A Genetic Programming Approach for Detection of Diabetes

Prof. M. A. Pradhan[1], Dr. G.R. Bamnote[2], Vinit Tribhuvan[3], Kiran Jadhav[4] Vijay Chabukswar[5], Vijay Dhobale[6]

[1]Asst. Prof., Department of Computer Engineering, AISSMS's College of Engineering, Pune, Maharashtra, India.
[2] Asso. Prof., Department of Computer Engineering, PRMIT & R, Amrawati, Maharashtra, India.
[3,4,5,6] Department of Computer Engineering, AISSMS's College of Engineering, Pune, Maharashtra, India.

Abstract:

Diabetes is a malfunctioning of the body caused due to the deficiency of insulin & has now-a-days gained popularity, globally. Although doctors diagnose diabetes using a blood glucose test, we cannot clearly classify the person as diabetic or not based on these symptoms. Also a pre-diabetic phase can alert the doctors and the patient about the depreciating health and can aware the patient about the concerned measures. In this paper, we propose 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.

Keywords: Classifier, Multi Class, Genetic Programming, GP, Diabetes Detection, Pre-diabetic.

1. Introduction

Diabetes has now-a-days gained global popularity. The 21st century with its sedentary lifestyle in the suburbs and a fast, social, urban lifestyle all endanger an individual's life and promote him towards diabetes. Diabetes is a malfunctioning of the body to produce insulin. Insulin is a hormone that helps the cells of the body take in sugar, or glucose, that is circulating in the blood stream and use it for energy. The American Diabetes Association estimates that 25.8 million children and adults in the United States—8.3% of the population—have diabetes and 7.0 million people are still undiagnosed[14]. Doctors diagnose 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 lab. Diagnosis of diabetes depends on many other factors and hence makes the medical practitioners job difficult, at times. One high blood sugar test is not always enough to diagnose someone with diabetes especially if the person has no other symptoms. 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. Many a times, doctors prefer a second opinion too. Bearing all factors in mind, a tool which enables the doctors to look at previous patients with similar conditions is necessary. The most important factors in diagnosis are data taken from the patients and an expert's opinion. This is a primary reason for the growth of artificial intelligence systems growing in health care industry [2]. Also a pre-diabetic phase can alert the doctors and the patient about the depreciating health and can aware the patient about the concerned measures. In this paper, we propose a multi-class genetic programming (GP) based classifier that will help the medical practitioner to confirm his/her diagnosis towards pre-diabetic, diabetic and non-diabetic patients.

2. Related Work

GP has already been used by a lot of authors to classify 2-class problems [3]–[8]. Karegowda et al. used neural networks and presented a hybrid model which uses Genetic Algorithms (GA) and Back Propagation Network (BPN) for classification of diabetes among PIMA Indians[18]. 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[15][17]. Muni, Pal, Das [22] proposed a method for multiclass classifier and introduced a new concept of unfitness for improving genetic evolution. Hasan Temurtas et al.[16] proposed a neural approach for classification of diabetes data and achieved 82.37% accuracy. Pradhan et al. used Comparative Partner Selection (CPS) along with GP to design a 2-class classifier for detecting diabetes[17]. Cheung used C4.5, Naive Bayes, BNND and BNNF algorithms and reached the classification accuracies 81.11%, 81.48%, 81.11% and 80.96%, respectively[20]. Ephzibah [19] 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 a 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 for diabetes dataset[21]. Having a look at multi-class classifiers, a few researchers [9] - [12], have had an attempt with it. Kishore et al. [9] proposed an interesting method which considers a class problem as a set of two-class problems[22]. Smart investigated the multi class approach using modified genetic operators and concluded that GP can be used to improve multi class problems [24]. Lim *et al.* presented an excellent comparison of 33 classification algorithms in [23]. They used a large number of benchmark data sets for comparison. None of these 33 algorithms use GP [22].

3. Introduction to GP

The field of evolutionary computation (EC) has a variety of alternate methods and approaches to problem solving. Some important approaches that have been applied to problems based on Genetic Algorithms (GA) classifier systems, evolutionary strategies and evolutionary programming. Both GP and GA are being used for image feature extraction, selection, and classifiers optimization. However, in recent years the field of Genetic Programming (GP) [1] has emerged as an effective means for evolving solutions to problems. GP can represent solution in the form of computer programs.

