Combination of SVM-GA Method based Forest Area Estimation in Kutai National Park of East Borneo

Syafaruddin *, Lapu Tombilayuk, Andani Achmad
Department of Electrical Engineering of Universitas Hasanuddin, 90245 Tamalanrea-Makassar, Indonesia
E-mail address: *syafaruddin@unhas.ac.id

Abstract

The paper presents the combination of support vector machine (SVM) and genetic algorithm (GA) methods for the forest area estimation in Kutai National park of East Borneo. The considering variables are reboization concerning natural green, forest fire, encroachment and illegal logging activities. The SVM method is superior in the data classification, while the GA method is well-known method in the estimation task. The input data that has been classified using SVM method will be the input chromosome for the GA method. As results, the computational effort and convergence are fast and the method has high accuracy estimation because only the important data is processed. The accuracy of area estimation result is compared with the actual data using mean absolute percentage error (MAPE).

Keywords: support vector machine (SVM), genetic algorithm (GA), data classification, forest estimation, MAPE.

1. Introduction

Indonesia has continuously experienced in deforestation area where the majority of forest damaging occurs in Borneo and Sumatera. The causes of deforestation are forest mismanagement, illegal logging, forest fire and forest opening for farming and mining purposes. The effects of forest damaging can be locally, such as ecologic disaster, flood, soil avalanche and can be globally as well, such as drought and other global warming effects [1]. The government is actually proactive to do the forest conservation program with policies, for instance the protection of primary natural and peat forests and the strict regulation of forest utilization for mining purpose [2]. The government regulation on forest conservation is highly protected by law; in fact, deforestation seems uncontrollable. The organization of forest watch Indonesia (FWI) reported that the forest in Java will be used up in 2020 with small width area remaining in Bali and Nusa Tenggara (0.08 million h.a), Sulawesi (7.2 million h.a), Sumatera (7.72 million h.a), Borneo (21.29 million h.a) and Papua (33.45 million h.a). The rate of deforestation is about 1.51 million h.a per year in the period of 2000-2009 with the highest rate of 0.55 million h.a per year occurs in Borneo [3].

Some group researchers focused on the method to slowing down the rate of deforestation. For instance, the group researcher from Bogor Institute of Agriculture has investigated the conservation level and wood production in forestry management unit in Bojonegoro, Java using previous 10 years data [4]. Data analysis is conducted by projection method considering the dynamic population percentage within certain area by periods. The dynamic population percentage is calculated to determine the causes of forest damaging variable and forest conservation variable with assumptions of minimum and maximum levels of each variable. Another approach method using local color segmentation algorithm based remote sensing measurement is proposed to calculate the width rate of the forest area [5]. This method follows the standard RGB color extraction and segmentation into color space of HSV, and then binary process by Freeman chain is conducted to calculate total area of object.
Similar method, the combination between the linear regression method and satellite image data is used for the forest area estimation [6]. In this method, the mean of variable regression is considered as the function of pixel value of the satellite image classification.

Several researchers have also investigated the trend conservation in certain forest site worldwide with different methods. The artificial neural network with multi-layer perceptron combined with Markov chains has been used to estimate the pressure on primary forest in Tam Dao National Park Region and secondary forest surrounding the park [7]. The method determines the forest pattern changes and suggests that the reboization is one of the important factors for the future forest conservation program. Again, the remote sensing method capability is used to estimate the Basal Area in West Oak Forests of Iran [8]. In Germany, researchers have proposed automatic estimation of forest inventory parameters based on lidar, multi-spectral and fogens data [9]. Remote sensing technology measurement is the basis data to determine the forest inventory, such as type, height, age, diameter and volume of woods.

Forest area estimation method has gained important awareness of researchers with variety methods. The statistic methods are still dominant in this topic. With the hierarchy of Bayesian model, the forest of area can be accurately classified up to 88% [10]. Another approach by the forest sampling method is used to estimate the forest canopy cove according to probability theory [11]. In addition, the forest cover estimation based the importance vegetation type has been proposed using K-means method on the time series data. The research is focused on the clustering of different type of vegetation [12]. Meanwhile, the estimation of forest canopy by comparison of the field measurement technique has been conducted by picture result of digital camera [13].

The results of the previous research motivate us to find the best method to estimate the forest area based on progress of reboization concerning natural green, forest fire, encroachment and illegal logging activities. The parametric method by mean mathematic model and statistic, such as autoregressive integrated moving average (ARIMA) is less suitable for this case study due to the difficulty in modeling irregular and variable number of data. On the other hand, the non-parametric method with exponential technique is only superior for the short-term forecasting and the results may not be confirmed optimal. In addition, the artificial neural network is facing difficulty to provide solution with time-series data. Therefore, the proposed method utilizes the combination between the support vector machine and genetic algorithm (SVM-GA) methods. The SVM-GA methods are suitable for time series data processing according to the pervious historical data information and it has been successfully implemented in many engineering problems [14].

