Diabetes Detection using Genetic Programming

Rashmi Sonawane
ME CE Student
JSPM's BSIOTR
Wagholi, Pune. 412207

Sonali Patil
Asst. Professor
JSPM's BSOITR
Wagholi, Pune. 412207

ABSTRACT
Diabetes is a flopping of the body caused due to the absence of insulin and has gained popularity, globally. Physicians analyze diabetes using a blood glucose test; we cannot visibly categorize the person as diabetic or not based on these indicators. A pre-diabetic stage can aware the doctors and the patient about the denigrating health and can conscious the patient about the concerned measures. So proposed work intend a multi-class genetic programming (GP) based classifier design that will help the medical practitioner to confirm his/her diagnosis towards pre-diabetic, diabetic and non-diabetic patients.

This system will design in two phases, first phase consist generation of a single feature from available features using Genetic Programming from the training data. The second phase consists of use the test data for checking of the classifier. Analysis of diabetes can be complemented by this GP based classifier.

General Terms
Genetic Programming Algorithm for Classification

Keywords

Data Mining, Genetic Programming, Diabetic, Non-Diabetic, Pre-Diabetic, Classification.

1. INTRODUCTION
Physicians analyze diabetes using a blood glucose test. A blood sample is drawn and the concentration of sugar in the plasma of the blood is analyzed in a laboratory. Diagnosis of diabetes depends on many other factors and hence makes the medical practitioners' job difficult, at times. Also, many a times it is noticed that in extreme cases, the doctor has also to depend upon his previous knowledge and experience to diagnose the patient. Most of the times, doctors desire a second estimation. In addition to it all, the methods that are currently being used to diagnose diabetes prove to be costly, when a huge mass of people is to be tested. Bearing all factors in mind, a tool which enables the doctors to have a perspective of previous patients with similar conditions is necessary. The most important factors in diagnosis are data taken from the patients and an expert's opinion. A vital focus is that a pre-diabetic phase can alert the doctors and the patient about the depreciating health and can aware the patient about the concerned measures.

2. RELATED WORK
In this section we discuss about the related work done by the authors for Detection of Diabetes. A number of classification techniques have been used for the classification of diabetes disease. Many more algorithms have been proposed for the classification of diabetes data with accuracy between 59.5% and 77.7%.

Zhechen Zhu, Asoke K. Nandi, Muhammad Waqar Aslam, 2013 [1] introduced new Adapted Geometric Semantic (AGS) operators in the case where Genetic programming (GP) is used as a feature generator for signal classification. Also to control the computational complexity, a devolution scheme is introduced to reduce the solution complexity without any significant impact on their fitness.

M. W. Aslam and A. K. Nandi, 2010.[2] Proposed a system using Genetic Programming and achieved an accuracy of 78.5±2.2%. They have given the use of GP and GP with CPS and have given a comparison between the two techniques. Aslam and Nandi had pioneered the idea of CPS in order to emphasize the importance of phenotype in GP.

Muni, Pal, Das, 2004 [3] discuss GP approach to design multi class (m>2) classifiers is proposed by MuniPalDas. It needs only a single GP run to evolve an optimal classifier for a multiclass problem. For a m-class problem, a multi tree classifier consists of trees, where each tree represents a classifier for a particular class. This paper also proposes algorithms for various operations like crossover, mutation and others for a generalized n-class classification problem. To obtain a better classifier they have proposed a new scheme for OR-ing two individuals. They have used a heuristic rule-based scheme followed by a weight-based scheme to resolve conflicting situations.

M. A. Pradhan, 2011[4] converse Genetic Programming design towards the development of classifier for detection of diabetes. GP, along with Comparative Partner Selection (CPS) has been used in this paper. CPS has been used to improve the performance of the classifier. Diagnosis of diabetes can be supplemented by this GP based classifier which uses the PIMA Indians Dataset inclusive of eight attributes.

