jtm

by Arn Jtm Arn Jtm

Submission date: 16-Nov-2023 08:21AM (UTC+0700)

Submission ID: 2229571601

File name: TURN.pdf (261.51K)

Word count: 2539

Character count: 13210

CART Classification on Ordinal Scale Data With Unbalanced Proportions using Ensemble Bagging Approach

A. INTRODUCTION

CART is one of the algorithms in data exploration techniques with decision tree techniques. CART was developed to classify nominal, ordinal, and continuous response variables. CART can also select the variables that are most important in determining the results (Breiman, 1996; Siahaan et al., 2017). The main problem that often becomes a challenge in classification analysis is unbalanced class proportions. Unbalanced class proportions is a condition where there is an unbalanced proportion between classes in the data. Unbalanced class proportions can be defined as a condition in a data set where there is a large class while other classes are only represented by a few objects (Sun et al., 2007). Unbalanced class proportions in the classification process can cause the classification results on minor data to be covered by the prediction of major data or in other words the classification results of minor data to be incorrect (Fitriani et al., 2021).

One way to overcome the problem of data imbalance is to use an ensemble algorithm. An ensemble approach is an algorithm that combines various predictions into one final prediction. One of the most commonly used ensemble methods is ensemble bagging. Bagging is an ensemble method for training data on a subset of random samples from the original dataset (Ngo et al., 2022). This subset is generated through a bootstrap resampling process that involves random sampling with returns from the original data. This research is applied to under-five nutritional status data on Height/Age assessment with three categories namely stunting, normal, and high.

The expected proportion of stunted toddlers is much smaller than the proportion of toddlers with normal or high height. Therefore, there will be a class imbalance in the data with stunting cases. Stunting is a condition of child growth failure due to malnutrition in the first thousand days of a child's life since being in the mother's womb. Stunting can have immediate and long-term impacts that result in decreased productivity, decreased intellectual ability, increased risk of infection and infectious diseases in adulthood, and even death. The government is trying to carry out a stunting prevention program by appointing areas to become pioneers of the stunting prevention acceleration program. One of the areas that became a pioneer of the accelerated stunting prevention program is Malang District. Bappeda Malang District designated 32 villages as priority villages for accelerating stunting prevention in 2021, one of which is Sumberputih Village, Wajak District. Thus, this study develops a CART classification method on ordinal-scale data with unbalanced proportions through an ensemble bagging approach in the case of nutritional status of toddlers in Sumberputih Village.

B. METHODS

1. Classification and Regression Trees (CART)

CART is a machine learning that is used to perform classification analysis for categorical and continuous response variables. The results of CART itself depend on the scale of the response variable. If the response variable is continuous, the resulting tree model is regression trees, while if the response variable is categorical, the resulting tree model is classification trees (Breiman et al., 1984). The purpose of CART itself is to get an accurate group of data to characterize a classifier (Bramer, 2016). There are three stages of the classification tree formation process with CART, namely:

a. Node Breaking

The selected variables and threshold values are selected based on criteria that maximize data cleaning or reduce impurity in each resulting group. The selected variables and threshold values are chosen based on criteria that maximize data cleaning or reduce impurity in each resulting group. In this study, the Gini Impurity Index criterion was used to decide which variables to use as separators.. Gini Impurity is calculated by equation (1) (Daniya et al., 2020).

$$i(t) = 1 - \sum_{j=1}^{n} p^2(j|t) \qquad (1)$$

where $i(t)$ is the gini impurity at node y and $p(j|t)$ is the probability of class j at node t.

b. Class Labeling

Class labeling is the process of identifying vertices to determine the dominant class of a vertex. Class labeling is done to find out the characteristics of each vertex formed. The largest class probability indicates that the class dominates the node. The calculation of the dominant class probability is presented in equation (2).

$$p(j_0|t) = \max_{\mathrm{j}} p(j|t) = \max_{j} \frac{N_j(t)}{N(t)} \qquad (21)$$

with

$p(j_0|t)$: probability of class j_0 at node t (dominant class probability)

$p(j|t)$: probability of class j at node t

$N_j\,(t)$: the number of observations of class j at node t

$N(t)$: the number of observations at vertex t

c. Pruning

Tree pruning is done to prevent large trees from forming. Large classification trees cause high complexity. After pruning is done, an optimal classification tree will be formed. The pruning method performed is Minimal Cost-Complexity Pruning in equation (3).

