International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Design and Development of the Novel Genetic Algorithm Framework for Chronic Kidney Disorder Classification

K. Sindhya, Dr. R. Rangaraj

Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, Tamil Nadu, India

ABSTRACT

Chronic Kidney Disorder (CKD) is a progressive loss of the renal functions. Classifying the disease with features such as blood pressure, albumin, sugar helps in diagnosing the disease. Machine learning helps in getting accuracy for the classification task. This paper implements the framework for discriminating the CKD based on genetic algorithm. Performance is compared with J48 classifier. Genetic algorithm seems to provide optimal solution for the CKD Classification.

Keywords: CKD, Genetic Algorithm, J48 Classifier, Performance Optimization.

I. INTRODUCTION

Chronic kidney disease (CKD), also known as chronic renal disease, is a progressive loss in renal function over a period of months or years. The symptoms of worsening kidney function are not specific, and might include feeling generally unwell and experiencing a reduced appetite. Often, chronic kidney disease is diagnosed as a result of screening of people known to be at risk of kidney problems, such as those with high blood pressure or diabetes and those with a blood relative with CKD [1].

This disease may also be identified when it leads to one of its recognized complications, such as cardiovascular disease, anemia, or pericarditis. It is differentiated from acute kidney disease in that the reduction in kidney function must be present for over 3 months.

CKD Features

People with CKD suffer from accelerated atherosclerosis and are more likely to develop cardiovascular disease than the general population. Patients afflicted with CKD and cardiovascular disease tend to have significantly worse prognoses than those suffering only from the latter [2].

Table 1: Feature Description

Features used and its description		
age	-	age
bp	-	blood pressure
sg	-	specific gravity
al	-	albumin
su	-	sugar
rbc	-	red blood cells
pc	-	pus cell
pcc	-	pus cell clumps
ba	-	bacteria
bgr	-	blood glucose random
bu	-	blood urea
sc	-	serum creatinine
sod	-	sodium
pot	-	potassium
hemo	-	hemoglobin
pcv	-	packed cell volume
wc	-	white blood cell count
rc	-	red blood cell count
tn	-	hypertension
dm	-	diabetes mellitus
cad	-	coronary artery disease
appet	-	appetite
pe	-	pedal edema
ane	-	anemia
class	-	class

II. Machine Learners

Machine Learning is mainly useful in cases where algorithmic/deterministic solutions are not available i.e. there is a lack of formal models or the knowledge about the application domain is scarce. The algorithms have been developed in diverse set of disciplines such as statistics, computer science, robotics, computer vision, physics, and applied mathematics. Advantages of machine learning over statistical models are accuracy, automation, speed, customizability and scalability.

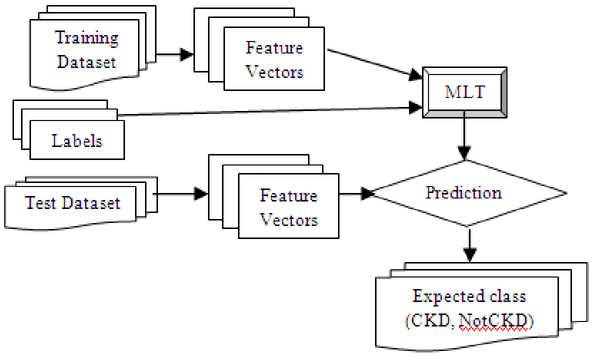

Figure 1: Machine Learning

As medicine plays a great role in human life, automated knowledge extraction from medical data sets has become an immense issue. Research on knowledge extraction from medical data is growing fast. All activities in medicine can be divided into six tasks: screening, diagnosis, treatment, prognosis, monitoring and management. As the healthcare industry is becoming more and more reliant on computer technology, machine learning methods are required to assist the physicians in identifying and curing abnormalities at early stages. Medical diagnosis is one of the important activities of medicine [3]. The accuracy of the diagnosis contributes in deciding the right treatment and subsequently in cure of diseases.

III. Genetic Algorithm and CKD

Genetic Algorithms (GA) is used as an effective search method, when the search space contains complex interacting parts. Simply saying a genetic algorithm (GA) is a search heuristic that imitates the process of natural evolution.

