Classification of Status of the Region on Java Island using C4.5, CHAID, and CART Methods

R A Syaraswati¹, I Slamet², and B Winarno¹
¹Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia
²Statistics Department, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia
E-mail: isnandarslamet@staff.uns.ac.id

Abstract. The indicator of region economic success can be measured by economic growth, presented by value of Gross Regional Domestic Product (GRDP). Java island has the biggest GDP contribution toward the Indonesian government, but not all of the region gives equality contribution. The C4.5, CHAID, and CART methods can be used for classifying the status of the region with nonparametric approach. The C4.5 and CHAID methods are non-binary decision tree, meanwhile the CART methods is binary decision tree. The purposes of this paper are to know how the classification and to determine the factors that influence on classification of the region. The dependent variable is status of the region which is divided into four categories based on Klassen typology. The result shows factors that have the biggest contribution on classification of status of the region on Java island based on C4.5 method are economic growth rate, electricity, gas, and water sector, and area. The factors that have the biggest contribution based on CHAID method are growth rate, manufacturing sector, and electricity, gas, and water sector, while based on CART method are growth rate, manufacturing sector, and electricity, gas, and water sector.

1. Introduction
Classification problems are very common in business, economic, health, and education. In economic, indicator of region economic success can be measured by economic growth, presented by value of Gross Regional Domestic Product (GRDP). Java island has the biggest GDP contribution toward the Indonesian government, but not all of the region gives equality contribution. To overview the growth pattern we can used Klassen typology. This technique divide the region into four quadrants, i.e quadrant I, II, III, and IV. There are many factors which's influence the division of the region, so that region is shifted into quadrant I, II, III, or IV. Therefore, overview is needed to solve classification problem. Classification method can be used to solve it.

There are two classification methods, parametric and nonparametric approach. Parametric approach has some assumptions that should be completed. Nonparametric approach does not need any assumption but it has high accuracy and easy to interpreted. One of classification method with nonparametric approach is classification tree. This classification is divided into two groups i.e binary and non-binary decision tree. C4.5 and CHAID (Chi-squared Automatic Interaction Detection) are non-binary decision tree, while CART (Classification and Regression Tree) is binary decision tree.
In this paper we use C4.5, CHAID, and CART methods. The C4.5 method uses entropy concept and can be used for discrete and continuous data. CHAID is an iterative technique that examines the predictors (e.g. classification variables) individually and utilizes them in the order dictated by their statistical significance (Gallagher et al. [5]). CART analysis is a powerful technique with significant potential (Lewis and Roger [10]).

The C4.5, CHAID, and CART methods have similar characteristic. Therefore, in this paper we applied classification of the status of the region on Java Island using C4.5, CHAID, and CART methods.

2. Decision Tree

2.1. C4.5 Method
Quinlan [12] made improvements to ID3 and extended to C4.5 in 1993. The C4.5 method can handle both of continuous and discrete attribute.

(1) Determine the root node. Root node is chosen by the biggest gain ratio. Let $S$ denote number of cases, $n$ is number of classes, and $p_i$ is the proportion of $S_i$ instances $S$. Therefore, the entropy of attribute $S$ is calculate as

$$ Entropy(S) = \sum_{i=1}^{n} p_i \times \log_2 p_i. $$

The information by attribute $A$ is defined as

$$ Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) $$

where $A$ is attribute, $|S_i|$ is number of cases in $i$ partition, and $|S|$ number of cases of $S$. Therefore, the information gain ratio of attribute $S$ is defined as

$$ GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} $$

where Split Info $(S,A)$ is defined as

$$ SplitInfo(S, A) = \sum_{j=1}^{n} p_j \times \log_2 p_j. $$

with $p_j$ is the proportion of $S_j$ instances $S$.

(2) Determine the branch node.

(3) Iterate until all of the cases have a same class. The splitting of classification tree is stopped when all of the records have a same class, there is no more splitting attribute on record, and there is no empty record.

2.2. CHAID Method
CHAID is concerned with predicting a single variable, known as the dependent variable, based on a number of other variables, referred to as predictor variables (Gallagher [5]). The main part of CHAID method as follows.

(1) Chi-squared test (contingency table). Contingency table is data analyzing to defined relationship between several variables. Chi-squared test is used as follows

$$ \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} $$

where $O_{ij}$ is row $i$ and column $j$ observation, $e_{ij}$ is row $i$ and column $j$ expected frequency, $r$ is number of row, and $c$ is number of column.

