A Multigene Genetic Programming Model for Thyroid Disorder Detection

Joseph Ackora-Prah
Department of Mathematics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

Fidelis Nyame Oheneba-Osei
Department of Mathematics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

Perpetual Saah Andam
Department of Mathematics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

Daniel Gyamfi
Department of Mathematics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

Samuel Asante Gyamerah
Department of Mathematics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

Copyright © 2015 Joseph Ackora-Prah et al. This article is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract

Two common diseases of the thyroid gland, which releases thyroid hormones for regulating the rate of the body’s metabolism, are hyperthyroidism and hypothyroidism. Before a patient is classified as being normal or suffering from hyperthyroidism or hypothyroidism, there are a lot of information and tests conducted on the patient by existing models and these are costly in terms of time and money. We present the detection of thyroid disorder based on attributes collected from patients. A mathematical model is developed using Multigene Symbolic Regression Genetic Programming technique. The results show that the model is good and is even able to reduce the number of attributes used to classify a patient as Normal, Hyperthyroidism and Hypothyroidism.

Keywords: Hyperthyroidism, Hypothyroidism, Multigene Symbolic Regression Genetic Programming technique

1 Introduction

The human being’s neck has a thyroid gland which contains endocrine. This endocrine gland is responsible for the secretion of hormones that control the body’s metabolism and growth. A well functioning thyroid produces the right amount of hormones needed to keep the body’s metabolism working at a normal rate (not too fast or too slow). Thyroid disorders can range from a small, harmless goiter (enlarged gland) that needs no treatment to life-threatening cancer. The most common thyroid problems involve abnormal production of thyroid hormones. Over-production of thyroid hormone results in a condition known as hyperthyroidism. Insufficient hormone production leads to hypothyroidism. The ability to classify such an abnormality for proper diagnoses and treatment is crucial.

Hypothyroidism (under-active thyroid), has many causes. Some of the causes are prior thyroid surgery, exposure to ionizing radiation, chronic inflammation of the thyroid (autoimmune thyroiditis), iodine deficiency, lack of enzymes to make thyroid hormone, and various kinds of medication,[6]. Hyperthyroidism, may also be caused by inflammation of the thyroid, various kinds of medications, and lack of control of thyroid hormone production. The seriousness of thyroid disorders should not be underestimated as thyroid storm (an episode of severe hyperthyroidism) and myxedema coma (the end stage of untreated hypothyroidism) may lead to death in a significant number of cases.[6]. Recent studies have shown that evolutionary algorithms (EAs), such as genetic Programming (GP), can successfully be used as a model for thyroid disorder,[8].

Zhang [2] applied genetic programming to solve a symbolic regression. His objective was to find a function fitting the given input-output data as close as
possible. In his experiment, he first observes that the function and terminal sets had a powerful impact on the performance of the GP. Too many functions and terminals increased the search space exponentially, resulting in long training times and more local minimums, so the algorithm was more likely to be trapped by some suboptimal solutions. On the other hand, too few functions failed in finding a fitting function. He concluded that the size of population, the selection method and selection percentage are main factors that affect the variety of generations.

Muhammed et al. [7], used genetic programming (GP) and a variation of genetic programming called GP with Comparative Partner Selection (CPS) for diabetes detection. They used GP to produce an individual from training data and converted the available features to a single feature that has different values for healthy and patient (diabetes) data. They claim that their proposed system was able to achieve $78.5 \pm 2.2\%$ accuracy. Their results showed that GP based classifier can assist in the diagnosis of diabetes disease.

Again Ahlam et al. [1] used a multigene genetic programming technique to detect diabetes based on set of attributes collected from patients. Their technique showed significant advantages of an evolving nonlinear model which can be used for prediction. Their developed GP model was evaluated using Pima Indian data set and showed higher capability and accuracy in detection and diagnosis of Diabetes.

In this work, we develop a mathematical model using MGGP for the detection of thyroid disorder based on attributes collected from patients. Section 2 deals with the preliminary concepts of GP, where the concept of GP and hence Multigene symbolic regression GP is introduced. We present the Multigene Genetic Programming algorithm to the Thyroid problem in section 3. Section 4 also deals with how the Multigene Genetic Programming model was developed and finally the analysis of the developed model is presented in section 5.

2 Preliminary concepts

2.1 Genetic Programming (GP)

GP is basically a new area of artificial intelligence research. It is an algorithm inspired by biological evolution to find computer programs that perform a user defined task. It is a machine learning technique used to optimize a population of computer programs according to a fitness landscape determined by a program’s ability to perform a given computational task.

