International Journal of Advanced Smart Convergence Vol.9 No.4 184-191 (2020)
http://dx.doi.org/10.7236/IJASC.2020.9.4.184

IJASC 20-4-24

Finding a plan to improve recognition rate using classification analysis

SeungJae Kim[1], SungHwan Kim[2]

[1]Assistant Professor, Department of Convergence Honam University, Korea
[2] Research professor, College of IT convergence engineering and SW Center University Business unit, Chosun University, Korea
cdma1234@hanmail.net, shkimtop@chosun.ac.kr

Abstract

With the emergence of the 4th Industrial Revolution, core technologies that will lead the 4th Industrial Revolution such as AI (artificial intelligence), big data, and Internet of Things (IOT) are also at the center of the topic of the general public. In particular, there is a growing trend of attempts to present future visions by discovering new models by using them for big data analysis based on data collected in a specific field, and inferring and predicting new values with the models. In order to obtain the reliability and sophistication of statistics as a result of big data analysis, it is necessary to analyze the meaning of each variable, the correlation between the variables, and multicollinearity. If the data is classified differently from the hypothesis test from the beginning, even if the analysis is performed well, unreliable results will be obtained. In other words, prior to big data analysis, it is necessary to ensure that data is well classified according to the purpose of analysis. Therefore, in this study, data is classified using a decision tree technique and a random forest technique among classification analysis, which is a machine learning technique that implements AI technology. And by evaluating the degree of classification of the data, we try to find a way to improve the classification and analysis rate of the data.

Keywords: *Machine Learning; Decision Tee; Random Forest: Classification: Recognition*

1. Introduction

With the emergence of the 4th Industrial Revolution, core technologies that will lead the 4th Industrial Revolution such as artificial intelligence (AI), big data, and Internet of Things (IOT) are also at the center of the topic of the general public. However, no matter how much data you have and how well you can analyze it, if the data for analysis is not classified properly, you will not be able to obtain sophisticated results no matter how much you analyze it, and this result will be unreliable. In other words, the data must be classified very well in accordance with the hypothesis test for the purpose of analysis. Among big data analysis (BDA) techniques, classification analysis (CA) is also one of the machine learning (ML) techniques [1]. CA belonging

Manuscript Received: November. 27, 2020 / Revised: December. 5, 2020 / Accepted: December. 5, 2020
Corresponding Author: cdma1234@hanmail.net
Tel: +82-62-940-5639, Fax: +82-62-940-5005
Assistant Professor, Department of Convergence Honam University, Korea

to ML includes Decision Tree (DT) [2-4], Random Forest (RF) [5-7] and Support Vector Machine (SVM) [8-10] and logistic regression analysis (LRA). In BDA, CA uses this analysis method to train and test data to classify data. The numerical values obtained through CA are analyzed statistically to determine the degree of classification of the data. In this study, when market utilization data according to customer types is given, data on the degree of market utilization according to independent variables are classified through CA of DF and RF techniques. And by evaluating the degree of classification of the data, we try to find a way to improve the CA rate of data by DF and RF.

2. Definition of classification analysis

2.1 Machine Learning Technique

ML is one of the fields of artificial intelligence that allows machines to make decisions and infer like humans. It is a technology that gives them the ability to infer by learning data having a specific meaning (ML). ML can be divided into three categories: supervised learning (SL), unsupervised learning (UL), and reinforcement learning (RL). First, SL finds a prediction function from specific data and predicts the result of a new input value. Second, UL learns random data by itself. Third, RL reinforces learning through errors. The ML techniques to be applied in this study are DT and RF.

2.2 Decision Tree

a. Decision Concept

It is one of the treadmill models. Compared to other classification and analysis techniques, the rules are relatively quick, simple and easy to understand. In addition, the decision tree is an analysis method that performs classification and prediction by charting decision rules, and it is easy to understand and explain because the process of classification or prediction is expressed by inference rules based on a tree structure.