(2) Bonferroni correction. Bonferroni correction is a multiple comparison correction when several dependent or independent statistical tests are being performed simultaneously.
2.3. CART Method

CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure (Lewis and Roger [10]). The CART method consists of four basic steps as follows.

(1) Tree constructing. Tree is constructed using recursive splitting of node. Let $k, k = 1, 2, ..., m$ is class where $m$ is number of classes. Gini splitting index is as follows.

$$i(t) = 1 - \sum_{i=1}^{m} [P(k|t)]^2,$$

where $[P(k|t)]$ is relative frequency of class $j$ in node $t$. Splitter is chosen by the biggest goodness of split ($\phi(s,t)$), as follows

$$\phi(s,t) = i(t) - p_Lit_L - p_Rit_R$$

where $p_L$ is left node proportion, $it_L$ is left node heterogenous function, $p_R$ is right node proportion, and $it_R$ is right node heterogenous function.

(2) Stopping the tree constructing process. At this step, maximal tree has been produced. The process is stopped when there is only one observation in each of the child nodes, all observations within each child node have the identical distribution of predictor variables, making splitting impossible, or an external limit on the number of levels in the maximal tree has been set by the user (depth option) (Lewis and Roger [10]).

(3) Tree pruning. In order to generate a sequence of simpler and simpler trees, each of which is a candidate for the appropriately-fit final tree. For tree pruning, the method of cost-complexity pruning is used (Lewis and Roger [10]).

(4) Optimal tree selection.

3. Research Methodology

This paper used C4.5, CHAID, and CART methods to classify the status of the region on Java island. This paper used 2014 BPS data. The dependent variable is status of the region which's divided into four categories (i.e K I, K II, K III, and K IV) based on Klassen typology. Agriculture sector ($X_1$), mining and quarrying sector ($X_2$), manufacturing sector ($X_3$), electricity, gas, and water sector ($X_4$), construction sector ($X_5$), commerce, hotel, and restaurant ($X_6$), transportation and communication ($X_7$), finance sector ($X_8$), services ($X_9$), economic growth rate ($X_{10}$), original local government revenue ($X_{11}$), the literacy rate ($X_{12}$), percentage of poor people ($X_{13}$), and area ($X_{14}$) are independent variables.

The first step is to divide the 113 region on Java island into four quadrant based on Klassen Typology. Next, construct a classification tree with C4.5. Starting with counting Entropy value, Gain, and then GainRatio to find a best splitting attribute. The higher attribute of gain ratio is a root node. Then, we make branches for every attribute until it on the same class. To construct a classification tree with CHAID method, we begin with merging the independent variables with chi-square test, splitting, and stopping when no more significant independent variable. To construct CART classification tree, we begin with split the node, stop splitting, prune, and choose the optimal classification tree. After that, the classification tree can be interpreted.

4. Result

4.1. Klassen Typology

Klassen typology divided the status of the region into four quadrant. Quadrant I shows an advanced and fast growing region, quadrant II shows an advanced but pressured region, quadrant III shows a rapidly growing region, and quadrant IV shows a relatively underdeveloped region. There are 21 regions in quadrant I, 9 region in quadrant II, 42 region in quadrant III, and 41 regions in quadrant IV.
4.2. C4.5 Method

There are four steps to construct decision tree with C4.5 method.

(1) Determine the root node. To determine the root node, we begin with calculating Entropy, Gain, and then Gain Ratio. The result of calculating Gain Ratio is presented in Table 1.

| Attribute | Gain Ratio | Attribute | Gain Ratio | Attribute | Gain Ratio |
|-----------|------------|-----------|------------|-----------|------------|
| $X_1$     | 0.0876     | $X_6$     | 0.0781     | $X_{11}$  | 0.0870     |
| $X_2$     | 0.1075     | $X_7$     | 0.0787     | $X_{12}$  | 0.0391     |
| $X_3$     | 0.1182     | $X_8$     | 0.0626     | $X_{13}$  | 0.0851     |
| $X_4$     | 0.3190     | $X_9$     | 0.0315     | $X_{14}$  | 0.1628     |
| $X_5$     | 0.0728     | $x_{10}$  | 1.0000     |           |            |

Table 1 shows that $X_{10}$ has the biggest Gain Ratio, therefore $X_{10}$ is the root node.