In GP, solutions are expressed as syntax trees rather than as lines of codes. In order to solve a problem, it is important to specify the steps such as Terminal set, Function set, Fitness measure, parameters like population sizes, genetic operators etc. and termination criterion.
However the two major control parameters in GP are the population size and the maximum number of generations to be run when no individual reaches the termination criterion. These two parameters depend on the difficulty of the problem to be solved. Generally, populations of 500 or more trees give better chances of finding a global optimum. For a small number of design variables, a starting population of 100 has proven to be sufficient. We illustrate a symbolic regression example for GP by predicting the numeric value of an output variable $y$ from 2 input variables $x_1$ and $x_2$. One possible symbolic representation for $y$ in terms of $x_1$ and $x_2$ would be

$$y = \frac{x_1 + x_2}{5}$$

Figure 1 demonstrates how this expression may be represented as a tree structure.

![Figure 1: Representation of a numeric expression using tree structure](image)

With this tree representation, the genetic operators of crossover and mutation are posed in a manner that allows the syntax of resulting expressions to be preserved. Figure 2 also illustrates a valid crossover operation where the two parent expressions are given in equations 1 and 2. The two offspring are given in equations 3 and 4.
A multigene individual consists of one or more genes, each of which is a traditional GP tree. The overall model is a weighted linear combination of each gene. The optimal weights for the genes are obtained using ordinary least square method to regress the genes against the output data. Mathematically, a multigene regression model can be written as:

\[ y = b_0 + b_1 \times \text{tree} + b_2 \times \text{tree} + \cdots + b_m \times \text{tree}_M \]

where \( b_0 \) is bias (offset) term, \( b_1, \ldots, b_m \) are gene weights and \( M \) is the number of genes (trees).
2.2.1 Least Square Estimates

The prediction of the y training data is estimated by:

\[ \hat{y} = b_0 + b_1 t_1 + b_2 t_2 + \cdots + b_g t_G \]  

(5)

where \( t_i \) is the \((N \times 1)\) vector of outputs from the \( i \)th tree(gene) comprising a multigene individual.

Next, define G as a \((N \times (G + 1))\) gene response matrix as follows in \( G = [1 \ t_1, \ldots, t_G] \) where the 1 refers to a \((N \times 1)\) column of ones used as a bias(offset) input. Now (5) can be rewritten as:

\[ \hat{y} = Gb \]

The least squares estimate of the coefficients \( b_0, b_1, b_2, \ldots, b_g \) formulated as a \(((G + 1) \times 1)\) vector can again be computed from the training data as

\[ b = (G^T G)^{-1} G^T \]

(6)

We note that the optimality of the estimates of \( b \) is only true if a number of assumptions are met such as independence of the columns of \( G \) and normally distributed errors. In practice the assumptions are rarely strictly met, and the columns of the gene response matrix \( G \) may be collinear (e.g. due to duplicate genes in an individual). Besides our matrix is too large which could end up to be singular and so therefore the Moore-Penrose pseudo-inverse through singular value decomposition (SVD) is used in (6) instead of the standard matrix inverse.

2.2.2 Singular value decomposition (SVD)

Generally, the SVD finds application in problems involving large matrices, with dimensions that can reach into the thousands as it can turn a singular problem into a non-singular one. SVD is based on a theorem from linear algebra which says that a rectangular matrix \( A \) can be broken down into the products of three matrices- an orthogonal matrix \( U \), a diagonal matrix \( S \), and the transpose of an orthogonal matrix \( V \).[10]

2.2.3 Theorem SVD:

**Theorem 1** Let \( A \) be a real valued \( m \times n \) matrix, where \( m \geq n \). Then \( A \) can be decomposed as follows: \( A = U S V^T \), where \( U^T U = I \), \( V^T V = I \). The columns of \( U \) are orthonormal eigenvectors of \( A A^T \), the columns of \( V \) are orthonormal eigenvectors of \( A^T A \) and \( S \) is diagonal matrix containing the square roots of eigenvalues from \( U \) or \( V \) in descending order. [11]
Hence (6) becomes

\[ b = U S^{-1} V^T \]  

(7)

3 The Multigene Genetic programming Algorithm to the Thyroid Problem

The data used was retrieved from the UCI Machine learning Repository [12]. The data has 7200 instances and consists of twenty one (21) input variables and one (1) output variable which has a value ‘1’, ‘2’ or ‘3’, where ‘1’ means the person is normal, ‘2’ implies the person is suffering from hyperthyroidism and ‘3’ means the person is suffering from hypothyroidism. The input variables (terminal set) consist of information and tests conducted on the patient as shown in Table 1.

