Journal of Engineering and Applied Sciences 12 (19): 4837-4840, 2017
ISSN: 1816-949X

Predictive Modeling Analysis Impact of Predictor Variables Towards Dependent Variable

Warnia Nengsih
Department of Computer, Politeknik Caltex Riau University, Pekanbaru, Indonesia

Abstract: Predictive modeling is one of the concepts to find a pattern or a learning model for the next test data. One implementation of this modeling is the decision tree concept. Data used in the simulation is vacant land data. Indicator analysis was conducted to determine patterns or learning models produced from test results using predictor variables towards dependent variable as seen from variable selection as the root and number of variables. Thus, it can be obtained a result that number of variables that used affect the pattern or learning model resulted. Capturing the root to obtain the decision tree does not affect learning model that obtained, so any variable that is used as a root produces the same learning model. The accuracy of variable selection also affects the patterns or learning models resulted. The fewer and inaccuracy in choosing the predictor variables affect the pattern or learning model resulted. Therefore, determination of the used variables must meet the principles of validity.

Key words: Predictive modelling, predictor variable, decision tree, accuracy, principles, learning

INTRODUCTION

Predictive modeling is one of modeling types in mining data field. This model will find a pattern or knowledge that gained from learning data. One of mining method data that embrace predictive modeling is a decision tree method that is part of the classification techniques.

A decision tree (Quinlan, 1993) is a formalism for expressing such mappings and consists of tests or attribute nodes linked to two or more sub-trees and leafs or decision nodes labeled with a class which means the decision. A test node computes some outcome based on the attribute values of an instance where each possible outcome is associated with one of the subtrees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate subtree. When a leaf is eventually encountered, its label gives the predicted class of the instance.

Algorithm for tree construction:

1. Start at the root node
2. For each X, find the set S that minimizes
the sum of the node impurities in the two child nodes and choose the split {X*εS*} that gives the minimum overall X and S
3. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn. A measure of node impurity based on the distribution of the observed Y values in the node

The theory of a decision tree has the following main parts: a "root" node is the starting point of the tree; branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called "interior" nodes. "Leaf" or "terminal" nodes are nodes that do not have child nodes and represent a possible value of targetvariable given the variables represented by the path from the root (Cong and Tsokos, 2009).

Large training data sets have millions of tuples. Decision tree techniques have restriction that the tuples should reside in memory. Construction process becomes inefficient due to swapping of tuples in and out of memory. More scalable approaches are required to handle data (Changala *et al*., 2012). In general, a larger dataset helps the algorithms to make better decisions about variables. Larger test samples also allow for more accurate error estimates (Watanuma *et al*., 2006).

An improved learning algorithm based on the uncertainty deviation is developed. Rationality of attribute selection test is improved. An improved method shows better performance and stability.

Equivalence between multiple layer neural networks and decision trees is presented. Mapping advantage is to provide a self configuration capability to design process. It is possible to restructure as a multilayered network on given decision tree. A comparison of different types of neural network techniques for classification is

presented. Evaluation and comparison is done with three benchmark data set on the basis of accuracy (Jeatrakul and Wong, 2009).

In this modeling, the use of variable x or predictor variable greatly affects the outcome of y or the dependent variable. Errors in the determination of the variable and amount of variable used may affect the result pattern will certainly have an impact on the next test data. Starting from these problems, conducted a comparative analysis study of predictive modeling using decision tree with the test indicator is based on the selection of variables as the root and the number of variables used. Data used as the simulation is field data.

MATERIALS AND METHODS

Research method and experimental

Decision tree method: Decision tree method is a part of classification techniques. In this study, decision tree method used to determine the appropriate variable to apply on the process of multiple regression calculation. The following is a formula to calculate entropy:

$$\text{Entropy}(t) = -\sum_{j} p(j|\,t)\log_2 p(j|\,t)$$

To determine the gain ratios, use the following equation:

$$\text{GAIN}_{\text{split}} = \text{Entropy}(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n}\text{Entropy}(i)\right)$$

Parent node, splitting p into k partitions, n_i number of records in i partition tree is due to the fact that some of the reduced subsets at the non-leaf nodes do not necessarily contain examples of every possible value of the branching attribute (Freeman, 1987). As pointed out by Quinlan (1993), large decision trees are difficult to understand because each node has a specific context established by the outcomes of tests at antecedent nodes and the structure of the decision tree may cause individual sub-concepts to be fragmented. Rewriting the tree to a collection of rules, one for each leaf in the tree, would not result in anything much simpler than the tree, since there would be one rule for every leaf. However, since the antecedents of a rule may contain irrelevant conditions, the rule can be generalized by deleting these superfluous conditions without affecting its correctness.