It is used to generate useful solutions to optimization and search problems. Genetic algorithms fit in to the larger class of evolutionary algorithms (EA), which produce solutions to optimization problems using methods inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

This GA works on search from general-to-specific rather than from simple-to-complex hypotheses. Here the GA is used to create the classification binary trees. Instead of using the binary strings, a natural representation of CKD is done with the binary tree structures with open source web mining tool GATree [6].

Since it has the ability to search complex space and find the conditionally dependent and irrelevant attributes it is possible to create a discriminating function [4].

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation.

Over successive generations, the population "evolves" toward an optimal solution. You can apply the genetic algorithm to solve a variety of optimization problems that are not well suited for standard optimization algorithms, including problems in which the objective function is discontinuous, nondifferentiable, stochastic, or highly nonlinear. The genetic algorithm can address problems of *mixed integer programming*, where some components are restricted to be integer-valued.

- The genetic algorithm uses three main types of rules at each step to create the next generation from the current population:
- *Selection rules* select the individuals, called *parents*, that contribute to the population at the next generation.
- *Crossover rules* combine two parents to form children for the next generation.
- *Mutation rules* apply random changes to individual parents to form children.
- The genetic algorithm differs from a classical, derivative-based, optimization algorithm in two main ways, as summarized in the following table.

Table 4.2: Classical Algorithm vs. Genetic Algorithm

Classical Algorithm	**Genetic Algorithm**
• Generates a single point at each iteration. The sequence of points approaches an optimal solution.	• Generates a population of points at each iteration. The best point in the population approaches an optimal solution.
• Selects the next point in the sequence by a deterministic computation.	• Selects the next population by computation which uses random number generators.

This paper proposes the GA for Chronic Kidney Disorder classification. The genetic algorithms can be used to evolve the decision trees for the closely related target concept neglecting the irrelevancy [7].

CKD classification has been done with the GA and the reason is to evolve accurate and as well as simple decision trees. Creating complex decision trees may consume time and space complexity, which decreases the performance of the decision trees. In this paper two kinds of decision tree are created.

Fig. 2 depicts the flow diagram of the GA based method for CKD [9] classification. Initially start with a population, in this experiment the population value is set to 100, 50 and 30 respectively.

Since the nature of genetic algorithm is evolutionary and because of the dynamic nature the three values are offered and tested. It is observed that when the population P_i =100 the system yields higher accuracy. After that fitness is evaluated and genetic operators are applied. Finally a good individual that better classify the CKD and NotCKD is yielded.

3.1 Overview of the Genetic Algorithm

Initially process is start with a population value, in this experiment the population value is set to 100, 50 and 30 respectively. Genetic operations are performed on the remaining dataset based on the inference obtained from the initial population. It is observed that when the population is set to 100 the system yields higher accuracy. After that, fitness is evaluated and genetic operators are applied. Finally a good individual that better classify the CKD and NO CDK is yielded and unfit individuals will be discarded.

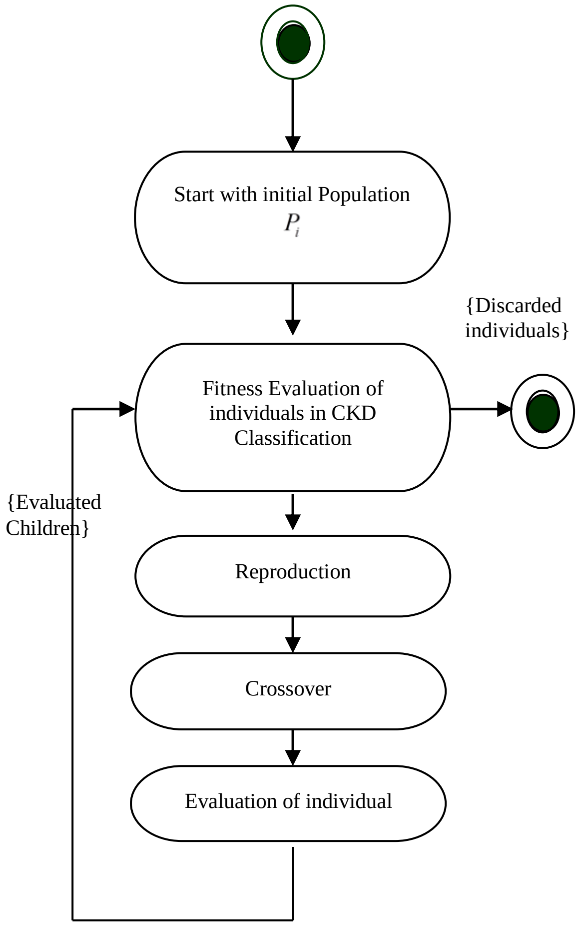

Figure 2: Overview of GA Based CKD Classification

Table 4.3: GA Experimental Settings

Generations – 100,50 and 30 (3 Iterations)
Cross over probability – 0.8
Mutation probability -0.01
Interface update – 500 millisecond
Crossover heuristic – Between good sub-
Mutation heuristics – Mutate a bad node
Percent of Gnome replacement – 0.75
Error rate – 0.6
10 fold standard cross validation

