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

by Wisoedhanie Widi Anugrahanti

Submission date: 18-Jul-2018 08:28PM (UTC+0800)
Submission ID: 983437138
File name: 30-1-146-2-10-20171109.pdf (597.59K)
Word count: 2279
Character count: 12312

RESEARCH ARTICLE
URL of this article: http://heanoti.com/index.php/hn/article/view/hn1208

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 improved 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 cased 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, excersise, 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 economiy, 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 1. Goodness of Split Value of predictor variables

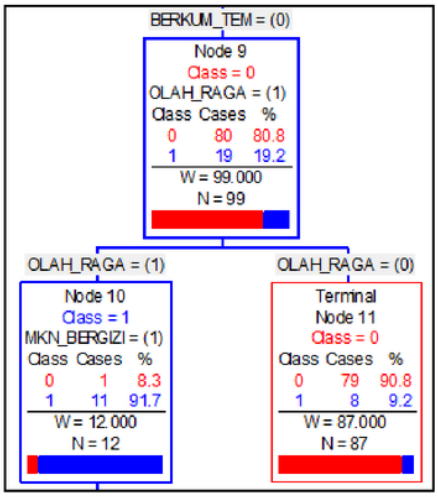
No	Variable	Goodness of split/ Improvement	N Left	N Right
1	Society participation	0.1659142	294	402
2	Get together with friends	0.1551249	381	312
3	Worship	0.1300618	400	296
4	Medical examination	0.1224152	245	451
5	Establish communication	0.1014271	199	497
6	Give attention	0.0882353	177	519
7	Exercise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

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:



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

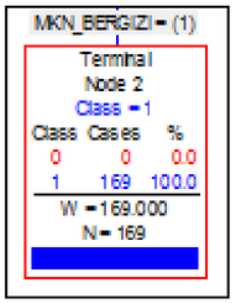
Figure 1. Class labelling

Figure 1showed 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 (exercise) = 8/87 = 0.091
P (not exercising) = 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:

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

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

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

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 :

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

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

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

At parent node 13 obtained :

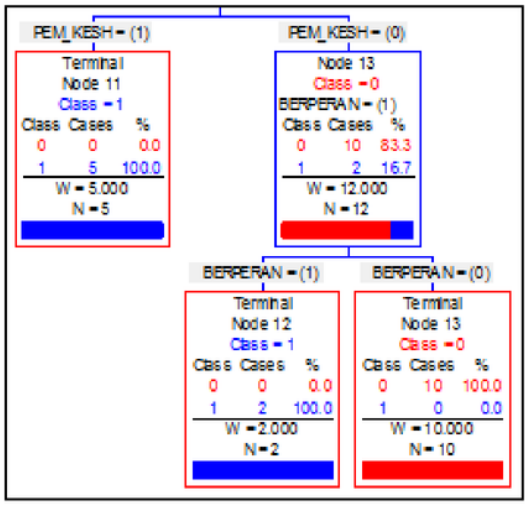
$$r\ (parent\ node\ 13) = 1 - \max P(j|simpul\ ter\ 13) = 1 - 1 = 0$$

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

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

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

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

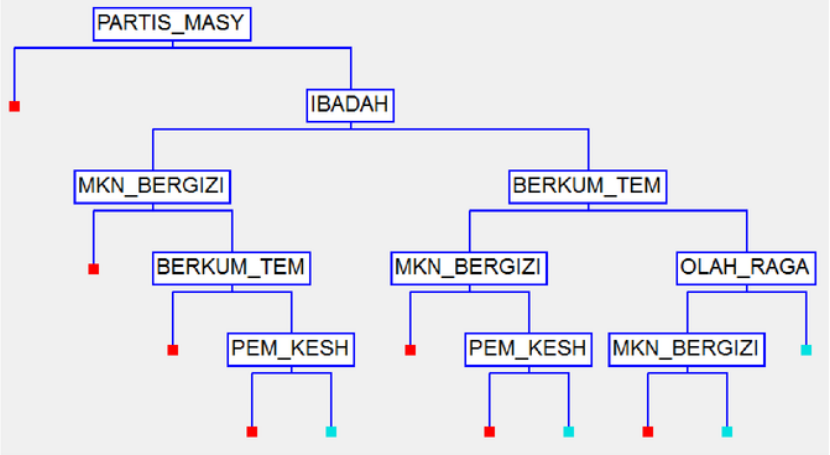


PEM_KESH=Medical examination; BERPERAN=Role

Figure.3 Pruning classification tree

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;

Participation in community activities (100,00000), medical examination (98,95988), providing nutritious food (68,60476), worship (57,36587), exercising (56,65865), get together with friends (54,90971) .