3. IMPLEMENATION PARTICULARS
3.1 Proposed System
Genetic Programming uses Evolutionary Computation and trains computational programs to take human-like decisions. Proposed system evaluates and classify diabetic patients, based on the previous knowledge discover into the system. In association with Data Mining, Genetic Programming has been used to classify a patient as pre-diabetic, diabetic or non-diabetic. This system not only provides a multiclass classifier of diabetes, but will also act as a second opinion to doctors and medical practitioners. This will help us to save important time in concern with the patients. This multiclass GP approach can efficiently co-ordinate doctors, especially the ones with no or little experience, to take major diagnostic decisions. These soft computing methodologies are complementary and although a cent per-cent accuracy is not expected, convincing results for a multiclass (pre-diabetic, diabetic, non-diabetic) classifier are promised, as multiclass classifiers are still in search of better and quicker results. Genetic Programming basically deals with optimization problems where the best result is not always expected but the better is always rewarded.Also a real time dataset of diabetes is to be used, which will differentiate the system from the previous diabetes classifiers which used the PIMA Indians

dataset.

Figure 1 Depict the architecture of the proposed system.

Eight attributes values are given as input then load the PIMA Indian Diabetes Dataset and Real time dataset, used to test the data. System is tested using train data and finally Result will show in three classes.

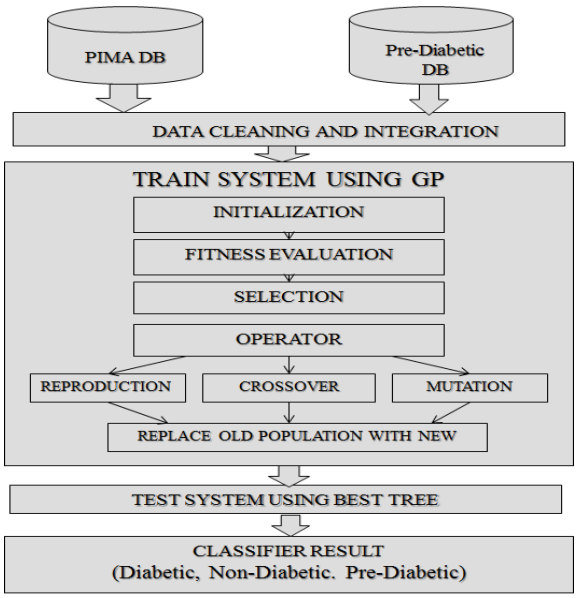


Fig.1. A schematic of multiclass classification representation

3.2 Roll of GP to Classification

3.2.1 Syntax Tree Representation

In GP, many times programs are expressed as syntax trees instead lines of code. max(x + x, x + 3 * y), this is the tree representation of program. These (x, y and 3) are variables and constants in the program and that are terminal or leaves of the tree. Arithmetic operations (+, * and max) are internal nodes called functions. The sets of allowed terminals and functions form the primitive set of a Genetic Programming system. In GP, the syntax trees are represented in prefix notation.

3.2.2 Initialization

The individual's initial population is randomly generated in GP.

Full Method

The full method is named as it generates full trees, i.e. in this method all leaves are at the same depth, the depth of a node is the number of edges that need to be traversed to reach the node starting from the tree's root node which is assumed to be at depth 0. The depth of a tree means the depth of its deepest leaf. In the full method, the initial individuals are generated so that they do not exceed a user specified maximum depth.

Grow Method

The generation of individuals in grow method resembles the full method in concern with the depth. In grow method the creation of trees are more varied sizes and shapes. Nodes get selected from primitive set (i.e., functions and terminals) until the depth limit is reached. After reaching to the depth limit, only terminals may be chosen just as in the full method.

Ramped half and half

Because neither grow nor full method provides a very wide array of sizes or shapes on their own. Koza proposed a combination called ramped half-and-half. Half the initial population is constructed using full and half is constructed using grow. This is done using a range of depth limits to help ensure that we generate trees having a variety of sizes and shapes.