Genetic Programming utilizes similar characteristics with GAs in the fundamental processes of evolution and natural selection in order to construct solutions to problems. However, unlike GAs which use fixed sized binary strings, GPs use a variable sized tree structure. An initial population of individuals is generated and tested against the problem at hand. An objective fitness value is assigned to individuals based upon their ability to solve the problem, and then fitness proportionate selection is used to select which individuals will pass their genetic material into the next generation using genetic operators like crossover, mutation, and reproduction operations. An example of a GP tree has been shown in the figure below.

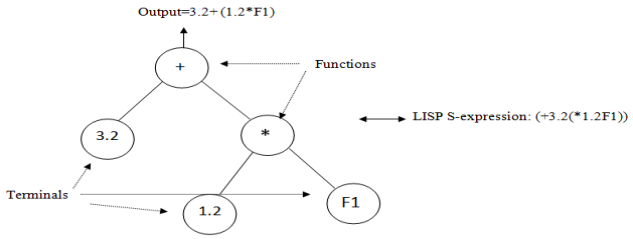

Fig. Example of Genetic Programming

3.1. GP Operators

3.1.1. Crossover

In crossover (or recombination), two programs are selected from the population, both are then copied to a mating pool. A crossover point is randomly chosen in each program, and the subtrees below the crossover points are swapped. The two programs, with swapped subtrees, are then copied to the new population. This is pictured in the figure below.

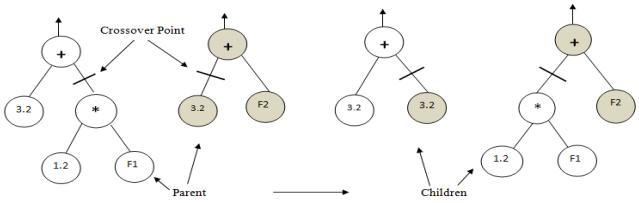

Fig. Crossover Genetic Operator

3.1.2. Mutation

In mutation, a single program is selected from the population and copied to a mating pool. A mutation point is chosen randomly, somewhere in the program, and the subtree below the mutation point is replaced with a new, randomly generated subtree. The new program is then copied into the new population. This is pictured in figure below.

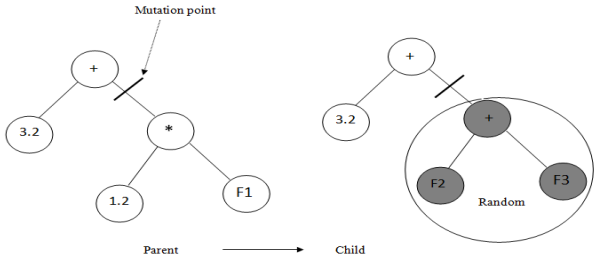

Fig. Mutation Genetic Operator

3.1.3. Reproduction

To ensure that the fitness of programs in a population is never less than that of previous generations, the reproduction, or elitism, operator is used. This consists of simply copying the best few programs of a generation's population directly to the next.

3.2. Selection in GP Evaluation

The selection process determines which individual should be passed to next generation. Fitness value is the selection key of the individuals. However, a number of methods help us to select the individuals based on the fitness values. The two primary methods used in GP are described below.

3.2.1. Roulette-wheel selection: In this method a roulette wheel is broken into a number of segments. Each program in the population takes one segment, but the size of each segment is relative to the fitness value of the program in the segment. The fittest programs take larger segments while less fit programs take smaller sections. The roulette-wheel is spun many times to select which individuals will be chosen for crossover, mutation, and reproduction operations. Every program has a chance of being selected multiple times, but the probability of more fit programs being selected is greater due to the larger segment of the wheel.

3.2.2. Tournament selection: Normally the selection of individuals is carried out over the entire population space; but in tournament selection, competition for selection is divided into a large number of localized competitions, called tournaments. In each tournament selection, a number of individuals between two and ten are selected at random from the population. The individuals within the tournament then compete for a chance to be selected to pass genetic material into the next generation. Usually only the best one or two individuals in the tournament, depending on tournament size, are selected. Each individual program compete in several tournaments, but those programs with higher fitness values would have a better chance to win more tournaments as compared to lower fitness.

3.3. Basic Terms of GP

3.3.1. Function Pool

It is a set of functions that will be used by the intermediate nodes in the structure of the tree. The function pool can contain different types of functions which are problem dependent. All these functions may have different number of inputs but always have a single output e.g. for logical problems logical functions like AND, OR, etc. are used. The function pool[2] that will be used is {+, -, *, /, square, √, sin, cos, asin, acos, log, abs, reciprocal}.

