Journal of Engineering and Applied Sciences 12 (Special Issue 4): 6672-6674, 2017
ISSN: 1816-949X

Naive Bayes and Decision Tree Modelling for Comparative Analysis Method

Warnia Nengsih
Department of Computer, Politeknik Caltex Riau, Umbansari Street Rumbai,
28265 Pekanbaru Riau, Indonesia

Abstract: Predictive modelling is one of data mining modelling to predict a value based on a pattern which has been formed by other values. Knowledge or pattern which has formed is gained from data training tabulation. Classification is one of some predictive modelling that applied this concept. There is a large number of methods that is applied on classification technique such as Naive Bayes and decision tree. Both of these methods have the same concept, however, they have different calculation stage. Even though both of these methods in conception are predicting a new value based on existing pattern, however, after being tested by random testing data from the same data training, gained different pattern.

Key words: Naive Bayes, decision tree, modelling dan comparative method, calculation, classification, random

INTRODUCTION

Classification is one of some predictive modelling which have several methods; where as in conceptual this method finds a value based on pattern that formed by existing data training. Naive Bayes and decision tree are some methods derived from the classification technique. Both of these model classification types have the same concept but they differ in the calculation process as well as the pattern which created. Naive Bayes attempts to search the probability value from each variable that is used to the existing label.

Comparative analysis of these two models is conducted in order to acknowledge the pattern that created by using the same data. This is to identify whether there is any difference between the pattern that produced and the result of prediction from data experiment which is included.

MATERIALS AND METHODS

Theoretical framework

Naive Bayes method: The design and manufacturing domain is a natural candidate for data-mining applications because it contains extensive data. Besides enhancing innovation, data-mining methods can reduce the risks associated with conducting business and improve decision-making (Bressan and Vitria, 2003).

It might appear that classification tasks are only a minuscule subset of procedural tasks, but even activities such as robot planning can be recast as classification problems (Aitkenhead, 2008). Among these approaches, the Naive Bayes text classifier has been widely used because of its simplicity in both the training and classifying stage (Michie, 1982). Although, it is less accurate than other discriminative methods (such as SVM), numerous researchers proved that it is effective enough to classify the text in many domains (Chakrabarti *et al*., 2003).

Naive Bayes is one of the methods from classification technique of data mining where this method is applying predictive model concept (Joachims, 1998). Data training is included in order to form the pattern which is used as a reference to form the new value. Naive Bayes method employs bayes method principal use conditional probabilities as the foundation. If P (X|Ci) can be identified by probability calculation on class, then class (label) from sampel X data is class (label) which have:

$$P(X|Ci)*P(Ci)$$

P(X|Ci)*P(Ci) maximum where P(X|Ci) variabel probability on every class and P(Ci) Probability class value.

Decision tree method: Decision tree is also one of the classification methods with the same concept in predictive modeling. The method emphasizes more on entropy value, gain value and decision tree formation (Aviad and Roy, 2011). The pattern which obtained from data training that has been processed is used as a reference to predict the value from data experiment:

$$\text{Entropy value } (S) = -P+\log_2 P_+ -p-\log_2 p-$$

Where:

S = Data sample value
P+ = Positive amount to certain criteria value
P- = Negative amount for certain criteria value

The various selection criteria have been compared empirically in a series of experiments (Balamurugan and Rajaram, 2009).

Design system: The following is a diagram block from naïve bayes comparative and decision tree (Fig. 1). Data training for both of these methods is data field. Data field is consequently processed by using Naive Bayes and decision tree and furthermore will be observed if they generate the same pattern, some percentage of similarity level and to observe prediction value they create. The following is a decision tree method flowchart (Fig. 2 and 3).

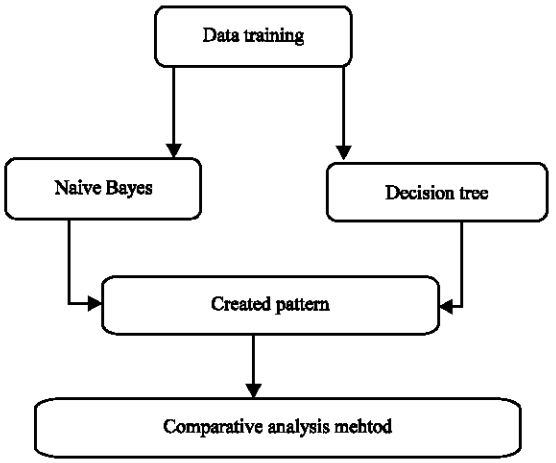

Fig. 1: Diagram block of comparative analysis method

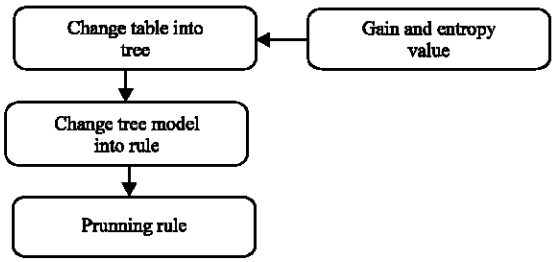

Fig. 2: Flowchart decision tree

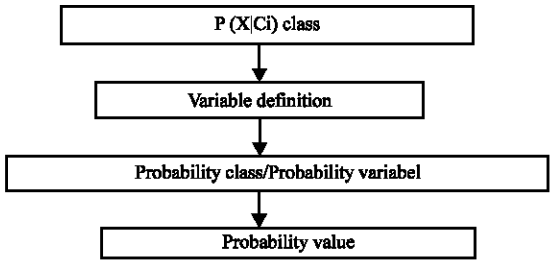

Fig. 3: Plot of naive bayes method

Training that exist in the table is changed into decision tree by finding the gain and entropy value. After that change the decision tree model and simplify rule from the rule which has been processed on the tree.