3.2.3 Selection

According to the survival of the fittest principle, the best individuals among the current generation are forwarded to the next generation with the expectation to evolve the best optimum result. This process of selection of the optimum individuals is done by tournament selection. Selection of individuals is carried out over the entire population space, but in tournament selection, competition for selection is divided into large number of localized competitions, known as tournaments. Number of individuals between two and ten are selected at random from the population in each tournament selection. The individuals within the tournament then compete for a chance to be selected to pass genetic material into the next generation. Depending on tournament size, generally the best one or two individuals are selected in the tournament. Every individual program attempt several tournaments, but the programs with higher fitness values will be win more tournaments as compared to lower fitness.

3.2.4 Fitness Function

The main task of the fitness measure is to evolve the mechanism for giving a high-level statement of the problem's requirements to the GP system. Fitness can be measured in many ways in terms of: the amount of error between its output and the desired output; the amount of time required to bring a system to a desired target state; the accuracy of the program in recognizing patterns or classifying objects; the compliance of a structure with user-specified design criteria.

3.2.5 GP Operators

Unlike the arithmetic operators, GP has its own set of operators. They are discussed in this section.

Reproduction

Basically it is a random selection process. By this operator, the system selects individuals haphazardly from the mating pool. This gives rise to a new generation of individuals (chromosomes). Then the new generation is forwarded to be operated over by the rest operators.

Crossover

In crossover / recombination operator, two programs are selected from the population and then both are copied to a mating pool. In each program, crossover point is randomly chosen and the sub-trees below the crossover points are swapped. The two programs, with swapped sub-trees, are then copied to the new population. This is shown in following figure.

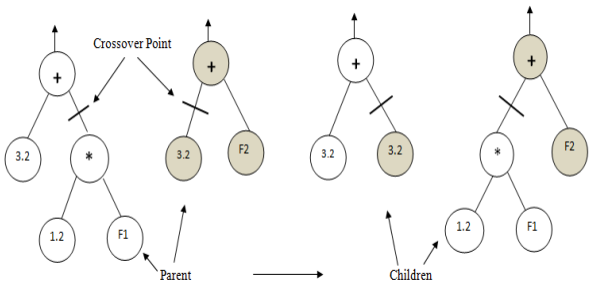


Fig.2. Crossover Genetic Operator Mutation

In mutation operator, a single program is selected from the population and that program is copied to a mating pool. A mutation point is chosen randomly, somewhere in the program, and the sub-tree below the mutation point is replaced with a new, randomly generated sub-tree. The new program is then copied into the new population. This is shown in following figure.

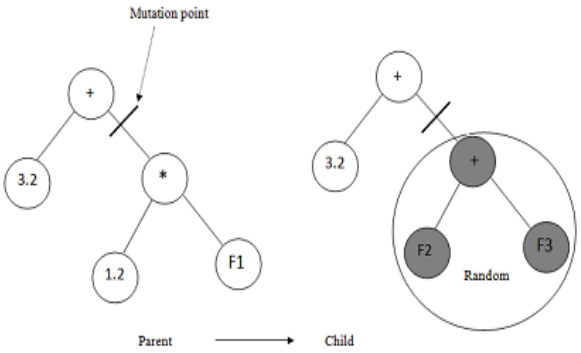


Fig.3. Mutation Genetic Operator

3.2.6 Function Set

Function set, also termed as Function pool, is typically directed by the nature of the problem domain. The function set may consist of somewhat the arithmetic functions (+, -, *, /). Table 1 shows sample of some of the functions one sees in the GP literature. Sometimes the primitive set includes specialized functions and terminals which are designed to solve problems in a specific problem domain.

Table.1. Primitives in GP function and terminal sets

Function Set	
Kind of Primitive	Examples (s)
Arithmetic	+, *, /
Mathematical	sin, cos, exp
Boolean	AND, OR, NOT
Conditional	IF-THEN-ELSE
Looping	FOR, REPEAT
.	.
.	.
Terminal Set	
Kind of Primitive	Example(s)
Variables	X, y
Constant values	3, 0.45
0-arity functions	Rand, go left

3.2.7 Classification

Classification represents the individuals in two or more than two categories. GP classifies these individuals so as to categorize them in either of the categories. In concern with Diabetes, many researches, like Rahman & Nandi, have come up with a classifier design for a two class classifier, which diagnoses the diabetic patient as diabetic or non-diabetic. Propose system focus to classify the patients in any of the three classes that are diabetic, Pre-diabetic and non-diabetic.

