Evolutionary Continuous Genetic Algorithm for Clinical Decision Support System

Prem Pal Singh Tomar & Ranjit Singh,
Faculty of Engineering
Dayalbagh Educational Institute
Agra, India
singhppst@rediffmail.com, rsingh_dei@yahoo.com

ABSTRACT
The Medical Multimedia based Clinical Decision Support System (MM-CDSS) using the continuous genetic algorithm (CGA) methodology presents a foundation for a new technology of building intelligent computer aided diagnosis systems. The system has trained set of 280 cases for Indian heart patients for four major heart diseases: coronary heart disease, rheumatic valvular heart disease, chronic cor pulmonale, and congenital heart disease, which contain 24 symptoms for each heart disease. Using the CGA methodology, 24 critical symptom values have been identified. The best chromosome is obtained through matlabR2007a simulation. This paper presents a evolutionary continuous genetic algorithm for MM-CDSS, supporting diagnosis of four major heart diseases from their symptoms and signs through employing Microsoft Visual Basic .NET 2005 along with Microsoft SQL server 2005 environment with the advantage of Object Oriented Programming technology. A Consultant physician’s interpretation was used to evaluate the system’s Sensitivity, Specificity, Positive Prediction Value and the Negative Prediction Value. The preliminary results showed promising usage for the MM-CDSS in terms of correct and accurate diagnosis for the inexperienced physician as well as consistent and timely diagnoses, in the study of diagnostic protocol, education, self-assessment, and quality control of four major heart diseases that were investigated.

Key words: Medical Multimedia based Clinical Decision Support System, Heart Diseases Diagnosis, Continuous Genetic Algorithm, Diagnostic Features, Physician.


1. INTRODUCTION

In clinical practice, making decision involves a careful analysis of harms and benefits associated with different treatment options. These decisions, often associated with high stake and important long term consequences, are frequently made in presence of limited resources and information and an incomplete clinical picture. Under such circumstances, a rigorous and objective analysis of outcomes and probabilities is essential to achieve the best possible decision given a specific clinical situation. Therefore, physician is required to be fully conversant with the diversity of possible patterns, recognize and diagnose them, timely and accurately. Hence, a physician who is not a specialist in the pathology of the heart diseases has to refer to textbooks and study past diagnosis before concrete diagnosis can be made and conclusion reached. Hence, there is the need for a system, which can assist the physician to reach timely and accurate decision.

For the last few decades, significant efforts have been made in the field of research dedicated to using different data mining and machine learning techniques to discover useful medical knowledge and rules [1-4]. In the different data mining and machine learning techniques, the genetic algorithms (GAs) have been accepted to be dominant in medical knowledge discovery. Genetic algorithms are basically a search algorithms developed, following the principles of natural selection and natural genetics [5, 6], and have been successful in complex knowledge discovery and rule extraction problems. For example, diagnosis of hypertension based on the applied genetic algorithm was presented by Adel [7] using geometrical parameters.

Anbarasi, Anupriya and Iyengar [8] proposed a procedure using genetic algorithm and Decision Tree data mining technique to reduce the number of tests which were needed to be taken by a heart patient. Kiran and Ramesh [9] proposed a framework supported with genetic evolution to predict the heart attack.

In this paper Section 2.0 briefly describes the methodology of proposed continuous genetic algorithm for critical diagnostic symptoms value. Symptom acquisition from the system user is provided in Section 3.0 Finally, Section 4.0 presents the Results and Discussion.

2. CGA METHODOLOGY DESCRIPTIONS

Continuous genetic algorithm which is a computerized stochastic search and optimization method that works by mimicking the evolutionary principles and chromosomal processing in natural genetics. Solutions from a population are used to form a new population. Solutions that will form new solutions are selected according to their fitness: the more suitable they are, the more chances they have to reproduce.

When the variables are continuous, it is more logical to represent them by floating-point numbers. In addition, since the binary GA has its precision limited by the binary representation of variables, using floating point numbers instead easily allows representation to the machine precision. This continuous GA (CGA) also has the advantage of requiring less storage than the binary GA because a single floating-point number represents the variable instead of N\text{bits} integers. The continuous GA is inherently faster than the binary GA, because the chromosomes do not have to be decoded prior to the evaluation of the cost function.

