Nonlinear system identification model of the spread of TB disease using the genetic algorithm and multilayer perceptron

A Damayanti\textsuperscript{1}, P Subiyanto\textsuperscript{2}, and A B Pratiwi\textsuperscript{3}

\textsuperscript{1,2,3}Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Email: \textsuperscript{1}aulid@fst.unair.ac.id

Abstract. System identification is the process of building a mathematical model of a dynamic system based on the observed input-output data. Artificial neural networks have been shown to complete the identification of nonlinear systems with various learning algorithms. Identification of nonlinear dynamic systems in the model of TB disease spread is needed to predict the spread of TB disease. The model for the spread of TB disease has three compartments and six parameters. The process of identifying models for the spread of TB disease by estimating parameters using the genetic algorithm and validation of the model using multilayer perceptron. The choice of multilayer perceptron architecture determines the minimum error in the nonlinear system identification process for the model of TB disease spread. From the results of the trial, it was shown that the multilayer perceptron and genetic algorithm were able to identify the spread model of TB disease with MMRE of 0.00347

1. Introduction

In this modern era, computers have become an indispensable need for various groups. Today, systems on computers have been able to do many things. One of them is the ability to identify nonlinear systems on various problems with certain algorithms. System identification is a way to regulate a system so that output on the system is in accordance with the target. The identification of non-linear systems is very complicated because computers are only able to solve linear calculations. Some problems can occur when predictions are made on a system that has never been trained before. These problems include prediction of the system for a less accurate time. However, in the presence of various methods, it will be able to model nonlinear systems accurately even though the exact solution of nonlinear problems is very difficult. One method that can be used to identify non-linear systems is the Artificial Neural Network method [1].

Artificial neural networks are attempts to model information processing based on the capabilities of biological nervous systems that exist in humans. So, it can be concluded that artificial neural networks are biological neural networks viewed from the point of view of information processing. So that we are able to design models that can then be simulated and analyzed [2]. Neural networks can be used to identify complicated nonlinear systems. The excellent performance of a neural network can be attributed to the nonlinear nature of the network. It has been proven that a continuous function can be arbitrarily well approximated by a feedforward network with only one hidden layer [3].

The conventional backpropagation method experiences slow local and global minimal convergence and randomly assigns initial weight values, which results in difficulties in obtaining the expected output. So that many researchers carry out hybrid processes with other methods, one of
which is Particle Swarm Optimization. According to [4], Particle Swarm Optimization is one of the best Artificial Intelligence methods for optimizing and estimating parameters. According to [5], genetic algorithm (GA) is a metaheuristic algorithm inspired by the process of natural selection and evolution. This algorithm can determine solutions by applying genetic operators to form solutions. So that by using GA, it can minimize operating costs and improve solutions [6].

In this research, the model discussed is a mathematical model of the spread of tuberculosis. Tuberculosis (TB) is an infectious disease caused by the bacterium Mycobacterium Tuberculosis, which can attack various organs, especially the lungs. If the disease is not treated or the treatment is not complete, it can cause dangerous complications to death. TB was reported to have existed in the world since 5000 BC, but progress in the discovery and control of TB has only occurred in the last two centuries [7]. In the field of mathematics, mathematical models play an important role in studying the dynamics of the spread of TB. The mathematical model is modeled in the form of a system of differential equations, in this case, most mathematical models are nonlinear in nature which are very difficult to find solutions in exact.

Therefore, in this study, the identification of the nonlinear system TB model parameter estimation process, using a genetic algorithm and multilayer perceptron. Genetic algorithms are used in the parameter estimation process, especially in the completion process of the TB model in the form of the ordinary differential equation, and optimized using multilayer perceptron. It is expected that the process of identifying nonlinear TB model systems can approach the actual model.

2. Methodology
In general, the stages to identify the nonlinear system are model constructing, model data determination, model parameter estimation, and model validation. In this research, the processes of nonlinear system identification are TB model constructing, parameter estimation, and model simulation.

