Diabetes Risk Factors Analysis Based on Genetic Factor Using Genetic-Pattern Matching Approach

K. Soundarrajan
Process Advisor, BARCLAYS SHARED SERVICES, DLF IT PARK, Chennai,Email: ksoundarms@gmail.com
L.Thenmozhi
Assistant Professor, Department of BCA, MGR College, Hosur. Email: thenmozhilakshmanan@gmail.com

----------**ABSTRACT**----------
Preventing the disease of diabetes is an on going area of interest to the healthcare community. Although many studies employ several Genetic-Pattern Matching Approach to assess the leading causes of diabetes, only small sets of clinical risk factors are considered. Consequently, not only many potentially important variables such as pre-diabetes health conditions are neglected in their analysis, but the results produced by such techniques may not represent relevant risk factors and pattern recognition of diabetes appropriately. In this study, we categorize our analysis into three different focuses based on the patients' healthcare costs. Predict and explain the causes of increasing diabetes in adult patients in each cost category. The preliminary analysis shows that high blood pressure, age, cholesterol, adult BMI, total income, sex , heart attack, marital status, dental checkup, and asthma diagnosis are among the key risk factors.

Keywords - Artificial Neural Network, pattern matching, genetic algorithm,
---

I.INTRODUCTION

The Preventing the disease of diabetes is an ongoing area of interest to the healthcare community. Based on the data from the 2011 National Diabetes Fact Sheet, diabetes affects an estimate of 25.8 million people in the US, which is about 8.3% of the population. Additionally, approximately 79 million people have been diagnosed with pre-diabetes [1]. Pre-diabetes refers to a group of people with higher blood glucose levels than normal but not high enough for a diagnosis of diabetes.

Increased awareness and treatment of diabetes should begin with prevention. Much of the focus has been on the impact and importance of preventive measures on disease occurrence and especially cost savings resulted from such measures. Many studies regarding diabetes prediction have been conducted for several years. The main objectives are to predict what variables are the causes, at high risk, for diabetes and to provide a preventive action toward individual at increased risk for the disease. Several variables have been reported in literature as important indicators for diabetes prediction.

Lindstrom and Tuomilehto (2003) develop the diabetes risk score model considering Age, BMI, waist circumference, history of antihypertensive drug treatment, high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables as categorical variables [2]. Park and Edington (2001) present a sequential neural network model for diabetes prediction. The authors indicate risk factors, in the final model, including blood pressure, cholesterol, back pain, fatty food, weight index or alcohol index [3]. Concaro et al, (2009) present the application of a genetic-pattern matching technique to a sample of diabetic patients. They consider the clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk in the model [4].

Although these studies employ several Genetic Algorithm to assess the leading causes of diabetes, only small sets of clinical risk factors are considered. Consequently, not only many potentially important variables such as pre- diabetes health conditions are neglected in their analysis, but the results produced by such techniques may not represent relevant risk factors and pattern recognition of diabetes appropriately.

This study seeks to fill this gap. Specially, the question arises "What are the most important risk factors to be included in prognostic analysis to prevent prevalence of diabetes?" To answer this research question, we examine whether more complex analytical models using several GP algorithm techniques can better predict and explain the causes of increasing diabetes. this project can be separated in to six phases: Genetic understanding, pattern understanding, GP analysis, common type, symptoms and causes, and Treatment

Headings

In This paper we discussed about the following terms **I.Introduction**, **II.Researchframework, III.Genetic understanding IV.Pattern understanding V.GP**

analysis with Algorithm1 and Algorithm 2, **VI.Common Types,Symptoms** and causes now we describe those things briefly.

II .RESEARCH FRAMEWORK

We provide the in-depth analysis on how Genetic-Pattern Matching approach can be a great help. After understanding the Genetic information of diabetes and developing the objectives of achieving prognostic analysis of diabetes though Genetic approach with instead of Pattern matching analysis, we begin our analysis by understanding the relevant data source, accessing data quality, and discovering first insights into the data. The next step is toward data preprocessing from the initial raw data to the final dataset, ready for the model development. This preprocessing step takes about 90% of time to understand, transform, construct, and format the relevant data. We then apply Genetic-Pattern (GP) Matching Approach to predict and explain factors that increase the prevalence of diabetes in the patient samples. However, we need to evaluate and assess the validity and the utility of our developed predictive models before deploying the Pattern Matching results into the Genetic information as stated in the objectives of the study. Figure 1 presents the overall framework of our models to address the research question from the data understanding to model deployment.