3.3 Algorithm

Multiclass Classifier Algorithm

1. GP begins with a randomly generated population of solutions of size N.
2. A fitness value is assigned to each solution of the population.
3. A genetic operator is selected probabilistically.

   Case i) If it is the reproduction operator, then Perform reproduction.

   Case ii) If it is the crossover operator, then Perform crossover.

   Case iii) If the selected operator is mutation, then Perform reproduction.
4. Continue Step (3), until the new population gets N solutions. This completes one generation.
5. Find Best Individual of new Population.
6. If that Best Individual has less Fitness value as compare to Old Best Individual, then do not replace Old population with New population , otherwise replace.
7. Terminate GP operation after predefine no. of generation or completion of time delay.
8. If all trainings are not correctly classified do or-ing operation.
9. To resolve conflict individual, perform weighting scheme.
10. END

3.4 State Flow Graph

The proposed model is depicted by tuple shown below:

$S= \{V, \sum, \partial, V_0, F\}$

Where, $V= \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9\}$

$\sum$= set of inputs = {number of times pregnant, plasma glucose concentration, diastolic blood pressure, Triceps skin fold thickness, serum insulin, blood mass index, diabetes pedigree function, Age}

∂= algorithm= {Genetic Programming}

V_0= initial state = v_1 {Log in and providing initial details}

F= final state = {diabetic, non-diabetic , pre-diabetic}

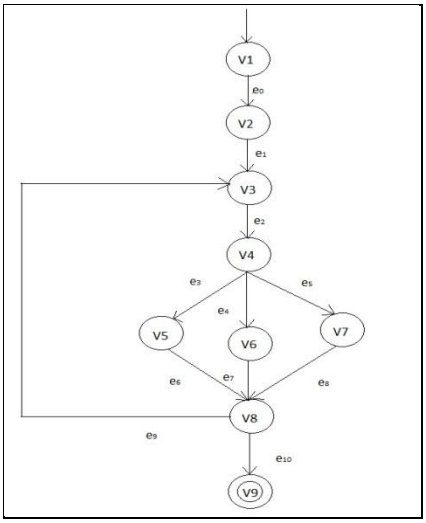


Fig. 4. State flow graph

4. RESULTS

4.1 Data Set

The PIMA Indian Diabetes Data Set is used for the

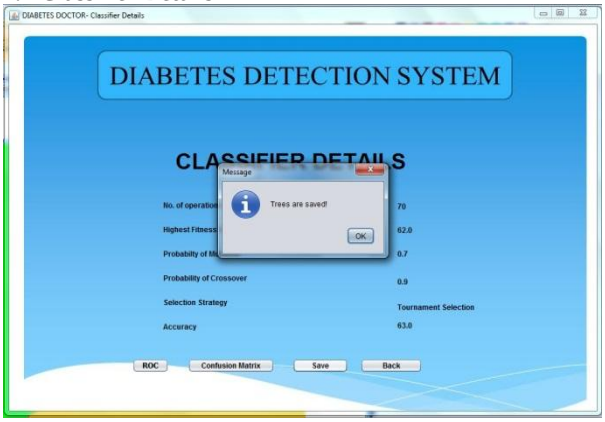
Classification of Diabetic and Non-Diabetic patient's classification. This data is retrieved from the website https://archive.ics.uci.edu. Real time data is used for Pre-Diabetic classification. This data is taken from hospital.