DT uses the concept of impurity to select branching criteria in the decision tree, which means the complexity of the data. In other words, it means how much different data are mixed within one category. This should be set so that the impurity of the child node is reduced compared to the impurity of the current node when the branch reference is set. Equation 1 is the impurity function.

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2 \qquad p_i\,(i = 1, 2, \ldots, k) \tag{1}$$

G is the amount of information acquired, S is the set of all events, and P is the proportion of classes in each group.

Table 1. Analysis process of 5 steps of decision tree

stage	Step-by-step analysis	
Step 1	Analysis process	**Creation of decision tree**
	According to the purpose of analysis, it has appropriate separation criteria and stopping rules.	
Step 2	Analysis process	**Pruning**
	Remove branches that have the potential to increase the classification error or have induction rules that are inappropriate.	
Step 3	Analysis process	**Feasibility evaluation**
	Analyze cross-validation using gains chart, risk chart, or verification data.	
Step 4	Analysis process	**Interpretation and prediction**

	Interpret the decision tree and establish a predictive model.	
Step 5	Analysis process	**Decision tree formation**
	In the above process, different decision trees are formed depending on how the separation criteria, suspension rules, and evaluation criteria are applied.	

2.3 Random Forest

a. Random Forest overview

RF is an ensemble ML model based on a DT. RF is an algorithm made by combining multiple decision trees, and some trees can be misclassified, but since it creates a large number of trees, the misclassification does not have a significant effect on prediction. Figure 1 shows the structure of a RF.

As shown in Figure 1, the internal operation process of the random forest is classified according to the structure of the decision tree. It is determined by several decision trees, not just one decision tree.

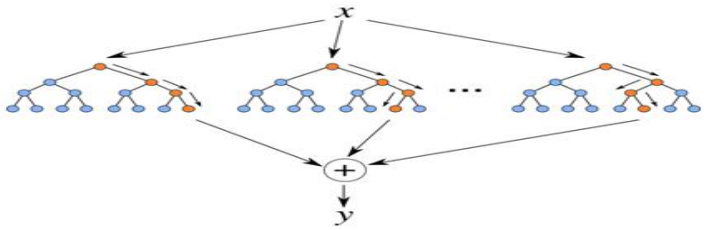


Figure 1. Structure of Random Forest

That is, the ensemble used in RF creates several decision trees by randomly selecting different data in a data set, and obtains results through a majority voting method. This method of randomly selecting data and variables is called bagging (bootstrap aggregation). In other words, bagging does not learn individual trees using the same whole data, but constructs a DT by sampling training data from the whole data. As such, the RF analysis is effectively executed on large amounts of data and has high accuracy even if it is executed without removing the variables that will cause errors even when using many variables.

In the end, votes are made on the results of each tree's classification in the RF, and the result with the most votes is selected as the final classification result.

b. Random Forest Model

The RF model constructs a resampling tree of training data from all data, and has a structure in which random sampling is performed from the training data. Figure 2 shows the internal structure of a RF.

As shown in Figure 2, the internal working structure of the random forest is classified by several decision trees. At this time, train data is not created as a single data set during the learning process for classification. Data sets are created as many as the number of decision trees in the internal structure of the random forest.