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \qquad (3)$$

Where the value of R(T) is shown in equation (4).

$$R(T) = \sum_{t\in\tilde{T}} r(t)p(t) = \sum_{t\in\tilde{T}} R(t) \qquad (4)$$

with

$R_\alpha(T)$: resubstitution of tree T at complexity α or cost-complexity pruning value

$R(T)$: resubstitution estimate

α : complexity parameter

$|\tilde{T}|$: the number of terminal or leaf nodes in tree T

$r(t)$: probability of making a wrong classification at node t

$p(t)$: probability of node t

2. Ensemble Bagging

Ensemble is a machine learning algorithm where several weak models are trained to solve a problem and combined to get better results (Cendani & Wibowo, 2022). Ensemble approach combines various predictions from each iteration into one final prediction (Siringoringo & Jaya, 2018). Ensemble techniques are able to provide predictions with very good accuracy (Efendi et al., 2020). The main idea of ensemble is to combine several sets of models that solve the same problem to get a more accurate model. (Friedman et al., 2000).

Bagging is an ensemble used to improve classification stability. Bagging uses based-models by performing parallel and independent learning on each based-model which is then combined to obtain the best results. The bagging process is depicted in Figure 1 which is a redrawing of Cendani & Wibowo (2022). This method is used as a tool to improve stability and predictive power by reducing the variance of a predictor of classification and regression methods whose use is not limited to improving estimators. One of the ensemble bagging methods is bootsrap aggregating (bagging). Bagging works by combining models trained using randomly generated data using bootstrap resampling.

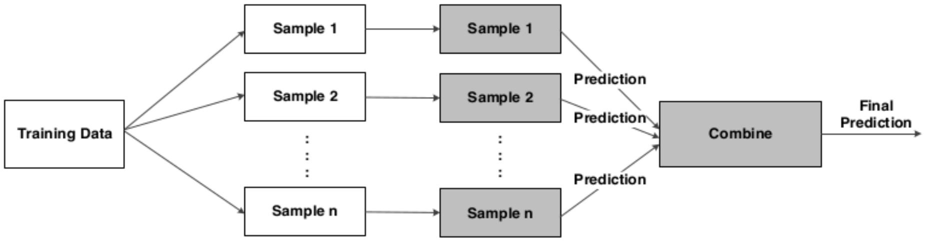

Figure 1. Bagging Ensemble Learning

Resampling is used as a tool to improve predictive consistency by reducing the variance of a predictor in classification. The basic idea of bagging with bootstrap resampling is to generate multiple versions of predictors which, when combined, produce better results for solving the same problem (Breiman, 1996). The idea behind ensemble bagging is to reduce the variance of the predictive model by training several similar predictive models on different subsets of data obtained from resampling (De Prado, 2018; du Plooy & Venter, 2021). The main steps in the ensemble bagging algorithm are as follows.

a. Resampling with returns (boostrap)

Bootstrap sampling of $\mathcal{L}_B$ as many as n from dataset $\mathcal{L}$

b. Independent model training

Each subset generated from the resampling process is used to train an independent predictive model. In this study, the bagging algorithm will be applied in logistic regression and CART analysis.

c. Aggregate prediction

After all models are trained, the final prediction is done by combining the prediction results of each model with majority voting for classification or average for regression where each model has the same weight. The majority voting used is argmax in equation (5).

$$\hat{y}_B = \text{argmax}_j f(\boldsymbol{x}, \boldsymbol{\mathcal{L}_B}) \tag{5}$$

3. Performance

Classification performance is measured based on three criteria, including: accuracy, sensitivity, model specificity, and F1-Score. Accuracy measures how correctly a diagnostic test identifies and excludes certain conditions. In other words, accuracy is used to measure the goodness of the model. In diagnostic tests, the terms sensitivity and specificity are also known. Sensitivity and specificity in diagnostic tests are measures of the ability to identify objects precisely according to reality (Wong & Lim, 2011). Sensitivity is the percentage of answers given by the system that can be classified from information on all requested data, while specificity is the success rate of the system in recovering information data that can be classified correctly. In addition, the calculation of the F1-Score value used gives an idea of how well the model identifies the positive class correctly without giving many positive errors or negative errors.

