Bagging Approach for Increasing Classification Accuracy of CART on Family Participation Prediction in Implementation of Elderly Family Development Program

Wisoedhanie Widi Anugrahanti*, Arief Wibowo**, Soenarnatalina Meilanani***
*Faculty of Public Health, Airlangga University, Indonesia
Email: wisoedhanie.widi@gmail.com

ABSTRACT

Classification and Regression Tree (CART) was a method of Machine Learning where data exploration was done by decision tree technique. CART was a classification technique with binary recursive reconciliation algorithms where the sorting was performed on a group of data collected in a space called a node/ node into two child nodes (Lewis, 2000). The aim of this study was to predict family participation in Elderly Family Development program based on family behavior in providing physical, mental, social care for the elderly. Family involvement accuracy using Bagging CART method was calculated based on 1-APER value, sensitivity, specificity, and G-Means. Based on CART method, classification accuracy was obtained 97.41% with Apparent Error Rate value 2.59%. The most important determinant of family behavior as a sorter was society participation (100.00000), medical examination (98.95988), providing nutritious food (68.60476), establishing communication (67.19877) and worship (57.36587). To improve the stability and accuracy of CART prediction, used CART Bootstrap Aggregating (Bagging) with 100% accuracy result. Bagging CART classifies a total of 590 families (84.77%) were appropriately classified into implement elderly Family Development program class.

Keywords: Bagging Classification and Regression Tree, Classification Accuracy, Family Participation

INTRODUCTION

Background

Classification was one of the statistical methods used to group or classify a systematically organized data. In many cases classification could be assumed to be the number of categories or populations of an existing individual and each population was characterized by the size of its probability distribution (Anderson, 1984).

Classification And Regression Tree was a method of Machine Learning where data exploration method was done by decision tree technique. CART was a classification technique with binary recursive reconciliation algorithms where the sorting was performed on a group of data collected in a space called a node/ node into two child nodes (Lewis, 2000). CART would produce a classification tree if the response variable was categorical data, whereas if the response variable was a continuous data would be generated the regression tree (Breiman, et al., 1993). However, the results of classification trees tend to be unstable, because small changes in the learning data will affect the results of prediction accuracy. To improve the stability and predictability strength of tree classification could be used Bootstrap Aggregating (Bagging) method Classification And Regression Tree (Bagging CART).

Bootstrap Aggregating (Bagging) was one of the ensemble techniques introduced by Breiman that was used in several methods of classification and regression to reduce the variance of a predictor to improve the quality of prediction. Bootstrap was a resampling or retrieval of mutually free and repeated sample data used to predict the error rate of the loop (Breiman, 1993).

The projected increase in the average life expectancy of the Indonesian population by 71.7% in 2015-2020 provided the consequences of an increase in the old dependency ratio. The care and participation of the family in the care of physical, mental and social health was needed to realize the elderly devoted, independent, productive and beneficial to the family and society, which was the goal of the National Family Planning Coordinating Program in the activities of Elderly Family Development.
Purpose

Classifying participation in Elderly Family Development program based on family behavior in performing physical, mental and social care of elderly used Classification And Regression Tree (CART).

METHODS

This was a non-reactive study, which is a measurement which individuals surveyed did not realize that they are part of a study. This study used secondary data from the National Population and Family Planning Program Performance Indicator Survey 2015 that was about the treatment of families who had elderly in maintaining physical health, mental and social elderly in East Java 2015.

Population in this study was family which had elderly which amount 727 family. Response variable was family participation, with category 1 = implementing, 0 = not implementing. Predictor variable consists of 16 variables, namely family behavior in: provide nutritious food, exercise, maintain personal hygiene, keep the environment clean, medical examination, worship, taking part in the family, keep the feelings, give attention, establish communication, understand the needs, advise, get together with friends, society participation, participate in the economy, courses. Data were analyzed using Salford Predictive Modeler (R).