2.1. System Identification
System identification is an experimental approach to modeling a process or plant from unknown parameters. In this context, modeling means building a mathematical representation of the dynamic behavior of a system or process in either the time or frequency of the domain. System identification is also known as the evaluation process of building a model at any time or domain frequency. Evaluation can be done by testing into the model approach used by entering the existing data input. From the process, the results of the model output are obtained which are then compared with the system output data so that it can be evaluated by improving certain parameters in the model used. System identification includes several stages of the process, including experimental planning, selection of model structures to be used, parameter estimation, and model validation. The steps for identifying the system will continue until the results of a satisfactory model development are obtained [8].

2.2. Nonlinear Model of the Spread of Tuberculosis (TB)
Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis and can cause death. In general, bacteria attack the lungs causing pulmonary TB disease. But in certain cases, these bacteria can attack other organs that cause extrapulmonary TB disease. An individual can be infected with TB if the bacteria is inhaled and forms colonies in the body. The spread of this disease can occur if individuals infected with TB sneeze, cough, even talk. In general, the main symptoms of TB include shortness of breath, coughing up phlegm, more than one month of fever, feeling unwell (malaise), weight loss, and cold sweat [9].
The following formulas are for SIR type of tuberculosis spreading model:

\[
\begin{align*}
\frac{dS}{dt} & = -\frac{\beta}{N} IS - \mu S + \pi \\
\frac{dI}{dt} & = \frac{\beta}{N} IS - (\mu + \mu_i + \lambda) I \\
\frac{dR}{dt} & = \lambda I - \mu R
\end{align*}
\]

where \( N = S + I + R \). \( S \) is a human population that is susceptible for TB, \( I \) is a human population infected with TB, and \( R \) is a recovered TB population. The number of populations of \( S \) will increase due to birth by \( \pi \), where \( \pi \) is constant. The population of \( S \) will decrease due to death at the rate of \( \mu \). Direct contact with infected individuals causes individuals in susceptible populations to become infected and will never into population \( I \). This causes a decrease in population \( S \). The rate of TB spreading is \( \beta \). Decreased population \( I \) was caused by death due to other factors with the rate of \( \mu \) and death due to TB at a rate of \( \mu_i \). Individuals who are infected with TB can recover spontaneously at the rate of \( \lambda \) and enter the population \( R \). This also results in reduced population \( I \). Population \( R \) can be reduced due to death by the rate of \( \mu \) [10].

2.3. Genetic Algorithm (GA)
Genetic algorithm is a metaheuristic algorithm inspired from evolution theory process. The concept begins with determining solution set or so-called population. Each individual which is the solution of any problem is stated as chromosome. So, each individual in the population is called as chromosome. Every chromosome is formed from independent structure which is called gene. Gene declares the special characteristic of an individual. Every gene has its position in chromosome which is called as locus. This chromosome will develop through iteration, respectively. The iteration is called generation [11].

According to [12], fitness function is a function which provides a value for each individual in a population. This fitness function shows the quality of each solution which represented by the individual. In this study, the stages on genetic algorithm are as followed:

1. [Start] Generating random populations of \( n \) individuals. Each individual consists of a number of parameters from the model.
2. Calculate numerical solutions for models with four order Runge-Kutta for each individual. In this study the tb model equation is ordinary differential equation. The general form of solving ordinary differential equations with the fourth order Runge-Kutta method is as follows:

\[
y_{n+1} = y_n + \frac{1}{6}(k_1 + 2 k_2 + 2 k_3 + k_4)
\]

where

\[
\begin{align*}
k_1 &= h f(t_n, y_n) \\
k_2 &= h f(t_n + \frac{h}{2}, y_n + \frac{h}{2} k_1) \\
k_3 &= h f(t_n + \frac{h}{2}, y_n + \frac{h}{2} k_2) \\
k_4 &= h f(t_n + h, y_n + h k_3)
\end{align*}
\]