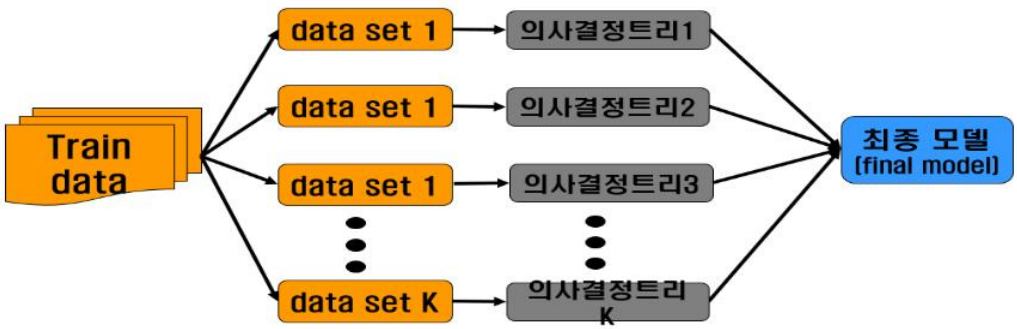


Figure 2. Internal structure of Random Forest

3. Classification Analysis Experiment

3.1 Subject and Method of Experiment

In this study, 1178 data were used for an arbitrary market in Gwangju Metropolitan City, and the research hypothesis for CA is to examine how consumption patterns differ according to customers using the market through CA. The criteria for consumption patterns examine the purchase desire according to the amount of time the customer stays in the market and the probability of being purchased together when purchasing a specific product.

CA on the consumption patterns of customers using the market analyzes 1178 data using R program, using DT and RF method among BDA methods. The internal operation result according to each technique is classified into the final low, middle, and high consumption patterns according to variables. In addition, the DT is divided into a train (training) set and a test (test) set through an internal operation among the total data, and after learning each, the classification rate is calculated. The smaller the difference, the better it can be said. Figure 3 is a part of the csv file of 1178 market usages used in the experiment.

As shown in Fig. 3, the number of data used in the experiment is 1178. he 1178 data are divided into a total of 6 independent variables. In addition, one of the six independent variables is a variable to distinguish large, medium, and small, and the interval is divided into high, middle, and low.

	A	B	C	D	E	F
1	freq	c_use	s_time	goods	respon	respon1
2	6	4.5	1.5	2.9	3.725	middle
3	7.7	6.7	2.2	3.8	5.1	high
4	6.3	4.7	1.6	3.3	3.975	high
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1176	5.1	1.5	0.2	3.4	2.55	low
1177	4.6	1.4	0.3	3.4	2.425	low
1178	5.4	4.5	1.5	3	3.6	middle
1179	5.5	3.7	1	2.4	3.15	low

Figure 3. Market usage data of 1178

3.2Classification Analysis using Decision Tree

For the DT analysis, 1178 experimental data were imported and classified by using the following code to classify them into a train set and a test set.

```
# 데이터 분할
set.seed(123)
ind <- sample(2, nrow(tree), replace = TRUE,  prob = c(0.7, 0.3))
train <- tree[ind==1, ]
test <- tree[ind==2, ]
str(test)
str(train)
```

a. Train

In the DT analysis, 833 data classified as train were classified into 5 branches by the pruning function, prune(), and the large categories were classified into three categories: low, middle, and high. Fig. 4 shows the classification criteria of each independent variable analyzed by the decision tree, and shows the classification results for 833 train sets.

As shown in Figure 4, pruning is performed by the prune() function by the internal function of the decision tree. As the pruning criterion by the prune() function was determined, it was classified into several. However,

by a large standard, you can see one on the left and two on the right.

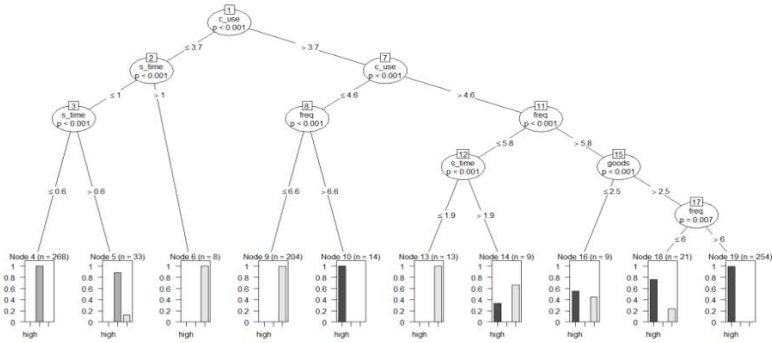


Figure 4. Classification of 833 train sets

The result of classification of consumption patterns according to the DT analysis is shown in Figure 5. Of the total 1178 data, train data was divided into 833. In the relationship between these variables, when high, middle, and low are regarded as cross-relationships, it can be seen that classification was done well in each classification process. In fact, it can be confirmed that the high classification rate was 0.9783% even through the calculation of the internal classification rate.