System design: There is diagram block illustrating the flow of predictive modeling analysis to find a comparison of the result of patterns or learning models by using two test indicators.

Figure 1 shows the design of test indicator based on the use of variables that serve as root. From the test results, it will be known if the used of variable as the root

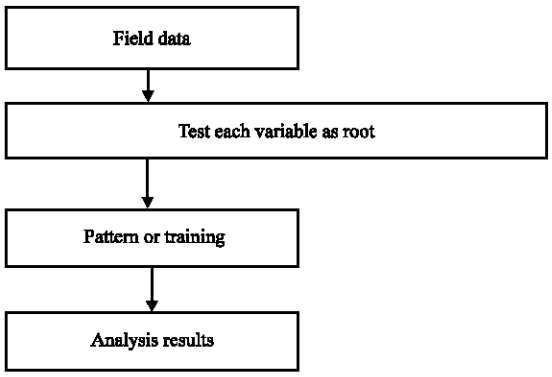

Fig. 1: The design of the test Indicator 1

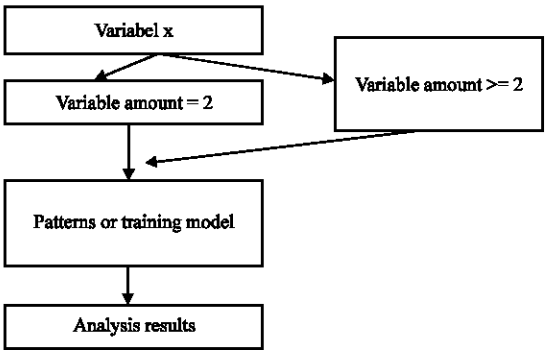

Fig. 2: Design of the test Indicator 1

in decision tree affects the pattern or learning models resulted. Figure 2 is a diagram block to analize the patterns obtained, based on the number of used variables indicator will it impact on the resulting pattern too.

RESULTS AND DISCUSSION

This test is using four X variables (predictor variable) and 1 Y variable (Dependent Variable). Variables predictor used is the cost (x1), location (x2), facility (x3) and environment (x4) and Purchase Variable (y). Simulations using primary data by the amount of data 20.

From these x variables (predictors), each variable will be tested as root to know the pattern or training models resulted. Table 1 indicates that test of x1, x2, x3 and x4 produce a pattern or learning models resulted.

Root capturing from the variables used in decision tree does not affect result of learning models. From the test results, known that any variables that is used as root will produce the same learning model. Following is using the amount of complete variables (Table 2). If the number of variables used <2, the result of learning models can be

Table 1: Test results based on the root that were used

Cost (x1)	Tarining modal	Location (x2)	Tarining model	Facility (x3)	Tarining modal	Environment (x4)	Trai Mc
Data 1	Yes	Data 1	Yes	Data 1	Yes	Data 1	Yes
Data 2	No	Data 2	No	Data 2	No	Data 2	No
Data 3	Yes	Data 3	Yes	Data 3	Yes	Data 3	Yes
Data 4	Yes	Data 4	Yes	Data 4	Yes	Data 4	Yes
Data 5	No	Data 5	No	Data 5	No	Data 5	No
Data 6	Yes	Data 6	Yes	Data 6	Yes	Data 6	Yes
Data 7	Yes	Data 7	Yes	Data 7	Yes	Data 7	Yes
Data 8	No	Data 8	No	Data 8	No	Data 8	No
Data 9	Yes	Data 9	Yes	Data 9	Yes	Data 9	Yes
Data 10	Yes	Data 10	Yes	Data 10	Yes	Data 10	Yes
Data 11	No	Data 11	No	Data 11	No	Data 11	No
Data 12	Yes	Data 12	Yes	Data 12	Yes	Data 12	Yes
Data 13	Yes	Data 13	Yes	Data 13	Yes	Data 13	Yes
Data 14	No	Data 14	No	Data 14	No	Data 14	No
Data 15	Yes	Data 15	Yes	Data 15	Yes	Data 15	Yes
Data 16	Yes	Data 16	Yes	Data 16	Yes	Data 16	Yes
Data 17	No	Data 17	No	Data 17	No	Data 17	No
Data 18	Yes	Data 18	Yes	Data 18	Yes	Data 18	Yes
Data 19	Yes	Data 19	Yes	Data 19	Yes	Data 19	Yes
Data 20	No	Data 20	No	Data 20	No	Data 20	No