Then three standard genetic operations, i.e., selection, crossover, and mutation are performed to produce a new generation. Such procedures are repeated until the pre-specified number of generations is achieved, or the required accuracy is satisfied. The Continuous GA algorithm for selecting critical features values for diseases diagnosis is given below. The algorithm will be stopped when the number of iterations is exceeded. It will also terminate when a chromosome in the population can recognize all the instances in the data set successfully. When the algorithm is stopped, it will output the best chromosome in $N_{\text{pop}}$.

The pseudo code for Continuous Genetic Algorithm of GA platform is given as under:

Begin (1)

Generate initial population $N_{\text{pop}}$ with $R$ random chromosomes with $m \times n$ real number (range -1 to 1) matrix.

Evaluate Fitness of all chromosomes in population $N_{\text{pop}}$

While stopping condition is not true do

Begin (2)

Rank the chromosomes in $N_{\text{pop}}$ in ascending order of their fitness

Keep the best $R/2$ chromosomes in $N_{\text{pop}}$ as elite population

For $R/2$ chromosomes

Begin (3)

Select two parent chromosomes P1 and P2 with roulette wheel strategy

Crossover operation (P1, P2)

Mutation operation

End (3)

Evaluate offspring population

Replace $N_{\text{pop}}$ with elite population + offspring population

If terminate condition is True then

Escape

End (2)

Output and save the best chromosome as solution in the database

End (1)
2.1 Encoding of Diagnostic Features

Each heart disease case includes 24 important recorded diagnostic features as shown in table-1, which are believed as critical information and symptoms required for diagnosing the chosen 4 chronic heart diseases: coronary heart disease, rheumatic valvular heart disease, chronic cor pulmonale, and congenital heart disease.

The diagnostic features are encoded using the three-value ordinal scales in terms of the degree of the seriousness. The encoded value '0' represents the absence of the attribute, '0.5' represents intermediate level of the attribute and '1' represents the largest presence of the attribute as shown in table 2.

2.2 Setting of the Parameters

The Continuous GA algorithm (CGA) for selecting critical features values for 4 chronic heart diseases diagnosis is given below. The algorithm will be stopped when the number of iterations is exceeded. It will also terminate when a chromosome in the population can recognize all the instances in the data set successfully.

When the algorithm is stopped, it will output the best chromosome in \( N_{\text{pop}} \). The parameters used in the proposed CGA algorithm are listed in Table 1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Population size</td>
<td>64</td>
</tr>
<tr>
<td>Elite population selection rate (Xrate)</td>
<td>50%</td>
</tr>
<tr>
<td>Number of generations</td>
<td>100</td>
</tr>
<tr>
<td>Crossover probability, pc</td>
<td>0.5</td>
</tr>
<tr>
<td>Mutation probability, pm</td>
<td>2%</td>
</tr>
</tbody>
</table>

Table 1: Parameters used in the proposed continuous genetic algorithm

2.3 Initial Population Generation

Generate initial population \( N_{\text{pop}} \) with \( R \) random chromosomes with \( m \times n \) continuous value number (range -1 to 1) matrix, where \( m \) and \( n \) represents the Diagnostic symptoms /facts and Chronic Heart or Lung Diseases respectively. A chromosome \( C \) of \( R \) randomly generated chromosomes, having gene values varying in the range of (-1, 1) randomly as shown below.

\[
C = \begin{bmatrix}
g_{i1} & \cdots & g_{in} \\
g_{m1} & \cdots & g_{mn}
\end{bmatrix}, \quad g_{ij} \in [-1, 1],
\]

\( 1 \leq i \leq m \) and \( 1 \leq j \leq n \)

Where, gene \( g_{ij} \) represents the \( i \)\(^{th} \) Diagnostic symptom of the \( j \)\(^{th} \) Chronic Heart or Lung Disease. The value and the sign of a gene indicate the relationship between the Diagnostic Symptom and its corresponding Chronic Heart Diseases.

There are 4 diseases decision targets and 24 numbers of diagnostic features for each disease. A randomly generated chromosome \( C \), having gene values varying in the range of (-1, 1), is then represented as \[24 \times 4\] real number matrix as shown in Figure 1, the matlabR2007a command window.

Figure 1: A Chromosome ‘C’ in CGA Model
First, second, third and forth column in figure above represents the chronic cor pulmonale, congenital heart disease, rheumatic valvular heart disease, and coronary heart disease respectively. The value and the sign of a gene indicate the relationship between the diagnostic feature and its corresponding chronic heart disease.