As shown in Figure 5, train data is classified from 1178 to 833, and it can be seen that it has been trained. Of the 833 data, 10 out of 298 belonging to the high section were misclassified as middle. 4 out of 234 in the middle section were misclassified as high. 4 out of 301 in the low section were misclassified as middle.

```
> addmargins(tree2)
tr.train  high   low middle   Sum
  high     288     0     10   298
  low        0   297      4   301
  middle     4     0    230   234
  Sum      292   297    244   833

> diag(tree2)
High   low middle
 288   297    230
> sum(diag(tree2))/sum(tree2) # 분류율 계산
  [1] 0.9783914
```

Figure 5. Train analysis of decision tree

b. Test

Out of the total market use data, 345 were classified as a test set, and if classified by DT analysis as in the previous train set, they were classified as shown in Figure 6.

As shown in Figure 6, test data is 345 data out of 1178 data, and since there are fewer data than train data, relatively few pruning was performed. There are 3 classifications by pruning.

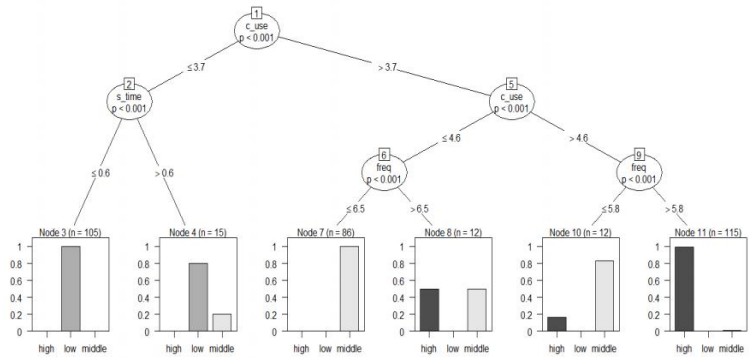


Figure 6. Classification of 345 test sets

As shown in Figure 7, out of the total 1178 data, test data were divided into 345 data. In the relationship between these variables, when high, middle, and low are considered as subject relations, it can be seen that classification is well done in each classification. In fact, it can be confirmed that the high classification rate was 0.965% even through the calculation of the internal classification rate.

As shown in Figure 7, it can be seen that train data is classified into 1178 to 345 trains. Of the 345 data, 7 out of 127 belonging to the high section were misclassified as middle. 2 out of 98 belonging to the middle section were misclassified as high. 3 out of 120 in the low section were misclassified as middle.

```
> addmargins(tree4)
te.tree   high    low   middle    Sum
 high      120      0        7    127
 low         0    117        3    120
 middle      2      0       96     98
 Sum       122    117      106    345
> diag(tree4)
   high    low   middle
    120    117       96
> sum(diag(tree4))/sum(tree4)
[1] 0.9652174
```

Figure 7. Test analysis of decision tree

3.3 Classification Analysis using Random Forest

For RF analysis, 1178 experimental data were imported and classified using the following code to classify them into a train set and a test set.

By using the "size=N*2/3" property, it was declared that only 70% of the train set data will be used and the remaining 30% of the test set data will be used.

```
# training/ test data : n = 150
set.seed(123)
N<-nrow(tree)
tr.idx<-sample(1:N, size=N*2/3, replace=FALSE)
```

As shown in Figure 7, out of the total 1178 data, test data were divided into 345 data. In the relationship between these variables, when high, middle, and low are considered as subject relations, it can be seen that classification is well done in each classification. In fact, it can be confirmed that the high classification rate was 0.965% even through the calculation of the internal classification rate.

a. Train

As for the classification result of the consumption type according to the RF analysis, as shown in Fig. 8, the result of separating and analyzing the train data into 785 out of the total 1178 data showed a very high classification rate of 0.998%.