3. [Fitness] Calculating fitness for each individual in population.
4. [New population] Form new individuals after passing the following genetic processes:
   a. [Selection] Choose the best individual with the fitness to be used as a parent. The stages for roulette wheel selection are as followed:
      i. Calculating fitness value for each chromosome, notated with $f_i$, $i=1, 2,…, \text{pop size}$.
      ii. Calculating total amount of fitness value of all chromosomes in population, notated by $F$.
      iii. Calculating selected probability from each chromosome ($p_i$),
          \[
          p_i = \frac{f_i}{F}.
          \]
      iv. Calculating cumulative probability of each chromosome ($q_i$),
          \[
          q_i = \sum_{j=1}^{i} p_j.
          \]
      v. Generating real number randomly with the interval of $(0, 1)$, notated as $r$. If $r \leq q_i$, then the first chromosome is selected; for the others, select the $i$-th chromosome if $q_{i-1} \leq r \leq q_i$ with $2 \leq i \leq n$.
   b. [Crossover] Parental crossover will bear new individual. This process considerate the value of crossover probability.
   c. [Mutasi] Parental mutation will bear new individual using parental characteristic. This process considerate mutation probability.
5. [Replacing] Combing dan comparing the fitness of the population generated on process 1 with the population generated on process 3. n individual with the best fitness will be maintained, the rest will be discarded.
6. [Testing] If N iteration is achieved or optimum, the process stops and the best solution is obtained. Else, back to process 2 until the best solution is achieved.

2.4. Multilayer Perceptron (MLP)
Multilayer perceptron is an artificial neural network that has one or more hidden layers between the input layer and the output layer. The architecture MLP can be seen in Figure 1. In MLP, an activation function is used for mapping the sum of input processing elements (weighted inputs) to its output. The activation function on each neuron uses a differentiable nonlinear function. The method used in MLP training is the backpropagation method. During the training there are two main phases, namely the forward phase and the backward phase. The architecture of multilayer perceptron can be seen in Figure 1.

![Figure 1. Multilayer Perceptron Architecture with n Layers](image-url)
In the feedforward, every input unit will receive input signal and spread it to every hidden unit. After that, the hidden unit will calculate its activation and deliver the signal to every output unit. Every output will then calculate the activation too to produce response towards the given input. In the training process, each of the output compares their activation with target value to determine the error. Based on this error, the factor used to distribute the error is calculated and the output will return to the previous layer. Using the following way, the factor of hidden unit will be used to update the weights connecting hidden layer to output layer [13].

In this study, the stages of the process of identifying the nonlinear system that spreads the model of TB disease using multilayer perceptron as follows:

a) Initialize parameters, namely learning rate (\( \alpha \)), target error (\( \varepsilon \)), maximum iteration (max_epoch)

b) Discretizing the model of the spread of TB using the Euler method.

c) Substituting parameter estimation solutions obtained in the GA process to discrete TB models.

d) The network architecture used consists of an input layer, a hidden layer, and an output layer. Following is the design of the number of units in each layer:

i. The number of neurons in the input layer is 2 units.

ii. The number of neurons in the hidden layer is 2 units, according to the number of neurons in the input layer.

iii. The number of neurons in the output layer is 1 unit.

e) The conversion coefficient obtained from step c) becomes the weight and bias used in the input layer to the hidden layer. While the weights and biases used in the hidden layer to the output layer are random real numbers between [-1,1].

f) Perform a feedforward process with the following steps:

i. Add the weights and biases from the input layer to the hidden layer and then use the binary sigmoid activation function to calculate the output in the hidden layer.

ii. Add the weights and biases from the hidden layer to the output layer and then use the binary sigmoid activation function to calculate the output at the output layer.

g) Backpropagation of error is a process of updating weights and bias.

h) Check the stop condition. If the MSE obtained is less than the error limit or has reached a maximum iteration, the process is complete.

i) Repeat steps g) to i) to achieve the desired stop condition.

j) After the stop condition is reached, we get the optimal weight and bias for the model validation process. The spread validation process model of TB disease uses a feedforward process in multilayer perceptron