2.3 Chromosome Fitness Evaluation

Evaluate Fitness (A value associated with a chromosome that assigns a relative merit to that chromosome) of all the chromosomes in initial population \( N_{pop} \). The fitness function evaluates the performance of each chromosome to measure how close it is to the solution. A fitness function in CGAs is a particular type of objective function that quantifies the optimality (i.e. extent of “fitness” to the objective function) of a chromosome (solution) so that that particular chromosome may be ranked according to its fitness value against all the other chromosomes. In our diagnostic problem, it corresponds to the number of correct classifications over the whole dataset. Denote an instance \( A_t \) \((1 \leq t \leq X)\) in the data set with \( X \) instances as \( (a_{1t}, a_{2t}, \ldots, a_{kt}, \ldots, a_{mt}) \), where \( a_{kt} \) \((1 \leq k \leq m)\) is the value of Diagnostic Symptoms / facts \( k \) in instance \( A_t \).

The classification result \( f_{Rt} \) of a chromosome \( C_1 \) to instance \( A_t \) can be represented as:

\[
f_{Rt} = \max \{ a_{1t} | a_{1t} = \sum_{k=1}^{m} \xi_k a_{kt}, 1 \leq k \leq m \} \]

where, \( \xi \) = indication function for the instance \( X \) = instances in the data set

If the classification result \( f_{Rt} \) is the same as the ground truth \( \zeta_t \) of the instance \( A_t \), the indication function for the instance, denoted as \( \xi \), has a value of 1; otherwise, the value is 0. Thus, for chromosome \( C_1 \), its fitness value is obtained as:

\[
f(C) = \frac{\sum_{t=1}^{X} f_{Rt}}{X} \]

Where,

\( f(C) \) = Fitness value of chromosome \( C \)

\( \xi \) = indication function for the instance

The steps of Selection operation are as under:

It will also terminate when a chromosome in the population can recognize all the instances in the data set successfully. If the termination condition is not satisfied, then select and keep the best \((Z=R/2)\) chromosomes in initial population \( N_{pop} \) as elite population. The chromosome with the best fitness is kept from generation to generation i.e. only the best chromosomes are selected to continue, while rests of the chromosomes are deleted. Where \( 1 \leq Z \leq R/2 \). The selection rate, denoted by \( X_{rate} \), is the fraction of \( N_{pop} \) that survives for the next step of mating.

2.4 Genetic operations

Genetic operators used in genetic algorithms maintain genetic diversity. Genetic diversity or variation is a necessity for the process of evolution. Genetic operators are analogous to those which occur in the natural world: reproduction (i.e. selection), crossover (i.e. recombination), and mutation.

2.4.1 Selection of Parent Chromosomes

Select two parent chromosomes \( P_1 \) and \( P_2 \) with Roulette wheel strategy. Two chromosomes are selected from the mating pool of elite chromosomes to produce two new offspring chromosomes. Pairing takes place in the mating period until elite population offspring are born to replace the discarded chromosomes.

Assume the population \( N_{pop} \) has \( R \) chromosomes, for each chromosome \( C_y \), where \( 1 \leq y \leq R \). The average fitness function \( f(C_{xy}) \) is given by:

\[
f(C_{xy}) = \frac{\sum_{y=1}^{R} f(C_y)}{R} \]

The expected count (ec) of a chromosome \( y \) is given by

\[
\sigma e (C_y) = \frac{f(C_y)}{f(C_{xy})} \]

The probability of selection,\( \gamma_y(C_y) \), is computed as:

\[
\gamma_y(C_y) = \frac{\sigma e (C_y)}{R} \]

In Roulette wheel selection, a chromosome \( C_y \) is selected if a uniformly generated random number \( \mu \) in \([0, 1]\) satisfies the following equation:

\[
\sum_{k=0}^{y-1} \gamma_z(C_y) < \mu \leq \sum_{k=0}^{y} \gamma_z(C_y) \]

where \( y = 0 \) for \( k = 0 \)
1. Calculate the Average Fitness Value from the fitness function
2. Calculate the Expected Count = Fitness Function / Average Fitness Value
3. Probability of Selection ($\gamma_s$) = Expected count / No. of chromosomes
4. Calculate the cumulative probability ($cp_x$) = $\gamma_s$ + $cp_{x-1}$
5. Chromosome selection is done by seeing the random number $\mu$ in cumulative probability ($cp$)

As shown in the Figure 2, the first column denotes the chromosome identification number and its fitness value in second column according to the elite chromosome table, third, forth, fifth, sixth and seventh column denotes the expected count value, probability of selection, cumulative probability, random generated number and selected chromosome for mating process of continuous genetic algorithm.