This study seeks to fill this gap. Specially, the question arises "What are the most important risk factors to be included in prognostic analysis to prevent prevalence of diabetes?" To answer this research question, we examine whether more complex analytical models using GP Algorithms and its techniques can better predict and explain the causes of increasing diabetes. This project can be separated in to six phases: Genetic understanding, pattern understanding, GP analysis, common type, symptoms and causes, and Treatment.

III .GENETIC UNDERSTANDING

In this paper we discussed about the genetic approach of the human with chromosome links which is used to identify the possible identification of diabetic symptoms

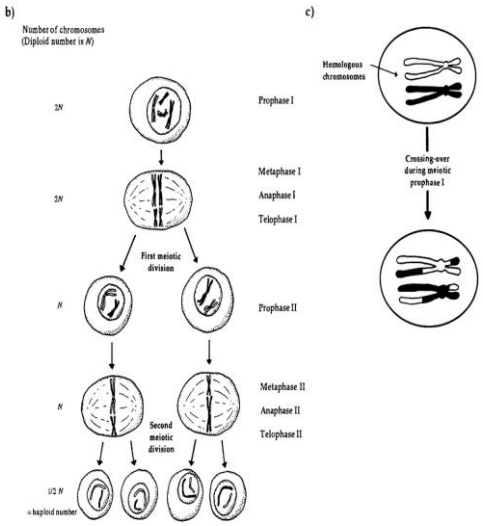

Figure 1:Human Genetics

Thus the above figure is used to know the chromosome and its genetic forum.

IV.PATTERN UNDERSTANDING

Chromosome I = gene1 gene 2…….gene j…..gene L When a binary encoding has been used every gene 1<=J<=L Є {0,1}. As an example assume L=6, then a chromosome can look like: Chromosome I =110101 B. Initialization Initialize the genes of all individuals randomly with 0's and 1's(assuming a binary encoding for simplicity). Evaluations Calculate the fitness of each individual by decoding each chromosome and applying the fitness function to each decode individuals. The decoding creates a phenotype based on a genotype. Select a specific individuals from the population to be the parents that will used to create new individuals there is many methods are used to choose those parents the most popular is the roulette wheel selection (RWS) which select the individuals with higher fitness with a higher probability("selection of the fitter individuals"). In this research we choose the parents that have specific fitness[19]Individuals from the set selected-parents are mated at random and each pair created offspring using 1-point crossover or 2-point crossover. Mutation is a random change of one or more genes. Every chromosome is simply scanned gene by gene and with a mutation rate Pm a gene is changed/swapped, i.e. 0 1 and 1 0 the probability for a mutation is usually kept small, i.e. Pm = 1/L such that we can expect one mutated gene per chromosome[20]. Criterion A simple and easy to implement stopping criterion is to stop the simple GAs if no improvement of the best solution has been made for a (large) predefined number of generations, where one generation is one turn through the do-until loop in algorithm.

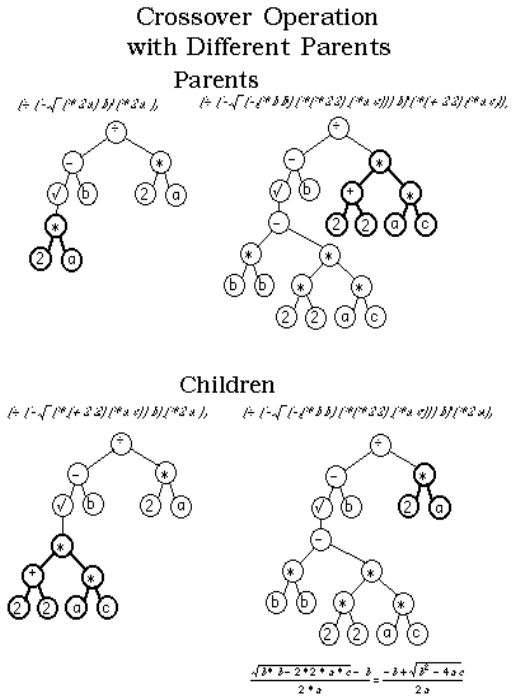

Figure 1:pattern growth

Artificial Neural Network (ANN) is a mathematical and computational model for pattern recognition and data classification through a learning process. It is a biologically inspired analytical technique, simulating biological systems, where learning algorithm indicates how learning takes place and involves adjustments to the synaptic connections between neurons. Data input can be discrete or real valued; meanwhile the output is in a form of vector of values and can be discrete or real valued as well.