3. The Experimental Result

Data on TB disease used in this study was obtained from Dinas Kesehatan Provinsi Jawa Timur Jl. Ahmad Yani Surabaya for 6 years starting from 2010 to 2016. Data obtained in the form of data on patients who suspect TB, TB infection, and TB recovery. The data is used in the process of estimating the parameters of TB spreading, namely the parameters of the (\( \mu \)) mortality rate, (\( \pi \)) birth rate, (\( \beta \)) TB transmission rate, (\( \mu_t \)) mortality rate due to TB virus, and (\( \lambda \)) cure rate from three of compartment of model TB. Parameter estimation of TB spreading model using the genetic algorithm. The best parameters are obtained at 500 generation with a various pm, pc and learning rate. The result of parameter estimation as shown in Table 1. The identification of nonlinear systems with estimated parameter obtained at 0.5 pm, 0.4 pc, 0.25 learning rate, and 500 generations, with an error of 0.00347 as shown in Figure 2. The result of the model simulation of the identification process can be seen in Figure 3.

| Experiment to | pm | pc | Error | Parameter of model | \( \pi \) | \( \beta \) | \( \mu \) | \( \mu_t \) | \( \lambda \) |
|---------------|----|----|-------|-------------------|--------|--------|--------|--------|--------|
| 1             | 0.4| 0.4| 0.06755|                   | 345    | 0.85653| 0.38965| 0.40562| 0.22187|
| 2             | 0.4| 0.4| 0.18999|                   | 6543   | 0.7876 | 0.1765 | 0.38981| 0.5643 |
| Experiment to | pm | pc  | Error | Parameter of model |
|---------------|----|-----|-------|-------------------|
|               |    |     |       | $\pi$ | $\beta$ | $\mu$ | $\mu_c$ | $\lambda$ |
| 3             | 0.5| 0.4 | 0.00347 | 7436 | 0.45956 | 0.40383 | 0.49049 | 0.60307 |
| 4             | 0.5| 0.5 | 0.05444 | 212  | 0.83182 | 0.29299 | 0.39286 | 0.20495 |
| 5             | 0.5| 0.5 | 0.17643 | 5436 | 0.6540  | 0.1652  | 0.3207  | 0.5210  |

Figure 2. The best result of Parameters Estimation of the Model
Figure 3. Result of identification of the Model (a). susceptible (S) population. (b). infected (I) population. (c). recovery (R) population

4. Conclusions
The process of identifying models for the spread of TB disease by estimating parameters using the genetic algorithm and multilayer perceptron. The choice of multilayer perceptron architecture determines the minimum error in the nonlinear system identification process for the model of TB disease spread. From the results of the trial, it was shown that the multilayer perceptron and genetic algorithm were able to identify the spread model of TB disease with MMRE of 0.00347.

References
[1] Soelaiman R and Rifa’i R 2009 Identifikasi Sistem Nonlinier dengan Menggunakan Recurrent Neural Network dan Algoritma Dead-Zone Kalman Filter. Industrial Electronic Seminar.
[2] Rojas R 1997 Neural Network: A Systematic Introduction, Springer-Verlag, Berlin
[3] Cybenko G 1989 Aproximations by superpositions of a sigmoidal function, Math Control Signam System 2. 303-314
[4] Neto M, Petry G G, Aranildo R L, and Ferreira T A, 2009 Combining Artificial Neural network and Particle Swarm System for Time Series Forecasting, International Joint Conference on Neural Networks, pp. 2230-2237
[5] Otbiko M 1998 Introduction to Genetic Algorithms, Czech Technical University Prague
[6] Ren C 2012 Applying Genetic Algorithm to Vehicle Routing Problem. Heilongjiang University, China
[7] Kemenkes 2011 Pedoman Pelayanan Gizi pada Pasien Tuberkulosis, Kementrian Kesehatan RI Dirjen Bina Gizi dan KIA, Jakarta
[8] Oliver Nelles 2001 Nonlinear System Identification, Springer-Verlag, Berlin
[9] World Health Organization 2017 Fact Sheets on Tuberculosis. http://www.who.int/en/news-room/fact-sheets/detail/tuberculosis
[10] Diekmann, O. and J.A.P. Heesterbeek. 2000. Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation. Simon Levis, Princeton University, USA
[11] Gen M and Cheng R 1997 Genetic Algorithms and Engineering Design, John Wiley and Sons Inc., New York
[12] Beasley D, Bull D R and Martin R R 1993 An Overview of Genetic Algorithm: Part I Fundamentals, University Computing, Cardiff
[13] Simon S Haykin (2008) Neural Networks and Learning Machines Third Edition, Pearson Education, New Jersey