The steps involved in GA based algorithm are:

Input:

1) Training Data - Tdata
2) Total Population - P_i
3) Number of individuals - NI[P_i]
4) Maximum Generations of the population - MGP[P_i]

Output:

The individual that discriminates the CKD with higher accuracy.

Algorithm:

Step 1: Start with a randomly generated population P_i with Mutation, P_{Mut} =0.01 and Crossover, P_{cross} =0.99

Step 2: Assess the fitness value of each individual $F(I)$ in the population $I \in P_i$.

$$Fitness(SR) = \max_{Ri \in Aq}(\sigma(rr, ar_i)) \quad (1)$$

Where SR – Search results or individual, rr - relevant results and ar - all results,

$$Fitness(SR) \rightarrow [0....1] \quad (2)$$

The fitness may range from 0 to 1

Step 3: Select individuals to reproduce based on their fitness given. Compute the average fitness of all value

$$P_{\max} = \left\{ \max_{F_i} \left| F_i \in P_i \right. \right\} \quad (3)$$

Step 4: Apply crossover with probability P_{cross} =0.99

Step 5: Apply mutation with probability P_{Mut} =0.01

Step 6: Replace the population by the new generation of individuals after the evaluation

Step 7: Go to step 2

The above algorithm is an iterative one. The algorithm generates N population. Here the N is set to 100, 50 and 30 for three iterations. Decision tree is reproduced by combining the best genomes of parents to create their respective children in different iterations. The best individual in population is determined with fitness value, which is used in selection process to choose parents. The process continues until all individuals are selected from the total populations which efficiently control the time taken for tree evolution. Crossover is carried out between good sub-trees. The probability for replacing a sub-tree with another sub-tree is set to 0.8. The GA based decision tree evolved from root to leaf and hence the characteristics of children is determined by the parent. Bad node mutation is adopted with probability value 0.01, which refers the chance of a node to be altered with a new node. We have implemented that basic/standard and improved DCS algorithms using MATLAB under 32 bit Vista operating system. Experiments are conducted on a laptop with Intel(R) CoreTM 2 Duo 2.00 GHz CPU, and 3 GB of RAM. The values of parameters of the proposed algorithm are selected based on some preliminary trials (Rangaraj and Cynthiya 2017).

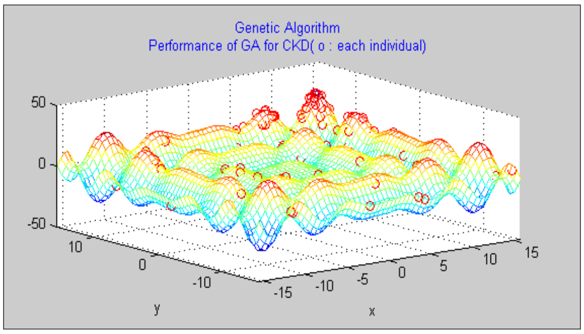

Figure 3: Performance of the GA in different populations

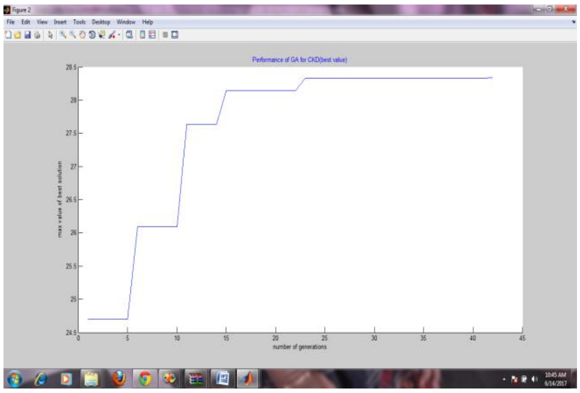

Figure 4: ROC value in GA when increasing the populations

IV. Results and Discussion

GA decision tree with higher number of initial population tends to have good inference on features and accuracy on classification. The maximum and minimum accuracy obtained in three different populations are plotted in Figure 4.5. It is evident from figure that, if number of initial population is reduced the performance degrades. Experimental results show that GA based classifier seems to be better performing for CKD classification.