RESULTS

Classification and Regression Trees (CART) was a classification method that used decision tree algorithms. Response variable used in this research is categorical, then the resulting tree was called classification tree. The formation of tree classification in this study as follows:

1. Sorting node

A split s would be used to select the vertex t into two vertices that was the left node (t_L) and the right node (t_R) by maximizing the value of \( \varphi (s, t) = i(t) - P_L i(t_L) - P_R i(t_R) \). The result of goodness of split of this study could be seen as follows:

<table>
<thead>
<tr>
<th>No</th>
<th>Variable</th>
<th>Goodness of split/ Improvement</th>
<th>N Left</th>
<th>N Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Society participation</td>
<td>0.1659142</td>
<td>294</td>
<td>402</td>
</tr>
<tr>
<td>2</td>
<td>Get together with friends</td>
<td>0.1551249</td>
<td>381</td>
<td>312</td>
</tr>
<tr>
<td>3</td>
<td>Worship</td>
<td>0.1300618</td>
<td>400</td>
<td>296</td>
</tr>
<tr>
<td>4</td>
<td>Medical examination</td>
<td>0.1224152</td>
<td>245</td>
<td>451</td>
</tr>
<tr>
<td>5</td>
<td>Establish communication</td>
<td>0.1014271</td>
<td>199</td>
<td>497</td>
</tr>
<tr>
<td>6</td>
<td>Give attention</td>
<td>0.0882353</td>
<td>177</td>
<td>519</td>
</tr>
<tr>
<td>7</td>
<td>Exercise</td>
<td>0.0824437</td>
<td>181</td>
<td>518</td>
</tr>
<tr>
<td>8</td>
<td>Maintain personal hygiene</td>
<td>0.0684008</td>
<td>142</td>
<td>554</td>
</tr>
<tr>
<td>9</td>
<td>Provide nutritious food</td>
<td>0.0620245</td>
<td>544</td>
<td>152</td>
</tr>
<tr>
<td>10</td>
<td>Keep the feelings</td>
<td>0.0608365</td>
<td>128</td>
<td>568</td>
</tr>
<tr>
<td>11</td>
<td>Keep the environment clean</td>
<td>0.0412844</td>
<td>90</td>
<td>606</td>
</tr>
<tr>
<td>12</td>
<td>Understand the needs</td>
<td>0.0344203</td>
<td>76</td>
<td>620</td>
</tr>
<tr>
<td>13</td>
<td>Role</td>
<td>0.0342054</td>
<td>104</td>
<td>592</td>
</tr>
<tr>
<td>14</td>
<td>Participate in the economy</td>
<td>0.0258467</td>
<td>58</td>
<td>638</td>
</tr>
<tr>
<td>15</td>
<td>Advise</td>
<td>0.0117086</td>
<td>27</td>
<td>669</td>
</tr>
<tr>
<td>16</td>
<td>Course</td>
<td>0.0008489</td>
<td>2</td>
<td>694</td>
</tr>
</tbody>
</table>

The best divider for node 0 or root node was the community participation variable, with the "Yes" sorting criteria on the left node (node 1) and "No" on the right node (node 2). The variable was chosen because it had the highest goodness of split / improvement value compared to other variables.

2. Class labeling

The process of labeling on the formed node was done based on the rule of the largest number of class members that was \( p(j_0 | t) = \max_j p(j | t) = \max_j \frac{N_j(t)}{N(t)} \).
Example calculation of labeling on the formation of tree classification could be seen in Figure 1 as follows:

```
Figure 1. Class labelling

OLAH_RAGA=exercise; BERKUM_TEM=Get together with friends;
MKN_BERGIZI=provide nutritious food

Figure 1 showed that, if we would give the class label or node 11, then \( j_0 = j \) with \( j = 1 \) (exercise) and \( j = 0 \) (not exercising). The labeling process at node 11 was done as follows:

\[
P (\text{exercise}) = \frac{8}{87} = 0.091
\]
\[
P (\text{not exercising}) = \frac{79}{87} = 0.908
\]

Based on these calculations, then node 11 was given a non-exercise class label because the proportion of the class was not exercised was greater than the proportion of the exercise class.

3. Pruning Termination

The maximal classification tree had 14 internal nodes and 17 terminal nodes. The process of termination of sorting is done as in Figure 2 which is at node 2. At node 2 there were 169 data which was homogeneous, so pruning process was stopped.