2. Configuration of proposed system

The time series estimation method has continuously attracted serious attention from scientific community. The method is basically part of computational intelligence utilizing historical data to solve estimation problems. This kind of technique is possible supporting from the advanced computational technique and information technology. One of the computational techniques based machine learning is the Support Vector Machine (SVM). This method is superior to classify the input data set from the minimum to the maximum values. The classification results are used as chromosome for the genetic algorithm (GA) operation. This combination accelerates the computational efforts and provides high accuracy estimation compared to the only GA process. In this case, the output genetic algorithm (GA) is the optimal estimated value of forest area. It also implies that the uncorrelated data classification is avoided, as results only the important inputs are considered by the implementation of SVM-GA method.
In this research, the main configuration of the estimation method for the forest area in Kutai National park, East Borneo is divided into the input database of including positive, like reboization concerning natural green and negative causes, such as forest fire, encroachment and illegal logging activities that affecting to the forest area, processing database using GA-SVM methods and the area forest estimation output. The database system is stored using MySQL software system. The genetic algorithm utilizes these data to initialize parameter and to generate the population. Later, the support vector machine method evaluates the fitness function in order obtain the data set. Then, the data set is reselected to obtain two chromosomes with the best fitness function by the genetic algorithm. By this approach, the only correlated data is used for the GA process, results in high accuracy estimation. The process and evaluation continue with crossover and mutation of new generation until it convergences at the best intent of fitness function.

The input database is taken directly from Kutai National Park office during the last 10 years (between 2003 and 2012) about data record of reboization, forest fire, encroachment and illegal logging activities. These data is quite random and irregular due to the cause combination between the nature and human activities. In these data, the average forest damaging area based on the negative causes is about 125.68 h.a, with the mean area of reboization is about 54.5 h.a. In 2013, the total forest area of Kutai National Park, East Borneo is about 198,629 h.a with damaging area about 711.8 h.a. Based on this reason, it is important to provide some tools to estimate the forest area for some time in the future, so that it becomes reference to measure and to accelerate the green activities inside the National Park.

The forest area estimation is performed by the implementation of combination between the support vector machine (SVM) and genetic algorithm (GA) method based on the storing data in the database system. The superiority of this method is in the capability of SVM method to divide vector space into hyperplane according to the data trend from small area to wide area classes. The algorithm combination for such estimation technique is able to provide better solution compared to other estimation techniques, such as artificial neural network because the error and generalization of the SVM method is not depending on the input vector. The complete process of SVM-GA methods is shown by the pseudo-code as follows:

1. Initialization process by the genetic algorithm to select the chromosome candidates of data stored in database.
2. Initialization of data set by running the SVM method to search all data set about reboization, forest fire, encroachment and illegal logging activities between 2003 and 2012 in the data base system.
3. Run the polynomial Kernel function to classify the non-linear data into two classes by $(X^T X_i + 1)P$. The results are about the influence causes, measured from the most to the less value impacts.
4. The training process of all selected data set by means the influenced parameters to the forest area. The selection is aimed to obtain the hyperplane of $y(x)f(x)=1$ that separates 2 classes. The candidates of support vector are:
   $y(x)f(x)\leq \beta + 1$ and $y(x)f(x)\geq1$ (1)
   where $X_a = X_a \cup X_b$ and $\beta$ is the arbitrarily determined value by users.
5. If the re-training process is conducted, then the previous training results are improved by taking only some data $(X \epsilon X_a)$ as the support vector candidates.
6. The data classification after training process is divided into $x_i,w+b \geq 1$ for the 1st class and $x_i,w+b \leq 1$ for the 2nd class.
7. Chromosomes selection. The best chromosome is usually selected by objective function evaluation with defined high probability. The roulette-wheel method is used in this step.
8. Fitness function evaluation by:

\[ \text{Fitness} = C - f(x) \text{ or Fitness} = \frac{C}{f(x) + \epsilon} \]  

where C is a constant and \( \epsilon \) is small number to avoid zero division.

In this research,

The fitness function = forest fire + encroachment + illegal logging – rebozation  

(3)

9. Selection process with Linier Fitness Ranking (LFR).

10. Cross-Over process. One-cut point method is used to exchange the gen of the parent chromosome with cross-over probability \( P_c \) of 0.25.

\[
\begin{align*}
&\text{Begin} \\
&k \leftarrow 0; \\
&\text{While (}k < \text{populasi) do} \\
&R[k] \leftarrow \text{random [0 - 1];} \\
&\text{If (}R[k] < P_c) \text{ then} \\
&\text{Select Chromosome[k] as parent;} \\
&\text{End;}
\end{align*}
\]

11. Mutation process. The mutation rate is specified at 0.1.

12. Population replacement by the new generation

13. If the new generation convergences to the optimal solution, the overall process stop; otherwise the process is repeated from no.9.

