the perception of medical classification. Investigation on genetic Algorithm for diagnosing diabetes using Pima Indian Diabetes data was done. In this we proposed a Threshold Genetic Algorithm to diagnose the diabetes using Minkowski Distance Method as fitness function to get optimized, effective and more accurate results. The proposed algorithm proves to be beneficial in the aspect of accuracy using

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

PIMA Indian Dataset with various distance methods. The results obtained through Minkowski distance method are comparatively better than other distance methods.

REFERENCES

[1].Cătălin Stoean, Ruxandra Stoean, Mike Preuss and D. Dumitrescu," Diabetes Diagnosis through the Means of a Multimodal Evolutionary Algorithm"
[2]. Denny Hermawanto, "Genetic Algorithm for Solving Simple Mathematical Equality Problem",Indonesian Institute of Sciences (LIPI), INDONESIA
[3]. Prof. M. A. Pradhan,, Dr. G.R. Bamnote, Vinit Tribhuvan, Kiran Jadhav,Vijay Chabukswar, "Vijay Dhobale6A Genetic Programming Approach for Detection of Diabetes", International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 6
[4]. Mitchell Melanie, "An Introduction to Genetic Algorithms "
[5]. http://www.webmd.com/diabetes/types-of-diabetes-mellitus
[6] Nonso Nnamoko, MSc; Farath Arshad, PhD; David England, PhD; Prof. Jiten Vora "Fuzzy Expert System for Type 2 Diabetes Mellitus (T2DM) Management Using Dual Inference Mechanism", Data Driven Wellness: From Self-Tracking to Behavior Change: Papers from the 2013 AAAI Spring Symposium
[6]. 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.
[7] J. K. Kishore, L. M. Patnaik, V. Mani, and V. K. Agrawal, ―"Application of genetic programming for multicategory pattern classification",‖ IEEE Trans. Evol. Comput., vol. 4, pp. 242–258, Sept. 2000.
[8] 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.
[9] T. Hasan, Y. Nijat, T. Feyzullah, "A comparative study on diabetes disease using neural networks", Expert system with applications, vol.36, pp.8610- 8615, May 2009.
[10] 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.
[11] Asha Gowda Karegowda , A.S. Manjunath , M.A. Jayaram ,"Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes" , International Journal on Soft Computing (IJSC), Vol.2, No.2, pp. 15-23, May 2011.
[12] 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, pp. 1-10, February 2011.
[13] Kemal Polata, & Salih Güne¸sa & Sülayman Tosunb, (2007), "Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre- processing", ELSEVIER,PATTERN RECOGNATION.
[14] Filipe de L. Arcanjo, Gisele L. Pappa, Paulo V. Bicalho, Wagner Meira Jr., Altigran S. da Silva, "Semi-supervised Genetic Programming for Classification", GECCO'11, July 12–16, 2011, Dublin, Ireland.
[15] Durga Prasad Muni, Nikhil R. Pal, and Jyotirmoy Das, ,"A Novel Approach to Design Classifiers Using Genetic Programming", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 8, NO. 2, pp. 183-196, APRIL 2004.
[16] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity and training time of thirtythree old and new classification algorithms", Mach. Learning J., vol. 40, pp. 203–228, 2000.
[17] William Richmond Smart, ,"Genetic Programming for Multiclass Object Classification", A thesis submitted to the Victoria University of Wellington, 2005.
[18]. Archana Singh , Avantika Yadav , Ajay Rana ,"K-means with Three different Distance Metrics", International Journal of Computer Applications (0975 – 8887), Volume 67– No.10, April 2013
[19]. Shankacharya, Devang Odedra, Subir Samanta, "Computational Intelligence _Based Diagnosis Tool for the Detection of Prediabetes and Type 2 Diabetes in India- The Review of Diabetic Studies", DIOI 10.1900/RDS.2011.9.55

BIOGRAPHY

Mrs.E.Sreedevi is presently Research Scholar in Department of Computer Science, S.V.University, tirupati. She received her MCA Degree from Jawaharlal Nehru Technological University, Hyderabad. She is currently pursuing her Ph.D from the same department under guidance of Prof.M.Padmavathamma. Her areas of interest include Artificial Intelligence , Neural Networks ,Data mining and Computer Networks..

Prof.M.Padmavathamma, Head of the Department, Dept.of Computer Science, S.V.University, tirupati. She is an eminent professor and awarded nearly 30 Ph.D and M.Phil., under her supervision. She delivered various keynote addresses on Network Security & Cryptography at various locations like Kuwait, China, Mauritius and Singapore .She is an editorial member and reviewer for various well-known journals. Her area of research includes Network Security, Privacy Preserving Data mining , Cloud Computing, Artificial Intelligence and Image Processing