```

MKN_BERGIZI=provide nutritious food

Figure 2. Pruning node termination
4. Pruning classification tree

The maximal classification pruning process begun by taking $t_L$ which was the left node and $t_R$ which was the right node of $T_{MAX}$ generated from the parent node $t$. If two child nodes and a parent node that satisfied the equation $R(t) = R(t_L) + R(t_R)$, then the child nodes $t_L$ and $t_R$ were trimmed. The maximal classification pruning process that had been done as in node 13 (Figure 3) as follows:

$$r(node\ 13) = 1 - \max P(j|node\ 13) = 1 - 0.833 = 0.167$$

$$P(node\ 13) = \frac{12}{696} = 0.017$$

$$R(node\ 13) = r(node\ 13) \times P(node\ 13) = 0.167 \times 0.017 = 0.002839$$

Then calculated value $R(t_L)$ and $R(t_R)$ in child node that was parent node 12 and parent node 13. At parent node 12 obtained:

$$r(parent\ node\ 12) = 1 - \max P(j|parent\ node\ 12) = 1 - 1 = 0$$

$$P(parent\ node\ 12) = \frac{2}{696} = 0.002$$

$$R(parent\ node\ 12) = r(parent\ node\ 12) \times P(parent\ node\ 12) = 0 \times 0.002 = 0$$

At parent node 13 obtained:

$$r(parent\ node\ 13) = 1 - \max P(j|simpuler\ 13) = 1 - 1 = 0$$

$$P(parent\ node\ 13) = \frac{10}{696} = 0.014$$

$$R(parent\ node\ 13) = r(parent\ node\ 13) \times P(parent\ node\ 13) = 0 \times 0.014 = 0$$

Based on calculations $R(parent\ node\ 12) + R(parent\ node\ 13) = 0 + 0 = 0$

The result was the same as the result of node 12 calculation was 0, so it could be done pruning at terminal node 12 and terminal node 13.

The Optimal CART produces 11 node divided into 4 non-executing class nodes and 7 class nodes executing, with sequential classification variables according to the value of the important variables as follows:
PARTIS_MASY=Society participation; IBADAH=Worship; MKN_BERGIZI=Provide nutritious food; BERKUM_TEM=Get together with friends; OLAH_RAGA=Exercise; PEM_KESH=Medical examination.

Figure. 4 Determinant variable classification

Evaluation of classification algorithmic performance could be measured through a confusion matrix. The confusion matrix contains information about the actual and predicted classes presented in Table 2 below:

Table. 2 Conversion Classification Matrix of CART

<table>
<thead>
<tr>
<th>Implementation of Elderly Family Development</th>
<th>Observed Class</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Class</td>
<td>Not Implementing</td>
<td>106</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Implement</td>
<td>18</td>
<td>572</td>
<td></td>
</tr>
</tbody>
</table>

\[ 1 - APER = \frac{TP + TN}{TP + FP + TN + FN} = \frac{572 + 106}{572 + 18 + 106 + 0} = 0.9741 \]

TP=True positive; TN=True negative; FP=False positive; FN=false negative

The result of classification using CART method shows the accuracy value of 97.41%. The application of Bagging (Bootstrap Aggregating) method on CART produces 100% accuracy, could be seen in Table. 3 Configuration Matrix on Bagging CART

Table. 3 Configuration Classification Matrix of CART Bagging

<table>
<thead>
<tr>
<th>Implementation of Elderly Family Development</th>
<th>Observed Class</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Class</td>
<td>Not Implementing</td>
<td>106</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Implement</td>
<td>0</td>
<td>590</td>
<td></td>
</tr>
</tbody>
</table>

\[ 1 - APER = \frac{TP + TN}{TP + FP + TN + FN} = \frac{590 + 106}{590 + 0 + 106 + 0} = 1.0 \]

TP=True positive; TN=True negative; FP=False positive; FN=false negative

DISCUSSION
The calculation of 1-APER in classification using Classification and Regression Trees (CART) method yields a value with an accuracy of 97.41%. The application of the Bagging method on CART gives an increase of accuracy of 2.59% expressed with 1-APER value on Bagging CART which is increased by 100%. Could be seen in table 3 The CART Confusion matrix results indicated that there were still incorrect data classified that was 18 families were classified in False Positive (FP), whereas in table 3 Bagging on CART confusion matrix there was no misclassification data.

Bagging implementation could increase the value of maximum accuracy because Bagging was able to overcome the instability of a single classification tree. Bagging could reduce the standard errors generated by a single tree by doing the average, so the assumptions will shrink, and the degree of alleged bias was unaffected (Hastie et al., 2008; Breiman, 1996; Berk, 2008).

CONCLUSION

There are six determinant variables of family participation in the implementation of Elderly Community Development program that was society participation, medical examination, providing nutritious food, worship, exercising, get together with friends (54,90971). The result of classification accuracy using Bagging on CART method of family participation in the implementation of Elderly Family Development program in East Java Province could increase accuracy by 2.59%.

REFERENCES