It begins by doing calculation on the every variable value on class, then finds the class probability value and makes calculation to probability value from each variable in order to gain probability value.

RESULTS AND DISCUSSION

The data that is used is data location which favored based on existing soil condition. The amount of the data is 58 records. Data training is first conversed into data transformation by dividing it into several groups (Fig. 4).

In order to test the probability that has been searched, the data testing that randomly made is necessary (Fig. 5).

Random testing table shows the class result which obtained by using 3 data experiment with the value of each class is L (record 21-30), L (Record 21-30) dan J (Record 1-10).

The following is tabulation result by using decision tree method (Table 1). Gain ratios value ilustrated in Table 2.

Even though both of these methods in conception are predicting a new value based on existing pattern, however, after being tested by random testing data from the same data training, gained different pattern.

Table 1: Entropy value

Y	Frek	PJ	log2.pj	Entropy value
J	10	0.5	-1.1	0.509709204
K	10	0.5	-1.1	0.509709204
L	1	0	-4.4	0.209157973
Total	21	1	H(t)	1.228576380

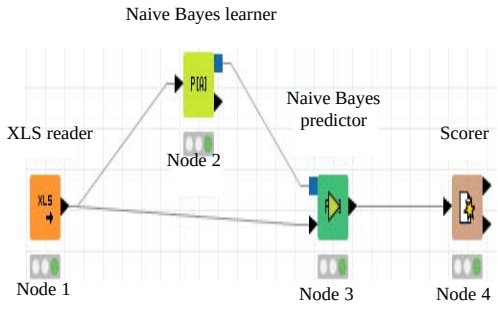

Fig. 4: Naive Bayes calculation

Table 2: Gain ratios value

Variables	J	K	L	P(J)	P(K)	P(L)	"-P(J) X Log2P(J)"	"-P(K) X Log2P(K)"	"-P(L) X Log2P(L)"	Total	P	Total XP	E (total XP)	Gain
X1														
A	4	0	0	1	0	0	0	0	0	0	0.190476	0	0.459594498	0.768982
B	0	1	0	0	1	0	0	0	0	0	0.047619	0		
C	4	0	0	1	0	0	0	0	0	0	0.190476	0		
D	2	4	1	0.285714	0.571429	0.142857	0.516387121	0.46134567	0.401050703	137878	0.333333	0.4595945		
E	0	5	0	0	1	0	0	0	0	0	0.238095	0		
A	4	3	0	0.571429	0.428571	0	0.46134567	0.523882466	0	0.98523	0.333333	0.3284094	0.669834379	0.558742
B	0	0	1	0	0	1	0	0	0	0	0.047619	0		
C	3	0	0	1	0	0	0	0	0	0	0.142857	0		
X2														
D	1	1	1	0.333333	0.333333	0.333333	0.528320834	0.528320834	0.528320834	1.58496	0.095238	0.1509488		
E	2	2	0	0.5	0	0	0.5	0.5	0	1	0.190476	0.1904762		
F	0	4	0	0	0	0	0	0	0	0	0.190476	0		
A	4	3	0	0.571429	0	0	0.46134567	0.523882466	0	0.98523	0.238095	0.2345781	0.537650888	0.690925
B	0	4	1	0	0.8	0.2	0	0.257542476	0.464385619	0.72193	0.238095	0.1718876		
X3														
C	1	0	0	1	0	0	0	0	0	0	0.047619	0		
D	2	1	0	0.666667	0.333333	0	0.389975	0.528320834	0	0.9183	0.142857	0.1311851		
E	3	0	0	1	0	0	0	0	0	0	0.142857	0		

X1	X2	X3	Y	P(X)|J	P(X)|K	P(X)|L	P(X)|M	P(X)|N	P(X)|O
A	C	E	L	0.036	0	0.04	0.06	0	0
B	E	A	L	0	0.006	0.01	0	0	0
A	D	A	J	0.016	0	0	0	0	0.015625

Fig. 5: Random testing

CONCLUSION

The data that is used is data location which favored based on existing soil condition. The amount of the data is 58 records. Data training is first conversed into data transformation by dividing it into several groups. Even though both of these methods in conception are predicting a new value based on existing pattern, however, after being tested by random testing data from the same data training, gained different pattern.

REFERENCES

Aitkenhead, M.J., 2008. A co-evolving decision tree classification method. Expert Syst. Applic., 34: 18-25.

Aviad, B. and G. Roy, 2011. Classification by clustering decision tree-like classifier based on adjusted clusters. Expert Syst. Appl., 38: 8220-8228.

Balamurugan, A.S.A. and R. Rajaram, 2009. Effective solution for unhandled exception in decision tree induction algorithms. Expert Syst. Appl., 36: 12113-12119.

Bressan, M. and J. Vitria, 2003. On the selection and classification of independent features. IEEE. Trans. Pattern Anal. Mach. Intell., 25: 1312-1317.

Chakrabarti, S., S. Roy and M.V. Soundalgekar, 2003. Fast and accurate text classification via multiple linear discriminant projections. VLDB J., 12: 170-185.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Kumar, A., M. Hanmandlu and H.M. Gupta, 2013. Ant colony optimization based fuzzy binary decision tree for bimodal hand knuckle verification system. Expert Syst. Appl., 40: 439-449.

Larose, D., 2006. Data Mining Methods and Models. John Wiley and Sons, New York, USA., ISBN-10: 0471666564, pp: 227-229.

Michie, D., 1982. Experiments on the mechanization of game-learning. 2-rule-based learning and the human window. Comput. J., 25: 105-113.