V.GP ANALYSIS

In this Genetic pattern analysis can be achieved by using of various data analyzing methods it can classified using the following algorithm

ALGORITHM 1

1: find db -> des;
2: find type1,type2;
3: if (type1==fnd)
4: move dbtype1;
5: else if
6: move dbtype2;
7: else
8:move nfnd
9:stop

Which is used to identified the diabetes type identification.Here ,db means diabetes and it can be mainly grouped into type1 and type2. If type 1 fnd means type1 found move it to diabetes type1 else if If type 1 fnd means type2 found move it to diabetes type2 else move it to not found under nfnd.

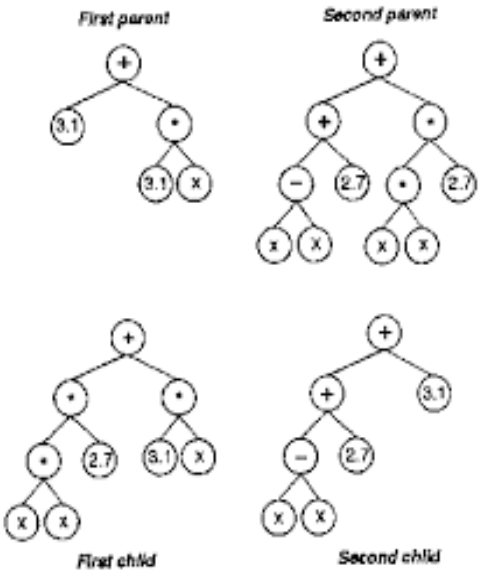

Figure 2:gp growth

ALGORITHM 2

1: 1: Procedure GP_fnd;
2: claim <-(GP_lvl,fnd , heridity(), type())
3: signed_claim <-(fnd, P(fnd))
4: a <-genitic(): (grdprt,prt, genitic(), signed_claim)
5: end procedure
6: procedure prt_path
7: **while** (no_prt)
8: find prt ()
9: pathfinder ()
10: genitic ()
11: end **while**
12: end procedure
13: procedure grdprt (m)
14: if is_claim (m) then
15: (-,-, signed_claim) <- m
16: (claim, type) <- type_claim
17: if t-found(claim, type) then
18: move lvl
19: else if incoherent_type(claim) then
20: (IDx,-,-,-)<- claim
21: trigger revocation procedure for IDx
22: return
23: end if
24: do with probability p
25: (claim, prt) <- signed_claim
26: (IDx,- , locx , timex) <- claim
27: locations <-pseudo_prt(prt, IDx, g)
28: for all L belongs to locations do
29: a ->L :(Ida ,L,is_grtprt_claim, prt_claim)
30: end for all

31: end do
32: else if is_grtprt_claim(m) then
33: (-,-,-, grtprt claim) <-m
34: (claim, grtprt) <- grtprt_claim
35: if type_fnd(fnd_claim) or replayed(claim) then
36: move lvl
37: else
38: (IDx,-, locx, timex)<-claim
39: if detect_type(m, (IDx; locx; timex)) then
40: trigger revocation procedure for IDx
41: else
42: store lvl_claim in memory
43: end if
44: end if
45: end if
46: end procedure

VI.COMMON TYPES ,SYMPTOMS AND CAUSES

Types in Diabetes Most common type of Diabetes is divided into two

Type 1 Diabetes

Type 2 Diabetes

It's called non-insulin dependent diabetes and the most common form of diabetes.: few years ago, it was rare to hear about a child with type 2 diabetes. It used to be thought that if diabetes occurred in childhood, it was type 1, or juvenile-onset, diabetes. Not anymore.

Symptoms: serious health complications.
Causes: Diabetes is a number of diseases that involve problems with the hormone insulin. Overweight, obesity and lack of physical activity are two of the most common causes of this form of diabetes.
Diabetes Prevention: possible by adopting some healthy lifestyle habits and paying attention to specific preventable diabetes complications associated with this disease.
Exercise : very important in managing diabetes. Combining diet, exercise, and medicine (when prescribed) can help control your weight and blood sugar level.
The complicity of the model is controlled by fit statistics calculated on the testing data. We use three different criteria to select the best model on the testing data. These criteria include false negative, prediction accuracy, and misclassification rate. False negative (Target = 1 and Outcome = 0) represents the case of an error in the model prediction where model results indicate that diabetes occurrence is not present, when in reality, there is an incident. The false negative value should be as low as possible. The proportion of cases misclassified is very common in the predictive modeling. However, the observed misclassification rate should be also relatively low for model justification. Lastly, prediction accuracy is evaluated among the models on the testing data. The higher the prediction accuracy rate, the better the model to be selected.