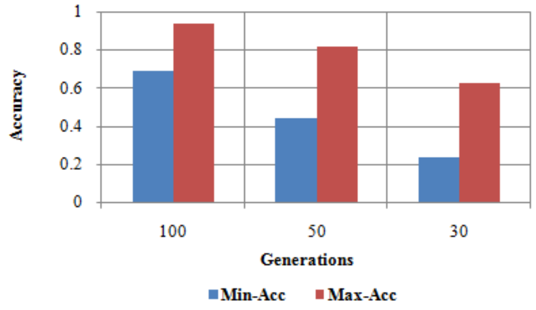

Figure 5: Minimum and Maximum Accuracy in Different Populations

V. CONCLUSION

Machine learning helps in classifying the CKD dataset to anticipate the disease. The features used are enlisted in the paper. The authors (Rangaraj and Cynthia 2017) perform an inference with the CKD features to find the best discrimination. GA algorithm has been implemented in MATLAB and the different populations are tested. Results show that the performance of the GA has been up to the mark for the classification of the CKD

VI. REFERENCES

[1]. Quinlan, J. R. C4.5,"Programs for Machine Learning", Morgan Kaufmann Publishers , 1993, pp.162-170.

[2]. Dataset:Dr.P.Soundarapandian.M.D.,D.M (Senior Consultant Nephrologist), Apollo Hospitals, L.Jerlin Rubini(Research Scholar) Alagappa University, Early stage of Indians Chronic Kidney Disease(CKD)

[3]. Amandeep Kaur, J K Sharma and Sunil Agrawal, "Optimization of Artificial Neural Networks for Cancer Detection", IJCSNS International Journal of Computer 112 Science and Network Security, VOL.11 No.5, May 2011.

[4]. A Martínez-Castelao; JL. Gorriz; J Bover; et al. (2014). "Consensus document for the detection and management of chronic kidney disease". Nefrologia. 34 (2): 243-62.doi:10.3265/Nefrologia.pre2014.Feb.12455

[5]. Moe S, Drüeke T, Cunningham J, Goodman W, Martin K, Olgaard K, Ott S, Sprague S, Lameire N, Eknoyan G (2006). "Definition, evaluation, and classification of renal osteodystrophy: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO)". Kidney Int. 69(11): 1945-53.

[6]. Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques." Artificial Intelligence Review 26.3 (2006): 159-190.

[7]. Redmon JH, Elledge MF, Womack DS, Wickremashinghe R, Wanigasuriya KP, Peiris-John RJ, Lunyera J, Smith K, Raymer JH, Levine KE (2014). "Additional perspectives on chronic kidney disease of unknown aetiology (CKDu) in Sri Lanka - lessons learned from the WHO CKDu population prevalence study". BMC Nephrology.

[8]. WEKA, "Machine learning toolkit", WEKA, www.cs.waikato.ac.nz/ml/weka/

[9]. http://en.wikipedia.org

[10]. Dr.R.Rangaraj, Ms.Cynthiya, "Classification of the Chronic Kidney Disorder using the Novel Genetic Algorithm and J48 Framework"