| \(X_1\) | \(X_2\) | \(X_3\) | \(X_4\) | \(X_5\) | \(X_6\) | \(X_7\) | \(X_8\) | \(X_9\) | \(X_{10}\) | \(X_{11}\) | \(X_{12}\) | \(X_{13}\) | \(X_{14}\) | \(X_{15}\) | \(X_{16}\) | \(X_{17}\) | \(X_{18}\) | \(X_{19}\) | \(X_{20}\) | \(X_{21}\) |
| Age | Sex | On thyroxine | Query on thyroxine | On antithyroid medication | Sick | Pregnant | Thyroid surgery | I\(I\)3\(I\) treatment | Query on hypothyroid | Query on hyperthyroid | Lithium | Goiter | Tumor | Hypopituitary | Psych | TSH | T3 | TT4 | T4U | FTI |

1. We input a data file containing \(m \times n\) matrix of input (predictor) variables, and \(m \times 1\) matrix of output (response) variables.
2. An initial population of candidate solutions are generated randomly based on step one by using Ramped-half-and-half method (e.g. If maximum tree depth is 5 and population is 100 then, on average, 20 will be generated at each depth (1/2 using grow and 1/2 using full) after we had chosen our function and terminal sets. The maximum depth of each gene is then set to 5 to control the complexity of the model.

3. The number of genes (Gmax) is set to 8 as a high Gmax may capture more non-linear behavior but there is the risk of over-fitting the training data and creating models that contain complex terms that contribute little or nothing to the model’s predictive performance. Conversely, setting Gmax to 1 is equivalent to performing the scaled symbolic regression (where we have only the bias term $b_0$ and a scaling term $b_1$)

4. Our fitness measure is to minimize the error between the predicted values and the actual observed values, hence we then evaluated the fitness of each individual solutions in the population ($p$) by using:

$$RMSE = \sqrt{\frac{\sum_{p=1}^{p} (F_p - \tilde{F}_p)^2}{p}}$$

where, $F_p$ = actual values observed and $\tilde{F}_p$ = predicted values.

5. We create a new population by repeating the following steps until the new population is complete.

(a) We set a tournament size of 7 and apply Lexicographic Tournament pressure as our selection method to select the more fitter solutions into the next generation, i.e

Lexicographic Tournament selection = pick $\max\{F(G_1), F(G_2), \ldots, F(G_7)\}$

if $F(G_1) = F(G_2) = \cdots = F(G_7)$

pick $\min\{S(G_1), S(G_2), \ldots, S(G_7)\}$

Where,

$F(G_n)$ = Fitness level of gene $n$

$S(G_n)$ = Size of gene $n$

(b) We apply two point high level crossover with a probability of 0.85 to the selected individual solutions.

(c) A subtree mutation is applied with the probability of 0.1
(d) Also we apply the direct reproduction operator (elitism) with a probability of 0.05. i.e. the entire individual is simply copied to the next generation without modification.

(e) We then place new offspring in the new population.

6. The new generated population is used to generate subsequent generations per the continuation of the algorithm.

7. If the end condition is satisfied, stop and return the best solution in current population.

8. Otherwise go to step 4 for fitness evaluation (loop).

The maximum number of generation we used is 250 and an initial population of size 1500.

4 The Developed Multigene Genetic Programming Model

GPTIPS, a Genetic Programming toolbox for use with MATLAB, framework was used to apply GP in the experiment designed in this research. The data set was loaded into GPTIPS framework[9], then a symbolic regression via GP was applied with parameter set (preparatory steps) shown in Table 2. The cross validation was tuned to 60% for training and 40% for testing. Initially the function pool that was used is

\{+, -, *, /, \sqrt{}, \text{square}, \text{sin}, \text{cos}, \text{arcsin}, \text{arccos}, \text{log}, \text{abs}, \text{reciprocal}, \text{tanh}\}

After several runs with different combination of function sets, the function set that achieved best solution is shown in Table 2 below. And the best eight genes together with their weights are given in Figure 3 through to Figure 6.
Table 2: GP Preparatory steps