(2) Determine the branch node. The branch node is the attribute that has the biggest Gain Ratio after the root node was deleted.

(a) First iteration. On $X_{10}$ first category, the biggest Gain Ratio is $X_4$ (0,1045). There are four categories in $X_4$, the fourth and fifth category classify the region into quadrant I. While first, second, and third category aren’t.

(b) Second iteration. On $X_4$ first category, the biggest Gain Ratio is $X_{12}$ (0,3212). There are two categories in $X_{12}$, the first category classify the region into quadrant I. While second category classify the region into quadrant III.

(c) Third iteration. On $X_4$ second category, the biggest Gain Ratio is $X_2$ (0,3506). There are four categories in $X_2$, the first and second category classify the region into quadrant I. While third and fourth category classify the region into quadrant III.

(d) Fourth iteration. On $X_4$ third category, the biggest Gain Ratio is $X_{14}$ (0,3021). There are three categories in $X_{14}$, the first and second category classify the region into quadrant I. While third category classify the region into quadrant III.

(e) Fifth iteration. On $X_{10}$ second category, the biggest Gain Ratio is $X_{14}$ (0,1579). There are three categories in $X_{14}$, the second and third category classify the region into quadrant IV. While first category isn’t.

(f) Sixth iteration. On $X_{14}$ first category, the biggest Gain Ratio is $X_5$ (0,3622). There are four categories in $X_{14}$, the first category classify the region into quadrant IV. While second, third, and fourth category classify the region into quadrant II.

All of branches node have classified the region, therefore process is stopped. The classification tree has 87.6% classification accuracy. The classification tree is presented in Figure 1.

Figure 1. C4.5 Classification Tree
4.3. CHAID Method

To construct decision tree with CHAID, we have to merge the independent variables to find the split values first. We apply chi-square test to merge the independent variables. After that, we determine which of the independent variables is the most significant in distinguishing among the dependent variable. The significant variables is given in Table 2.

| Parent Node | Independent Variable | $\chi^2$ | Split Values |
|-------------|----------------------|---------|--------------|
| 1           | $X_{10}$             | 113     | 1 and 2      |
| 2           | $X_4$                | 14,032  | 2,3,4,5 and 1|
| 3           | $X_3$                | 20,694  | 4 and 1,2,3  |
| 4           | $X_{14}$             | 9,808   | 3 and 1,2    |
| 5           | $X_{13}$             | 15,435  | 2,3 and 1    |
| 6           | $X_3$                | 14,983  | 4 and 1,2,3  |

Variable $X_{10}$, $X_4$, $X_3$, $X_{14}$, and $X_{13}$ are parent node in tree classification with CHAID. Classification tree is shown in Figure 2.

![Figure 2. CHAID Classification Tree](image)

4.4. CART Method

There are four steps to construct decision tree with CART Method.

(1) Comparing classification accuracy of data combination. The best data combination is 88:12 with 85.86% : 85.71%. Therefore, this combination data is used.

(2) Tree constructing. Tree constructing begins at the root node, which includes all variable in the learning data. To construct a tree we have to split the node with splitting criteria, index gini. The best splitter is $X_{10}$ with improvement of 0.2500. Variable $X_{10}$ end up being the root node. It split into two terminal node, terminal node 1 formed by $X_{10}$ first category and terminal node 2 formed by $X_{10}$ second category. Splitting process stops on terminal node 6 and 7. After that we get the maximum tree classification which has 10 nonterminal node and 9 terminal node.
(3) Tree Pruning. In order to generate a sequence of optimum tree, the method of cost-complexity pruning is used.

![Figure 3. Pruning Maximal Tree](image)

Optimal classification tree has resubstitution relative cost of 0.16683, complexity parameter of 0.031307, and Relative cost of 0.38889 +/- 0.05072.

(4) Optimal Tree Selection. We got optimal classification tree after prune the maximal tree classification. Optimal classification tree show by Figure 4.

![Figure 4. CART Classification Tree](image)

The variables that include on optimal classification tree are agriculture sector, mining and quarrying sector, manufacturing sector, economic growth rate, and original local government revenue.
5. Interpretation

5.1. C4.5 Algorithm
Classification tree with C4.5 method has 7 nonterminal node and 17 terminal node. Classification tree can predict accurately of 87.6%. There are 7 terminal node predicted the region into quadrant I, 3 terminal node into quadrant II, 4 terminal node into quadrant III, and 3 terminal node into quadrant IV. The factors that have the biggest contribution are economic growth rate, electricity, gas, and water sector, and area.