Table 2: Test results based on amount of variables used (x<2)

Cost (x1)	Location (x2)	Facility (x3)	Environment (x4)	Training model
>= 350.000/m	Strategic	Exist	God	Yes
>= 350.000/m	Strategic	Exist	Poor	No
>= 350.000/m	Strategic	Exist	Very good	Yes
>= 350.000/m	Strategic	No	Good	Yes
>= 350.000/m	Strategic	No	Poor	No
>= 350.000/m	Strategic	No	Very good	Yes
>= 350.000/m	Not strategic	Exist	Good	Yes
>= 350.000/m	Not strategic	Exist	Poor	No
>= 350.000/m	Not strategic	Exist	Very good	Yes
>= 350.000/m	Not strategic	No	Good	Yes
>= 350.000/m	Not strategic	No	Poor	No
>= 350.000/m	Not strategic	No	Very good	Yes
<= 350.000/m	Strategic	Exist	Good	Yes
<= 350.000/m	Strategic	Exist	Poor	No
< = 350.000/m	Strategic	Exist	Very good	Yes
<= 350.000/m	Strategic	No	Good	Yes
<= 350.000/m	Strategic	No	Poor	No
<= 350.000/m	Strategic	No	Very good	Yes
<= 350.000/m	Not strategic	Exist	Good	Yes
<= 350.000/m	Not strategic	Exist	Poor	No
<= 350.000/m	Not strategic	Exist	Very good	Yes
<= 350.000/m	Not strategic	No	Good	Yes
<= 350.000/m	Not strategic	No	Poor	No
<= 350.000/m	Not strategic	No	Very good	Yes

Table 3: Test results based on amount of variables used (x<2)

Cost (x1)	Location (x2)	Training model
>= 350.000/m	Strategic	Yes
<= 350.000/m	Not strategic	Yes
>= 350.000/m	Not strategic	No
<= 350.000/m	strategic	No

Primary data of 2016

in Table 2. The application of the variables used amount, also affect the result of learning models. Determination of variables used must suitable to fulfill the principle of validity so the determination of variable sync with affected variable.

CONCLUSION

Following is the conclusion of research that had been conducted:

- Amount of the variables that are used affect the pattern or learning models resulted
- Root capturing to produce training models on decision tree does not impact the result of training models. From test results, any variable that is used as root produces the same model training
- The accuracy of variable selection affects pattern or learning models resulted

- The fewer and the less accuracy in choosing variable, the worse pattern or learning models produced
- Determination of variables that are used must meet the principle of validity so the determination of predictor variable syncs with dependent variable

REFERENCES

Changala, R., A. Gummadi, G. Yedukondalu and U.N.P.G. Raju, 2012. Classification by decision tree induction algorithm to learn decision trees from the class-labeled training tuples. Intl. J. Adv. Res. Comput. Sci. Software Eng., 2: 427-434.

Cong, C. and C. Tsokos, 2009. Theory and applications of decision tree with statistical software. Commun. Appl. Anal., 20: 1-10.

Freeman, J.D.H., 1987. Applied Categorical Data Analysis. Marcel Dekker, New York, USA., ISBN:0-824-77752-2,.

Jeatrakul, P. and K.W. Wong, 2009. Comparing the performance of different neural networks for binary classification problems. Proceedings of the 8th International Symposium on Natural Language Processing (SNLP09), October 20-22, 2009, IEEE, Australia, Oceania, ISBN:978-1-4244-4138-9, pp: 111-115.

Quinlan, J.R., 1993. C4.5: Program for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA., USA., ISBN: 1-55860-238-0, Pages: 302.

Watanuma, T., T. Ozaki and T. Ohkawa, 2006. Decision tree construction from multidimensional structured data. Proceedings of the 6th IEEE International Conference on Data Mining Workshops (ICDM) 2006, December 18-22, 2006, IEEE, Kobe, Japan, ISBN:0-7695-2702-7, pp: 237-241.