The above pseudocode is translated into language programming of Borland Delphi 7. The mean absolute percentage error (MAPE) is used to validate the accuracy in output measurement. In addition, the white box testing is used to evaluate all computer logic programming by checking the logical iteration and to assess the overall data used in the simulation.

3. Simulation Results

The simulation results are explained as follows. In the inputting data process, the users may input data of year, region, and causes both positive and negative. The year data is specified from 2003 to 2012 with three definite regions where region I is called Suka Rahmat, region II is Sangatta and region III is Manamang. The last required data are the causes and their affected total area. The causes influenced to the forest area in the park are reboization, forest fire, encroachment and illegal logging activities.

<table>
<thead>
<tr>
<th>Year</th>
<th>Forest fire (h.a)</th>
<th>Encroachment (h.a)</th>
<th>Illegal logging (h.a)</th>
<th>Reboization (h.a)</th>
<th>Forest Area (h.a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>9.5</td>
<td>145.0</td>
<td>10.0</td>
<td>64.0</td>
<td>100.5</td>
</tr>
<tr>
<td>2004</td>
<td>33.0</td>
<td>95.0</td>
<td>6.0</td>
<td>50.0</td>
<td>184.5</td>
</tr>
<tr>
<td>2005</td>
<td>1.0</td>
<td>66.0</td>
<td>2.8</td>
<td>27.0</td>
<td>227.3</td>
</tr>
<tr>
<td>2006</td>
<td>1.0</td>
<td>50.0</td>
<td>5.0</td>
<td>68.0</td>
<td>215.3</td>
</tr>
<tr>
<td>2007</td>
<td>3.5</td>
<td>140.0</td>
<td>15.0</td>
<td>45.0</td>
<td>328.8</td>
</tr>
<tr>
<td>2008</td>
<td>5.0</td>
<td>77.2</td>
<td>4.5</td>
<td>65.0</td>
<td>350.5</td>
</tr>
<tr>
<td>2009</td>
<td>4.0</td>
<td>139.0</td>
<td>21.0</td>
<td>60.0</td>
<td>454.5</td>
</tr>
<tr>
<td>2010</td>
<td>1.5</td>
<td>111.0</td>
<td>11.0</td>
<td>59.0</td>
<td>51.9</td>
</tr>
<tr>
<td>2011</td>
<td>16.8</td>
<td>139.0</td>
<td>5.0</td>
<td>30.0</td>
<td>649.8</td>
</tr>
<tr>
<td>2012</td>
<td>23.5</td>
<td>90.0</td>
<td>25.5</td>
<td>77.0</td>
<td>711.8</td>
</tr>
</tbody>
</table>

After the input data process is complete, the user may check their data stored in the database. The typical initial data stored in database system is shown in Table 1. It is clearly shown that there are certain damaged forest area affected, forest fire, encroachment and illegal logging activities even though the reboization contributes positively in this matter. Total forest area in Kutai National Park, East Borneo is 198,629 h.a. However, it has been
recently found that the total forest area reduced to 197,917 h.a by considering the above causes. In the current display, the users may continue the process until they obtain the estimated area in future years.

If the process continues, then the users may receive information about the estimation of forest area in the following years. In this simulation, the maximum width of forest area (the chromosome value) of the generation is the estimation output. It is due the consideration of the 10 years previous data to predict the forest area in the year after. For example in Table 2, the chromosome value change from the initial value (data from 2003 to 2012) to other number specified that the damaged forest area in 2013 is 784.7 h.a. With the same consideration in the following generation, it will be the estimation results of the year 2014, and so on. With this application system, the estimated damaged forest area in the years of 2014, 2015 and 2016 are 905.1 h.a, 1,444.7 h.a and 2,211.2 h.a, respectively. These results are obtained with assumptions that there is no updating data from the initial data condition. The estimated results in 2016 may less than the above estimated number if the reboization activity increases and one of the negative causes can be pressed down.

Table 2: Typical data of damage forest area estimation in 2013 (first generation)