4.2 Results

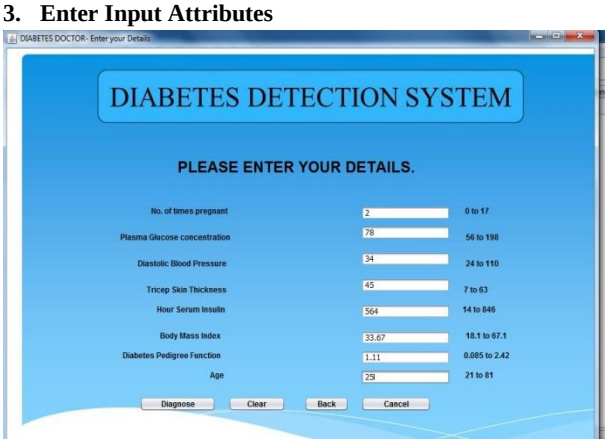
Classifier is generated by flow of algorithm after that eight attributes are given as a input to the system. That attributes are Number of Times Pregnant, Plasma Glucose Concentration, Diastolic Blood Pressure, Triceps Skin Fold Thickness, Serum Insulin, Body Mass Index, Diabetes Pedigree Function, Age and finally classifier result will be generated according to attributes.

1. Generation or loading of fittest tree

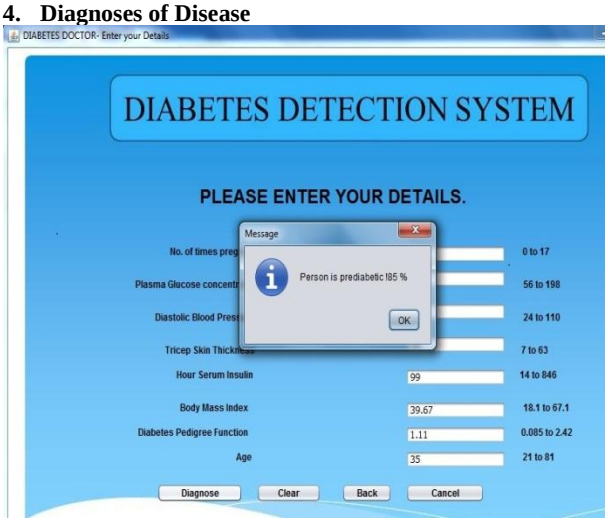
2. Classifier Details



3. Enter Input Attributes



4. Diagnoses of Disease



5. ACKNOWLEDGMENTS

The author wishes to thank Prof. Sonali Patil (Guide), Prof. A. C. Lomte (PG – Coordinator), Prof. G. M. Bhandari (HOD) and Dr. T. K. Nagaraj (Principal) for valuable guidance and encouragement. Author also thankful to National Institute of Diabetes and Digestive and Kidney Diseases for diabetes data and Unicare Hospital for Real time dataset.

6. CONCLUSION AND FUTURE WORK

A GP approach to design classifier for Diabetes Detection evolves an optimal classifier for a multiclass problem i.e. Diabetic, Non-Diabetic or Pre-Diabetic. It not only classifies a person as Diabetic or Non-Diabetic but also classifies as Pre-Diabetic and calculate in exact percent (%) of Pre-Diabetic level which leads the medical practitioner to take preventive steps towards his/her patient.

Future work will be a framework for multi disease classification can be implemented by undertaking minor changes in the current system. Extended research will be conducted to validate the propose method with more datasets. A client system model can be implemented with the server as the classifier system and the clients are used by remote end users.

7. REFERENCES

[1] Zhechen Zhu, Asoke K. Nandi, Muhammad Waqar Aslam, "Adapted Geometric Semantic Genetic Programming For Diabetes And Breast Cancer Classification", 2013 IEEE International Workshop On Machine Learning For Signal Processing, SEPT. 22–25, 2013, SOUTHAMPTON, UK.

[2] M.W. Aslam, A.K. Nandi, "Detection Of Diabetes Using Genetic Programming," *18th* European Signal Processing Conference (EUSIPCO-2010), Aalborg, Denmark, August 2010.

[3] 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, APRIL 2004, pp. 183-196.

[4] M. A. Pradhan, "Design of Classifier for Detection of Diabetes using Genetic Programming", International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011,pp.125-130.

[5] 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, pp. 1-10.

[6] Jung-Yi Lina,, Hao-RenKeb, Been-ChianChien,Wei-Pang Yang," Designing a classifier by a layered multi-population genetic programming approach", Pattern Recognition (2007) pp.2211 – 2225.

[7] 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.

[8] Guidelines on diabetes, pre-diabetes, and cardiovascular diseases, The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD), European Heart Journal.