5.2. CHAID Method
Classification tree with CHAID method has 6 nonterminal node and 7 terminal node. Classification tree can predict accurately of 88.47%. There are 1 region in quadrant I that misclassified into quadrant III, 3 regions in quadrant II that misclassified into quadrant IV, 7 regions in quadrant III that misclassified into quadrant I, and 2 regions in quadrant IV that misclassified into quadrant II. The factors that have the biggest contribution are economic growth rate, manufacturing sector, and electricity, gas, and water sector.

5.3. CART Method
Classification tree with CART method has 5 nonterminal node and 6 terminal node. Classification tree with testing data can predict accurately of 85.71%. There are no misclassification in quadrant I, 1 region in quadrant II that misclassified into quadrant IV, no misclassification in quadrant III, and 1 region in quadrant IV that misclassified into quadrant II. The factors that have the biggest contribution are economic growth rate, manufacturing sector, and electricity, gas, and water sector.

6. Conclusion
Based on the analysis, we can conclude.

(1) C4.5 classification tree can predict accurately of 87.6%, CHAID classification tree can predict accurately of 88.47%, and CART classification tree can predict accurately of 85.71%. Therefore, the CHAID method is the best method for classification of status of the region on Java island.

(2) The factors that influence the classification of status of the region on Java island based on C4.5 method are mining and quarrying sector, electricity, gas, and water sector, construction sector, economic growth rate, the literacy rate, and area. The factors influence based on CHAID method are manufacturing sector, electricity, gas, and water sector, economic growth rate, percentage of poor people, and area, while based on CART method the factors are agriculture sector, mining and quarrying sector, manufacturing sector, economic growth rate, and original local government revenue.

References
[1] Agresti, A., An Introduction to Categorical Data Analysis second edition, New Jersey, Jon Wiley and Sons, Inc, 2007.
[2] Bagozzi, R.P., Advanced Methods of Marketing Research. Blackwell Publishers Ltd, Oxford, 1994.
[3] Badan Pusat Statistik, Pertumbuhan Ekonomi Indonesia Triwulan III/2014 mencapai 2,96 Persen, 2014. http://bps.go.id/Brs/view/id/225 (diakses tanggal 17 April 2016)
[4] Breiman, L., Friedman, J.H, Olshen, R.A, and Stone, C.J., Classification and Regression Tree, Capman and Hall, New York, 1984.
[5] Gallagher, C.A., Monroe, H.M., and Fish, J.L., An Iterative Approach to Classification Analysis, 2000. http://www.casact.org/pubs/dpp/dpp90/90dpp237.pdf (diakses tanggal 24 November 2015)
[6] Kasmir, Analisis Laporan Keuangan, Jakarta, Rajawali Pers, 2008.
[7] Kass, G.V., An exploratory technique for investigating large quantities of categorical data. Appl. Statist. 29, (1980), no.2, 119-127.
[8] Kusrini dan Taufiq, Emha. *Algoritma Data mining*, Yogyakarta, Andi Offsee, 2009.
[9] Lakshmi, B.N, Indumathi, T.S, and Ravi, N., *An Empirical Study on Decision Tree Classification Algorithms*, International Journal of Science, Engineering and Technology Research (IJSETR), (2015), vol. 4, issue 11.
[10] Lewis, M.D dan Roger, J., *An Introduction to Classification and Regression Tree (CART) Analysis*, Department of Emergency Medicine Harbor-UCLA Medical Center Torrance, California, 2000.
[11] Myers, J.H., *Segmentation and Positioning for Strategic Marketing Decisions*, Chicago, American Marketing Association, 1996.
[12] Quinlan, J. R.*Induction of Decision Trees*, Machine Learning, (1986), vol.1, 81-106.
[13] Ramani, R.G and Shanthi, S., *Classifier Prediction Evaluation in Modeling Road Traffic Accident Data*. Computational Intelligence & Computing Research (ICCIC), (2012), 1-4.
[14] Sadono, Sukirno, *Ekonomi Pembangunan Proses masalah dan Dasar Kebijakan*, Jakarta, Kencana, 2006.