For example, in the classification of stunting toddlers, there are three categories, namely 1) Stunting, 2) Normal, and 3) high. The most common way to show classification results is by presenting them in the form of a confusion matrix to get accuracy, sensitivity, specificity, and F1-Score values as in Table 1.

Table 1. Confusion Matrix

Actual	**Predict**		
	Stunting	**Normal**	**High**
Stunting	a	b	c
Normal	d	e	f
High	g	h	i

Classification accuracy is calculated through accuracy using formula (6), sensitivity is calculated using formula (7), specificity is calculated using formula (8), and F1-Score is calculated using formula (9).

$$Accuracy = \frac{\text{a+e+i}}{\text{a+b+c+d+e+f+g+h+i}} \tag{6}$$

$$Sensitivity = \frac{\frac{a}{\text{a+(d+g)}} + \frac{e}{\text{e+(b+h)}} + \frac{i}{\text{i+(c+f)}}}{3} \tag{7}$$

$$Specificity = \frac{\frac{a}{\text{a+(b+c)}} + \frac{e}{\text{e+(d+f)}} + \frac{i}{\text{i+(g+h)}}}{3} \tag{8}$$

$$F1 - Score = 2\left(\frac{Sensitivity \times Specificity}{Sensitivity + Specificity}\right) \tag{9}$$

with

a : number of observations from group 1 that are correctly classified to group 1
b : number of observations from group 1 that are classified to group 2
c : number of observations from group 1 that are classified to group 3
d : number of observations from group 2 that are classified to group 1
e : number of observations from group 2 that are correctly classified to group 2
f : number of observations from group 2 that are classified to group 3

g : number of observations from group 3 that are classified to group 1
h : number of observations from group 3 that are classified to group 2
f : number of observations from group 3 and correctly classified to group 3

4. Research Data

The data used is secondary data from the research of Fernandes & Solimun (2023) which examines the factors that cause stunting in Wajak District. The sample in Fernandes & Solimun's research was mothers who had toddlers in Sumberputih Village. Sampling was carried out using stratified random sampling technique with a sample obtained of 100 respondents, all of which were used as samples in this study. Data on economic conditions (X_1), health services (X_2), children's diet (X_3), and environment (X_4) are predictor variables in the form of community perceptions that are assessed with Likert-scale indicators. The Toddler Nutritional Status variable (Y) is an ordinal scale response variable with categories 1 (stunting), 2 (normal), and 3 (high).

C. RESULT AND DISCUSSION

1. CART Classification Results

The results of classification of ordinal-scale data with CART are presented in Table 2.

Table 2. Confusion Matrix CART

Actual	Predict		
	Stunting	**Normal**	**High**
Stunting	1	3	0
Normal	6	7	0
High	1	0	1

Based on Table 2, the results show that CART is able to classify the nutritional status of 9 out of a total of 20 toddlers correctly. Based on the results in Table 2, the accuracy value is 45%, sensitivity is 44.3%, specificity is 41.7%, and F1-Score is 42.9%.

2. Bagging CART Classification Results

The results of classification of ordinal-scale data with Bagging CART are presented in Table 3.

Table 3. Confusion Bagging CART

Actual	Predict		
	Stunting	**Normal**	**High**
Stunting	2	2	0
Normal	0	14	0
High	0	1	1

Based on Table 3, the results show that CART is able to classify the nutritional status of 17 out of a total of 20 toddlers correctly. Based on the results in Table 3, the accuracy value is 85%, sensitivity is 94.1%, specificity is 66.7%, and F1-Score is 78%.