<table>
<thead>
<tr>
<th>Population size</th>
<th>1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of generation</td>
<td>250</td>
</tr>
<tr>
<td>Selection Method</td>
<td>Lexicographic tournament selection</td>
</tr>
<tr>
<td>Tournament size</td>
<td>7</td>
</tr>
<tr>
<td>Termination criteria</td>
<td>0</td>
</tr>
<tr>
<td>Maximum depth of each tree</td>
<td>7</td>
</tr>
<tr>
<td>Maximum Number of Genes</td>
<td>8</td>
</tr>
<tr>
<td>Function set</td>
<td>{+,-,*,tanh,cos,}</td>
</tr>
<tr>
<td>Crossover probability</td>
<td>0.85</td>
</tr>
<tr>
<td>Mutation probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Elite probability</td>
<td>0.05</td>
</tr>
</tbody>
</table>

\[1.226\tanh((X_2 + 9.515)(X_3 + X_8 X_17))\]

\[1.396\cos(\cos(X_{21}))\tanh(X_{17})\tanh(X_{18})\]

Figure 3: Gene 1 & 2
4.1 The Complete Model

After linear combination of all the eight (8) genes together with their weights, the overall simplified multigene symbolic regression model was harvested as
5 Analysis of Results

We did statistical analysis of the developed GP model to test the goodness of fit of our model. From Figure 7, the RMSE for the training data set was 0.10881 and had 91.1403% variation explained. The RMSE for the test data set was 0.1703 and had 78.6924% variation explained. The model gave the coefficient of determination ($R^2$) = 0.91151 and the Adjusted $R^2 = 0.9114$. This shows that there is a good correlation between the variable of interest ($y$) and the independent variables ($x_i$). The regression coefficients were estimated by the method of least squares and are shown in Figure 8 as Gene weights.

We also plotted a scatter diagram to have a visual indication of the degree of the association between the response variables, that is the predicted ($\hat{y}$) and the actual ($y$). The line drawn passed through the points plotted for the
A MGGP model for thyroid disorder detection

Figure 7: RMS error

Figure 8: $R^2$ and adjusted $R^2$

output (i.e. Normal(1), Hyperthyroidism(2) and Hypothyroidism(3)) as shown in figure 9. This means that there is a positive association between the $(y)$ and $(\hat{y})$.

The developed GP model used only nine (9) attributes out of the twenty one (21) to detect the thyroid disorder and these are $x_3$, $x_8$, $x_9$, $x_{15}$, $x_{17}$, $x_{18}$, $x_{19}$, $x_{20}$ and $x_{21}$. The heights of the bars in the histogram shown in figure 10.
corresponds to the degree or how much the predictor variables \( (x_i) \) influence the response variables \( (y) \). The longer the height of a predictor variable, the greater the influence (i.e. most important variable). Out of the nine (9) nine attributes that GP used, TSH\( (x_{17}) \) was the most influential factor on the response variable followed by \( x_3 \) (On thyroxine), \( x_{21} \) (FTI), \( x_{18} \) (T3), \( x_{19} \) (TT4), \( x_8 \) (Thyroid surgery), \( x_{20} \) (T4U), \( x_9 \) (I131 treatment) and \( x_{15} \) (Hypopituitary) in that order. These are shown in figure 10 as histogram.
6 Conclusion

In this work, a MGGP mathematical model was developed to provide a solution to the thyroid problem, that is the ability to diagnose a patient of the thyroid disorder (Hyperthyroidism and Hypothyroidism). Root mean square error (RMSE), $R^2$ and Adjusted $R^2$, scatter diagram and histogram were used to ascertain the goodness of the Model. Based on the statistical analysis, the model is found to be good for the thyroid disorder detection and is able to classify patient as normal (1), suffering from Hyperthyroidism (2) or Hypothyroidism (3). The developed MGGP model was able to reduce the attributes used in the classification and hence the model is cost effective compare to the traditional approach of using all the twenty one (21) attributes for diagnosing thyroid disorder. That’s the MGGP used 9 most important attributes instead of the whole 21 for the thyroid disorder detection.

Acknowledgements. We express our gratitude to the Department of Mathematics, Kwame Nkrumah University of Science and Technology for providing us resources to help complete this research successfully.

References


detection of diabetes using genetic programming, *European Signal Pro-

[8] L. Ozyimalz, T. Yildirim, Diagnosis of thyroid disease using artifi-
cial neural network methods, *In Proceedings of International Confer-
http://dx.doi.org/10.1109/iconip.2002.1199031

School of Computing Science, Newcastle University, UK, 2009.

[10] Emmett J. Jentilucci, Using the singular value decomposition, Chester F.
Carlson Center For Imaging Science, Rochester Institute of Technology,
(2003).


[12] UCI, Thyroid disease data set, 1987, Retrieved from
https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

Received: September 17, 2015; Published: November 10, 2015