VII. RESULTS AND DISCUSSIONS

After excluding variables with outliers and high missing values, we first develop groups as presented important risk factors of diabetes are different. The data is allocated to the training (70%) and testing (30%) partitions. The binary variable of patients with diabetes (Target = 1 for patients with diabetes and Outcome = 0 for patients without diabetes) is the output variable of the prediction models. After recoding all categorical input variables, the selected variables are tested whether the association between the input variables and the logit of binary target variable satisfy the linearity assumption. The problematic variables are then transformed to satisfy such assumption. Different models are constructed and compared in order to predict patients with diabetes.

VII.CONCLUSION

Training data is only used to extract models by the GP algorithms. Pattern recognition is one of the most important stages for any pattern recognition system. Patterns recognition method, which is based on genetic algorithm, has presented. The method has been tested with a lot of patterns, high recognition rate has recorded and its recognition rate is over 95%. This proposed method is analyzing causes diabetes disease and this is used to provide a way to preventing method but it can be used for recognizing other three forms(past, present and future) with an reason belong to genetic reason and food habits are the main things.

REFERENCES

Journal Papers:

[1] Delen, D., A. Oztekin, and Z.J. Kong, A machine learning-based approach to prognostic analysis of thoracic transplantations. Artificial Intelligence in Medicine, 2010. 49(1): p. 33-42.

Books:

[1] American Diabetes Association, "Diabetes Statistics," June 02, 2012, <http://www.diabetes.org/diabetes-basics/diabetes-statistics/>.
[2] J. Lindstrom and J. Tuomilehto, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," Diabetes Care, 26:3 (2003), 725-731.
[3] J. Park and D. W. Edington, "A Sequential Neural Network Model for Diabetes Prediction," Artificial Intelligence in Medicine, 23 (2001), 277-293.
[4] S. Concaro, L. Sacchi, C. Cerra, and M. Stefanelli, "Temporal Data Mining for the Assessment of the Costs Related to Diabetes Mellitus Pharmacological

Treatment," Proc. AMIA 2009 Symposium Proceedings, 2009, pp. 119-123.
[5] American Optometric Association, "Diabetes is the Leading Cause of Blindness Among Most Adults," July 11, 2010, <http://www.aoa.org/x6814.xml>.
[6]Eurekalert, "Insufficient sleep may be linked to increased diabetes risk," July 11, 2010, <http://lifesciencelog.com/cluster53092620/>.
[7] Jackson, J., Data mining: a conceptual overview. Communications of the Association for Information Systems, 2002. 8: p. 267-296.
[8] Turban, E., R. Sharda, and D. Delen, Decision Support and Business Intelligence Systems. 2011, Pearson.
[9] Delen, D., A. Oztekin, and Z.J. Kong, A machine learning-based approach to prognostic analysis of thoracic transplantations. Artificial Intelligence in Medicine, 2010. 49(1): p. 33-42.
[10] Delen, D., A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 2010. 49(4): p. 498-506.
[11] Delen, D., G. Walker, and A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 2005. 34(2): p. 113-127.
[11] Sturat, J .Russell, and Peter Norving, "Artificial Intelligence a modern approach",2nd Edition,Preentice Hall.2003.
[12] Koray Korkut , Bilat Alatas , "Mining Classification Rule by using Genetic Algorithms with non-random initial population and uniform operator ", Turk Jelec Engine,2004, Vol.12,pp:43-52.
[13] Shyu, M. and Leou, J., "A Genetic Algorithm Approach To Color Image Enhancement", Pattern Recognition, Vol.33, No.7, PP.871-880, 1998.
[14] Jasradj U. Dange, "Introduction to Genetic Algorithms" 2001.
[15] Phili Kohn, "Combing Genetic Algorithm and Neural Networks" M.Sc. Thesis,University of Tennessee,1994.
[16] Rasheed, Sh. A., "Genetic Algorithms Application in Pattern Recognition", Master's thesis, National Computer Center Higher Education Institute ,2000.
[17] Roger L. Wainwright, "Introduction to Genetic Algorithms Theory and Applications",Addison-Wesley publishing.1993.
[18] Schamidt, M. and Stidsen, T., "Hybrid Systems: Genetic Algorithms, Neural Networks and Fuzzy Logic", Aarhus University.

Chapters in Books:

[1] Franti, P., Kivjarvi, J., Kaukoranta, T. and Nevalainen, O., "Genetic Algorithm for Large-Scale Clustering Problems", The Computer Journal, 1997 Vol.40, No.9.
[2] Mauldin, M. L., "Maintaining Diversity in Genetic Search", The National Conference on Artificial Intelligence (AAAI- 84), August 1984.
[3] Sturat, J .Russell, and Peter Norving, "Artificial Intelligence a modern approach",2nd Edition,Preentice Hall.2003.