**2.4.2 Crossover of Parent Chromosomes**

Two parents Figure 3 and Figure 4, mate to produce two offspring. The basic operator for producing new chromosomes in the CGA is that of crossover (recombination). Similarly as its counterpart in nature, crossover operation produces new individuals that have some parts of both parent’s genetic material. The parents have produced a total of elite population offspring, so the total chromosome population is now back to $N_{pop}$.

If two parent chromosomes $P_1$ and $P_2$ are selected, they perform crossover operation with a crossover probability ($cp$) to generate two new chromosomes $ch_1$ and $ch_2$ through the following way:

$ch_1 = S.P_1 + (1-S).P_2$
$ch_2 = (1-S).P_1 + S.P_2$
Where \( S \) is a Stochastic matrix (Figure 5) with each element valued randomly between \([0, 1]\) and the operator \(\cdot\) denotes the element-by-element matrix multiplication.

Figure 3: Parent chromosome \(P_1\) in GA Model

Figure 4: Parent chromosome \(P_2\) in GA Model
An example of the crossover operation is illustrated below:

If two parent chromosomes P1 and P2 are selected, they perform crossover operation with a crossover probability pc to generate two new child chromosomes ch1 (Figure 6) and ch2 (Figure 7) by using the crossover operation.
2.4.3 Mutation Operation

Mutation is a random process where one allele of a gene is replaced by another to produce a new genetic structure. In CGAs, mutation is randomly applied with low probability value, and modifies elements in the chromosomes. To perform the mutation operation the random substitution method is adopted i.e. the chromosome chosen to mutate is replaced by a new randomly generated chromosome C (Figure 8), having gene values varying in the range of (-1, 1) of matrix size [m x n].
2.4.4 Offspring Fitness Evaluation
A fitness function evaluates the fitness value of each child chromosome to measure how close it is to the solution in the similar manner as explained. To maintain the size of the constant original population, the numbers of new individuals are added to the numbers of elite chromosomes.

2.4.5 Best Chromosome Selection
Continue the iteration till the average fitness of chromosomes increases. Having the maximum value of fitness function, the resulting chromosome will be the answer as best chromosome. As the best chromosome is the best candidate solution, and a gene is numerically encoded, the value of each gene indicates the relationship between each diagnostic feature and its corresponding goal disease. A positive gene indicates that the corresponding feature supports the diagnosis of the disease i.e. higher value indicate the higher effect of that feature on the diagnosis; or a negative gene indicates no effect of the feature for diagnosis. If the value of a gene is negative, it is set to “0” for convenience, as the feature itself in this case has no meaningful implication in the disease classification. The evolved best chromosome (Figure 9) by the proposed continuous genetic algorithm is saved to the database.

![Figure 9: Best Chromosome](image)

3. SYMPTOM ACQUISITION FROM THE SYSTEM USER
The details about the diagnostic symptoms values (either 1, 0.5 or 0) are collected from the system users using a specially graphically designed interactive interface in Microsoft Visual Basic.Net 2005 platform as shown in table 3 [Figure 2(1) – Figure 2(9)], for the design and development of a CGA Module with the advantage of Object Oriented Programming technology. The Microsoft SQL server 2005 is used to develop the database for different 4 chronic heart diseases.

The Diagnostic symptoms values are multiplied with each column of the best chromosome and then add the multiplied values to obtain the final diagnostic resultant value corresponding to a chronic heart disease that is displayed to the system user through a graphical user interface of CGA platform as shown in Figure 10.
4. RESULTS AND DISCUSSION

In this research work 280 diagnosed Chronic Heart disease cases suffering from Coronary Artery Heart Disease, Rheumatic Valvular Heart Disease, Chronic Cor Pulmonale Heart Disease and Congenital Heart Disease were acquired from specialized database available at Bio Medical Engineering Research (BMER) Lab of Faculty of Engineering, D.E.I., Dayalbagh, Agra, India, of duration April 2004 to May 2011 and from Dr. Varun Chaudhary (Physician), M.B.B.S., M.D., Heart, Chest and Allergy Research Laboratory, Agra, India.

The 273 subjects suffering from Chronic Heart disease were selected for medical investigations and diagnosis at Bio Medical Engineering Research (BMER) Lab of Faculty of Engineering, D.E.I., Dayalbagh, Agra, India and Heart, Chest and Allergy Research Laboratory, New Agra, Dayalbagh Road, Agra, India.