<table>
<thead>
<tr>
<th>Forest fire (h.a)</th>
<th>Encroachment (h.a)</th>
<th>Illegal logging (h.a)</th>
<th>Reboization (h.a)</th>
<th>Forest Area (h.a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>78</td>
<td>90</td>
<td>10</td>
<td>64</td>
<td>114</td>
</tr>
<tr>
<td>33</td>
<td>95</td>
<td>6</td>
<td>50</td>
<td>198</td>
</tr>
<tr>
<td>23.5</td>
<td>90</td>
<td>25.5</td>
<td>77</td>
<td>260</td>
</tr>
<tr>
<td>25</td>
<td>140</td>
<td>15</td>
<td>45</td>
<td>395</td>
</tr>
<tr>
<td>3.5</td>
<td>140</td>
<td>15</td>
<td>45</td>
<td>508.5</td>
</tr>
<tr>
<td>16.8</td>
<td>139</td>
<td>5</td>
<td>30</td>
<td>639.3</td>
</tr>
<tr>
<td>5.1</td>
<td>77.2</td>
<td>4.5</td>
<td>65</td>
<td>661</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>51</td>
<td>68</td>
<td>695</td>
</tr>
<tr>
<td>5</td>
<td>77.2</td>
<td>4.5</td>
<td>65</td>
<td>716.7</td>
</tr>
<tr>
<td>4</td>
<td>139</td>
<td>21</td>
<td>96</td>
<td>784.7</td>
</tr>
</tbody>
</table>

Fig. 1 Typical display of estimation results with graph estimation

The design of application program is complete with the display of graph estimation as shown in Fig. 1. In this figure, it is shown that in certain period the damage forest area decreases. For instance, there is significant reduction of forest area to about 939 h.a in 2017 and 358.4 h.a in 2020 from the previous year conditions. However, this condition is some kind of transition states because the total damaged area rises again to 1,883.9 h.a and 2,341.3 h.a in the years of 2018 and 2019, respectively. As previously mentioned that the current
forest area of Kutai National Park, East Borneo is 198,629 h.a. If the prediction process continues without any changes in the input data by means there is no positive action for the green activity, the forest area is used up in 2061 because the estimated damage area in this year has reached 200,129 h.a which is far beyond the current forest area. Obtaining such number is important for the local people, park authority and global society as the reference to do positive action for the forest conservation since the Kutai National Park of East Borneo is ‘the lung of world’.

4. Discussions

In the design of system application, the high accuracy estimation or prediction is the most important aspect to be considered in order to guarantee the outcomes are on the right value even the data input collection is abruptly changed. The accuracy assessment in this research is by measuring the Mean Absolute Percentage Error (MAPE) between the estimated area as the output of the designed system application and the actual area. In addition, the estimated output area is also compared with the conventional linear regression with similar MAPE performance index measurement.

![Fig. 2 Simulation result for the estimated and actual area comparison](image)

<table>
<thead>
<tr>
<th>Year</th>
<th>Estimated area (h.a)</th>
<th>Actual area (h.a)</th>
<th>Difference (h.a)</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>453.3</td>
<td>350.5</td>
<td>102.8</td>
<td>5.9</td>
</tr>
<tr>
<td>2009</td>
<td>503.3</td>
<td>454.5</td>
<td>48.8</td>
<td>2.1</td>
</tr>
<tr>
<td>2010</td>
<td>530.0</td>
<td>519.0</td>
<td>11.0</td>
<td>0.4</td>
</tr>
<tr>
<td>2011</td>
<td>641.0</td>
<td>649.8</td>
<td>8.8</td>
<td>0.3</td>
</tr>
<tr>
<td>2012</td>
<td>641.0</td>
<td>711.8</td>
<td>70.8</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The first scenario is the comparison between the estimated area and the actual area. For the easy comparison, the data of previous five years from 2003 to 2007 is arbitrarily selected as the initial input data (training data) because we have the figure actual data from 2008 to 2012 for the validation data. The result of simulation under such scenario is shown in Fig. 2. In the simulation results, we have comparing data between estimated area and actual area from 2008 to 2012. These data is evaluated by Mean Absolute Percentage Error (MAPE) equation as follows:

$$\text{MAPE} = \frac{\sum_{t=1}^{N} \left| \frac{A_t - E_t}{A_t} \right|}{N} \times 100\%$$

where $A_t$ is the actual area in the t year, $E_t$ is the estimate area in the t year and N is the period of data evaluation (N=5) in this research. The MAPE index performance calculation is summarized in Table 3. For the five years period of estimation, the average of MAPE is
about 2.1% with the maximum difference of 102.8 h.a in 2008 and minimum difference of 8.8 h.a in 2011. It means that the proposed system application to estimate the forest area of Kutai National Park, East Borneo is guarantee small. Therefore, the estimated forest area until the year of 2020 may present proper results.

5. Conclusion

The paper presents the design computer system application to estimate the forest area in Kutai National Park, East Borneo based the combination of support vector machine (SVM) and genetic algorithm (GA). The estimation process considering variables such as reboization concerning natural green, forest fire, encroachment and illegal logging as the input data in database system. Without any actions regarding to the initial data by means there is no significant effort to do the forest conservation program by the park authority, the forest area will be finished by the year of 2061. It is very dangerous situation to the world ecosystem because this park is ‘the lung of the world’. Provision initial data information is very urgently to save our environment. The accuracy of area estimation result is compared with the actual data using mean absolute percentage error (MAPE) with the average error of 2.1%.

References