3.3.2. Fitness Function

The most significant concept of genetic programming is the fitness function. Genetic Programming can solve a number of problems; it is the fitness function which connects GP to the given problem. Each individual is assigned a fitness value by this function. It determines how well a solution is able to solve the problem. It varies greatly from one type of the problem to another. It is chosen in such a way that highly fitted solutions have high fitness value. Fitness function is the only index to select a chromosome to reproduce for the next generation.

4. Proposed System

Genetic Programming uses Evolutionary Computation and trains computational programs to take human-like decisions. In our proposed system, we consider Genetic Programming to evaluate and classify diabetic patients, based on the previous knowledge imparted 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. The 'to-be diabetic' patients can also be warned and alerted and necessary steps can be taken by the doctor towards them. This will help us to save important time in concern with the patients. As the field of fuzzy systems and logical behaviour is rapidly growing, this multiclass GP approach can efficiently co-ordinate doctors, especially the ones with no or little experience, to take major diagnostic decisions. Evolutionary Computation techniques deal with enhancing optimization, as fuzzy systems with imprecision. These soft computing methodologies are complementary and although a cent percent 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. 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.

5. Conclusion

We have proposed a GP approach to design classifiers for diabetes detection. It evolves an optimal classifier for a multiclass problem i.e, pre-diabetic, diabetic, and non diabetic. The proposed system promises to evaluate quicker and better results than those discussed in the current paper.

References

[1] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.

[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] P. J. Rauss, J. M. Daida, and S. Chaudhary, "Classification of spectral imagery using genetic programming," in *Proc. Genetic Evolutionary Computation Conf.*, 2000, pp. 733–726.

[4] D. Agnelli, A. Bollini, and L. Lombardi, "Image classification: an evolutionary approach," *Pattern Recognit. Lett.*, vol. 23, pp. 303–309, 2002.

[5] S. A. Stanhope and J. M. Daida, "Genetic programming for automatic target classification and recognition in synthetic aperture radar imagery," in *Evolutionary Programming VII*, 1998, Proc. 7th Annu. Conf. Evolutionary Programming, pp. 735–744.

[6] I. De Falco, A. Della Cioppa, and E. Tarantino, "Discovering interesting classification rules with genetic programming," *Appl. Soft Comput.*, vol.23, pp. 1–13, 2002.

[7] G. Dounias, A. Tsakonas, J. Jantzen, H. Axer, B. Bjerregaard, and D.Keyserlingk, "Genetic programming for the generation of crisp and fuzzy rule bases in classification and diagnosis of medical data," in *Proc. 1st Int. NAISO Congr. Neuro Fuzzy Technologies*, Canada, 2002, Academic Press, [CD-ROM].

[8] C. C. Bojarczuk, H. S. Lopes, and A. A. Freitas, "Genetic Programming for knowledge discovery in chest pain diagnosis," *IEEE Eng.Med. Mag.*,vol. 19, no. 4, pp. 38–44, 2000.

[9] 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.

[10] T. Loveard and V. Ciesielski, "Representing classification problems in genetic programming," in *Proc. Congr. Evolutionary Computation*,May 27–30, 2001, pp. 1070–1077.

[11] B.-C. Chien, J. Y. Lin, and T.-P. Hong, "Learning discriminant functions with fuzzy attributes for classification using genetic programming," *Expert Syst. Applicat.*, vol. 23, pp. 31–37, 2002.

[12] R. R. F. Mendes, F. B.Voznika, A. A. Freitas, and J. C. Nievola, "Discovering fuzzy classification rules with genetic programming and co-evolution," in *Lecture Notes in Artificial Intelligence*, vol. 2168, Proc. 5th Eur. Conf. PKDD, 2001, pp. 314–325.

[13] R. Poli, "Genetic Programming for image analysis," in *Proc. 1st Int.Conf. Genetic Programming*, Stanford, CA, July 1996, pp. 363–368.

[14] American Diabetes Association http://www.diabetes.org/diabetes-statistics

[15] 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), 2008, pp. 214-221.

[16] T. Hasan, Y. Nijat, T. Feyzullah, "A comparative study on diabetes disease using neural networks", *Expert system with applications*, vol.36, May 2009, pp.8610-8615.

[17] 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 Dec. 2011 , pp.125-130.

[18] 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, May 2011, pp. 15-23.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms," *Mach. Learning J.*, vol. 40, pp. 203–228, 2000.

[24] William Richmond Smart, "Genetic Programming for Multiclass Object Classification", A thesis submitted to the Victoria University of Wellington, 2005.