Table 4 shows the diagnostic result for True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) from the developed CGA Model.

The following equations are used to calculate Sensitivity ($S_e$) and Specificity ($S_p$) [11] of CGA Model (Table 5):

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \%
\]
\[
\text{Specificity} = \frac{TN}{TN + FP} \times 100 \%
\]

REFERENCES


APPENDICES

The attributes and their encoding values of chronic heart disease diagnostic symptoms

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Chief Symptoms</th>
<th>Attributes and their corresponding encoding values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Edema</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>2</td>
<td>Breathlessness</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>3</td>
<td>Fatigue</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>4</td>
<td>Angina</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>5</td>
<td>Heart murmur</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>6</td>
<td>Syncope</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>7</td>
<td>Chronic cough(COPD)</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>8</td>
<td>Hepatomegaly</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>9</td>
<td>Hypoxia</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>10</td>
<td>Pulmonary hypertension</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>11</td>
<td>Hypertrophy &amp; dilation of RV</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>12</td>
<td>Cyanosis</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>13</td>
<td>Growth retardation</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>14</td>
<td>Fever</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>15</td>
<td>Sore throat</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>16</td>
<td>Arthalgia</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>17</td>
<td>Nausea</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>18</td>
<td>Pericardial rub</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>19</td>
<td>Anxiety</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>20</td>
<td>Family history</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>21</td>
<td>High BP</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>22</td>
<td>Diabetes (blood suger)</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>23</td>
<td>Smoking</td>
<td>Serious Medium 0.5 No</td>
</tr>
<tr>
<td>24</td>
<td>High cholesterol</td>
<td>Serious Medium 0.5 No</td>
</tr>
</tbody>
</table>

Table 3: Symptom acquisition from the system user

Figure (1): Interactive Interface for acquisition of Edema Diagnostic Symptom in CGA Model

Figure (2): Interactive Interface for acquisition of Breathlessness Diagnostic Symptom in CGA Model
Figure (3): Interactive Interface for acquisition of Fatigue Diagnostic Symptom in CGA Model

Figure (4): Interactive Interface for acquisition of Chest Discomfort Diagnostic Symptom in CGA Model

Figure (5): Interactive Interface for acquisition of Chronic Cough Diagnostic Symptom in CGA Model

Figure (6): Interactive Interface for acquisition of Hepatomagaly Diagnostic Symptom in CGA Model

Figure (7): Interactive Interface for acquisition of Hypoxia Diagnostic Symptom in CGA Model

Figure (8): Interactive Interface for acquisition of Pulmonary Hypertension Diagnostic in CGA Model
Figure (9): Interactive Interface for acquisition of Dilation of Right Ventricular Diagnostic Symptom in CGA Model

Figure 10: Diagnostic Results for Heart Diseases using CGA Model with Multimedia support tools
Table 4: Diagnostic Result from CGA Model

<table>
<thead>
<tr>
<th>Heart Diseases</th>
<th>Coronary Heart Disease</th>
<th>Rheumatic Valvular Heart Disease</th>
<th>Chronic Cor Pulmonale</th>
<th>Congenital Heart Disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary Heart Disease</td>
<td>76</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>Rheumatic Valvular Heart Disease</td>
<td>1</td>
<td>53</td>
<td>3</td>
<td>2</td>
<td>59</td>
</tr>
<tr>
<td>Chronic Cor Pulmonale</td>
<td>0</td>
<td>2</td>
<td>65</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td>Congenital Heart Disease</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>55</td>
<td>62</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>273</strong></td>
</tr>
</tbody>
</table>

Table 5: Sensitivity ($S_e$) and Specificity ($S_p$) Results for CGA Model

<table>
<thead>
<tr>
<th>Heart Diseases</th>
<th>CGA Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_e$ %</td>
<td>$S_p$ %</td>
<td></td>
</tr>
<tr>
<td>Coronary Heart Disease</td>
<td>92.68</td>
<td>98.95</td>
<td></td>
</tr>
<tr>
<td>Rheumatic Valvular Heart Disease</td>
<td>89.83</td>
<td>96.72</td>
<td></td>
</tr>
<tr>
<td>Chronic Cor Pulmonale Heart Disease</td>
<td>92.85</td>
<td>95.56</td>
<td></td>
</tr>
<tr>
<td>Congenital Heart Disease</td>
<td>88.71</td>
<td>97.15</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>91.01</strong></td>
<td><strong>97.09</strong></td>
<td></td>
</tr>
</tbody>
</table>