Appendix – A

Sample Dataset - Attributes and Values

@relation Chronic_Kidney_Disease

@attribute 'age' numeric
@attribute 'bp' numeric
@attribute 'sg' {1.005,1.010,1.015,1.020,1.025}
@attribute 'al' {0,1,2,3,4,5}
@attribute 'su' {0,1,2,3,4,5}
@attribute 'rbc' {normal,abnormal}
@attribute 'pc' {normal,abnormal}
@attribute 'pcc' {present,notpresent}
@attribute 'ba' {present,notpresent}
@attribute 'bgr' numeric
@attribute 'bu' numeric
@attribute 'sc' numeric
@attribute 'sod' numeric
@attribute 'pot' numeric
@attribute 'hemo' numeric
@attribute 'pcv' numeric
@attribute 'wbcc' numeric
@attribute 'rbcc' numeric
@attribute 'htn' {yes,no}

```
@attribute 'dm' {yes,no}
@attribute 'cad' {yes,no}
@attribute 'appet' {good,poor}
@attribute 'pe' {yes,no}
@attribute 'ane' {yes,no}
@attribute 'class' {ckd,notckd}

@data
48,80,1.020,1,0,?,normal,notpresent,notpresent,121,36,1.2,?,?,15.4,44,7800,5.2,yes,yes,no,good,no,no,ckd
7,50,1.020,4,0,?,normal,notpresent,notpresent,?,18,0.8,?,?,11.3,38,6000,?,no,no,no,good,no,no,ckd
62,80,1.010,2,3,normal,normal,notpresent,notpresent,423,53,1.8,?,?,9.6,31,7500,?,no,yes,no,poor,no,yes,ckd
48,70,1.005,4,0,normal,abnormal,present,notpresent,117,56,3.8,111,2.5,11.2,32,6700,3.9,yes,no,no,poor,yes,yes,ckd
51,80,1.010,2,0,normal,normal,notpresent,notpresent,106,26,1.4,?,?,11.6,35,7300,4.6,no,no,no,good,no,no,ckd
60,90,1.015,3,0,?,?,notpresent,notpresent,74,25,1.1,142,3.2,12.2,39,7800,4.4,yes,yes,no,good,yes,no,ckd
68,70,1.010,0,0,?,normal,notpresent,notpresent,100,54,24.0,104,4.0,12.4,36,?,?,no,no,no,good,no,no,ckd
24,?,1.015,2,4,normal,abnormal,notpresent,notpresent,410,31,1.1,?,?,12.4,44,6900,5,no,yes,no,good,yes,no,ckd
52,100,1.015,3,0,normal,abnormal,present,notpresent,138,60,1.9,?,?,10.8,33,9600,4.0,yes,yes,no,good,no,yes,ckd
53,90,1.020,2,0,abnormal,abnormal,present,notpresent,70,107,7.2,114,3.7,9.5,29,12100,3.7,yes,yes,no,poor,no,yes,ckd
50,60,1.010,2,4,?,abnormal,present,notpresent,490,55,4.0,?,?,9.4,28,?,?,yes,yes,no,good,no,yes,ckd
63,70,1.010,3,0,abnormal,abnormal,present,notpresent,380,60,2.7,131,4.2,10.8,32,4500,3.8,yes,yes,no,poor,yes,no,ckd
68,70,1.015,3,1,?,normal,present,notpresent,208,72,2.1,138,5.8,9.7,28,12200,3.4,yes,yes,yes,poor,yes,no,ckd
68,70,?,?,?,?,?,notpresent,notpresent,98,86,4.6,135,3.4,9.8,?,?,?,yes,yes,yes,poor,yes,no,ckd
68,80,1.010,3,2,normal,abnormal,present,present,157,90,4.1,130,6.4,5.6,16,11000,2.6,yes,yes,yes,poor,yes,no,ckd
40,80,1.015,3,0,?,normal,notpresent,notpresent,76,162,9.6,141,4.9,7.6,24,3800,2.8,yes,no,no,good,no,yes,ckd
47,70,1.015,2,0,?,normal,notpresent,notpresent,99,46,2.2,138,4.1,12.6,?,?,?,no,no,no,good,no,no,ckd
47,80,?,?,?,?,?,notpresent,notpresent,114,87,5.2,139,3.7,12.1,?,?,?,yes,no,no,poor,no,no,ckd
60,100,1.025,0,3,?,normal,notpresent,notpresent,263,27,1.3,135,4.3,12.7,37,11400,4.3,yes,yes,yes,good,no,no,ckd
62,60,1.015,1,0,?,abnormal,present,notpresent,100,31,1.6,?,?,10.3,30,5300,3.7,yes,no,yes,good,no,no,ckd
61,80,1.015,2,0,abnormal,abnormal,notpresent,notpresent,173,148,3.9,135,5.2,7.7,24,9200,3.2,yes,yes,yes,poor,yes,yes,ckd
60,90,?,?,?,?,?,notpresent,notpresent,?,180,76,4.5,?,10.9,32,6200,3.6,yes,yes,yes,good,no,no,ckd
48,80,1.025,4,0,normal,abnormal,notpresent,notpresent,95,163,7.7,136,3.8,9.8,32,6900,3.4,yes,no,no,good,no,yes,ckd
```