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

Implementation of Elderly Family Development		**Observed Class**	
		Not Implementing	Implement
Predicted Class	Not Implementing	106	0
	Implement	18	572

$$1 - APER = \frac{TP + TN}{TP + FP + TN + FN} = \frac{572 + 106}{572 + 18 + 106 + 0} = 0{,}9741$$

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

Implementation of Elderly Family Development		**Observed Class**	
		Not Implementing	Implement
Predicted Class	Not Implementing	106	0
	Implement	0	590

$$1 - APER = \frac{TP + TN}{TP + FP + TN + FN} = \frac{590 + 106}{590 + 0 + 106 + 0} = 1{,}0$$

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

Anderson, T. W. (1984). *An Introduction To Multivariate Statistical Analysis.* USA: Wiley.

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley and Sons Inc

Berk RA. (2008). *Statistical Learning from a Regression Perspective*. New York: Springer Science+Business Media

Breiman, L. (1996). Bagging Predictors, *Machine Learning*, Vol.24: 123-140.

Breiman, L., J. H. Friedman, R.A. Olshen, and C.J., Stone. (1993). *Classification And Regression Trees*, Chapman & Hall (Wadsworth, Inc), New York.

Direktorat Pengembangan Keluarga Nasional Badan Koordinasi Keluarga Berencana Nasional. (2010). *Materi Bina Keluarga Lansia (BKL).*

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learnig : Data Minning, Inference and Prediction. Second Edition* . New York: Springer-Verlag.

Hastie, Trevor. (2003). *Comparison of Learning Method.* Statistic Departement. Stanford University

Lewis, R. J. (2000). *An Introduction to Classification and Regression Trees (CART) Analysis. Annual Meeting of the Society for Academic Emergency Medicine.* California: UCLA Medical Center.

Sutton, C. O. (2005). *Classification and Regression Trees, Bagging and Boosting.* Handbook of Statistics

Wezel, M. P. (2007). Improved Customer Choice Predictions Using Ensemble Methods. *European Journal of Operational Research 181 (1)*, 436-452

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

ORIGINALITY REPORT

14%	11%	11%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Pankaj Bhaskar, Sheikh I Ahamed. "Privacy in Pervasive Computing and Open Issues", The Second International Conference on Availability, Reliability and Security (ARES'07), 2007 Publication	2%
2	doaj.org Internet Source	1%
3	dspace.uni.lodz.pl:8080 Internet Source	1%
4	scholar.sun.ac.za Internet Source	1%
5	www.biomedcentral.com Internet Source	1%
6	hal-audencia.archives-ouvertes.fr Internet Source	1%
7	support.sas.com	

	Internet Source	1%
8	www.m-hikari.com Internet Source	1%
9	ar.iiarjournals.org Internet Source	1%
10	biblioweb.u-cergy.fr Internet Source	1%
11	aliquote.org Internet Source	1%
12	pages.stern.nyu.edu Internet Source	1%
13	ejournal.unesa.ac.id Internet Source	1%
14	mpra.ub.uni-muenchen.de Internet Source	1%
15	Chiman Wong. "Classification of Imagery Movement Tasks for Brain-Computer Interfaces Using Regression Tree", Advances in Soft Computing, 2009 Publication	1%
16	Xia, Meng, Vijay Gupta, and Panos J. Antsaklis. "Networked State Estimation over a Shared Communication Medium", IEEE Transactions	<1%

on Automatic Control, 2016.
Publication

17	www.eurasip.org Internet Source	<1%
18	www-civil.eng.monash.edu.au Internet Source	<1%
19	journal.fsv.cuni.cz Internet Source	<1%
20	www.informs-sim.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	Off		

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

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6