3. Performance of Classification

Based on the results in table 2 and table 3, it can be concluded that the bagging CART method is better for classifying data with proportion imbalance problems. This is because the performance value of bagging CART is much greater than the conventional CART method. the accuracy value is 85% which means that the Bagging CART method is able to classify cases of nutritional status of toddlers correctly by 85%. Sensitivity is 94.1% that shows that the Bagging CART method can correctly classify the nutritional status category of positive toddlers by 94.1%. Specificity is 66.7% This shows that the Bagging CART method can correctly classify the nutritional status category of negative toddlers by 66.7%. F1-Score is 78% whic mean bagging CART has balance measurement between a model's ability to correctly identify positive cases (sensitivity) and its ability to avoid classifying negative cases as positive cases (specifity) sebesar 78%. This is in line with the research of Kumari et al. (2021) which states that the ensemble method is better at classifying data with unbalanced proportions compared to conventional classification methods. Ensemble methods are better than traditional classification methods for several reasons.

a. Improved accuracy:

Ensemble methods combine multiple models to produce a more accurate final model. The final model is often more accurate than any of the individual models used in the ensemble

b. Reduced overfitting

Ensemble methods reduce overfitting by combining multiple models that have been trained on different subsets of the data. This reduces the variance of the final model and makes it less likely to overfit the training data

c. Robustness

Ensemble methods are more robust than traditional methods because they are less sensitive to noise and outliers in the data. This is because the final model is based on the consensus of multiple models, rather than a single model

d. Flexibility

Ensemble methods can be used with any type of model, including decision trees, neural networks, and support vector machines

e. Interpretability

Ensemble methods can be more interpretable than traditional methods because they combine multiple models, each of which may have a different interpretation. This can help to explain the final model and make it more understandable to humans

D. CONCLUSION AND SUGGESTIONS

Classification performance is based on accuracy, sensitivity, specificity, and F1-Score values calculated from the three-category confusion matrix. bagging CART is better at classifying data with unbalanced proportions compared to ordinary CART. This is because the performance value produced by Bagging CART is the highest with accuracy, sensitivity, specificity, and F1-Score values of 85%, 94.1%, 66.7%, and 78%, respectively. In future research, it is expected that simulation studies can be carried out with various unbalanced proportions and different sample sizes.

jtm

ORIGINALITY REPORT

15%	8%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.jatit.org Internet Source	2%
2	Evi Ramadhani, Shafira Mustaqima, Bagus Sartono, Alfian F. Hadi, Winny D Safitri, Raudhatul M. Yusuf. "An analysis of the relationship between social protection program status and the incidence of food insecure households in Aceh province", AIP Publishing, 2022 Publication	2%
3	Koehler, G.J.. "Improper linear discriminant classifiers", European Journal of Operational Research, 19910125 Publication	1%
4	aman.ai Internet Source	1%
5	journal.uinjkt.ac.id Internet Source	1%
6	Submitted to University of Ulster Student Paper	1%

7	Submitted to Brunel University Student Paper	1 %
8	Submitted to University of Melbourne Student Paper	1 %
9	Helen Abbey. "Statistical procedure in developmental studies on species with multiple offspring", Developmental Psychobiology, 07/1973 Publication	1 %
10	Yulia Ery Kurniawati, Yulius Denny Prabowo. "Model optimisation of class imbalanced learning using ensemble classifier on over-sampling data", IAES International Journal of Artificial Intelligence (IJ-AI), 2022 Publication	1 %
11	heanoti.com Internet Source	1 %
12	Submitted to Aston University Student Paper	<1 %
13	Submitted to The University of Manchester Student Paper	<1 %
14	cris.vtt.fi Internet Source	<1 %
15	baixardoc.com Internet Source	<1 %

16	dokumen.pub Internet Source	<1 %
17	hdl.handle.net Internet Source	<1 %
18	www.earthdoc.org Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

jtm

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7