As shown in Figure 8, it can be seen that train data is classified from 1178 to 785 trains. Of the 785 data, 1 out of 289 belonging to the high section was misclassified as middle. All 223 out of 223 belonging to the middle section were correctly classified as middle. Of the 273 in the low section, all 273 were correctly classified as low.

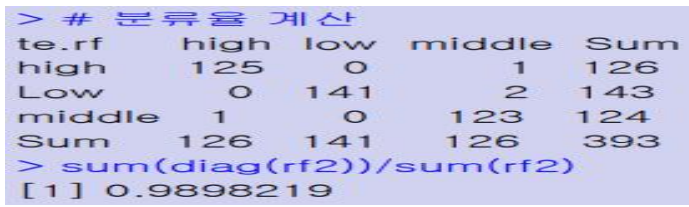
```
> # 분류율 계산
tr.rf   high   low   middle   Sum
high    288      0        1   289
low       0    273        0   273
middle    0      0      223   223
Sum     288    273      224   785
> sum(diag(rf1))/sum(rf1)
[1] 0.9987261
```

Figure 8. Random forest train analysis

b. Test

As for the classification result of consumption type according to the RF analysis, as shown in Fig. 9, the result of separating and analyzing the train data into 393 out of the total 1178 data showed a very high classification rate of 0.989%.

As shown in Figure 12, it can be seen that train data is classified into 1178 to 393 trains. Among the 393 data, 1 out of 126 in the high section was misclassified as middle. One out of 124 in the middle section was misclassified as high. Two out of 143 in the low section were misclassified as middle.



```
> # 분류율 계산
te.rf     high  low  middle  Sum
high       125    0       1  126
Low          0  141       2  143
middle       1    0     123  124
Sum        126  141     126  393
> sum(diag(rf2))/sum(rf2)
[1] 0.9898219
```

Figure 9. Random forest test analysis

4. Conclusion

In the era of the 4th industrial revolution, at the time when trying to develop various technologies using AI in all fields of society, the best of consumption and the maximization of profits are being optimized. In order to obtain sophisticated BDA results, data must be well classified according to the purpose to be analyzed. From this point of view, this study conducted an experiment to investigate the methodology of which method of two CA methods, DT and RF, among the various CA of ML techniques, classify data with less misclassification rate, more efficient and stable.

In this experiment, DT also had a high classification rate, but misclassification appeared due to interference between data and data. However, in the case of RF, the error rate is significantly lower than that of DT, so in the big data field dealing with more data, use RF rather than DT. It is determined that the data will be classified.

In the future, we will compare and analyze the performance between DT and RF after examining the performance using CA that was not covered in this study

.

References

[1] Tom M. Mitchell, “The discipline of machine learning(Vol. 9),” Carnegie Mellon University, Shcool of Computer Science, MachineLearning Department, 2006.
DOI: https://www.researchgate.net/publication/268201693_The_Discipline_of_Machine_Learning

[2] Young Jin Kim, Joung Woo Ryu, Won Moon Song and Myung Won Kim, “Fire Probability Prediction Based on Weather Information Using Decision Tree,” Journal of KIISE, JOK: software and application“, Vol.40, No.11, 2013.11.
DOI: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02283313

[3] Yoo, Se Hee, Park, Il Su and Kim, Yoomi, “A Decision-Tree Analysis of Influential Factors and Reasons for Unmet Dental Care in Korean Adults,” Health and Social Welfare Review, Vol.34, No.4, *pp.294-335*, 2017
DOI: https://dx.doi.org/10.15709/hswr.2017.37.4.294.

[4] Yoon, Tae-Tok and Lee, Jee-Hyong, “Design of Heuristic Decision Tree (HDT) Using Human Knowledge,” Korean Institute of Intelligent Systems, Vol.19, No.1*, pp.161-164*, 2009. 4
DOI: https://doi.org/10.5391/JKIS.2009.19.4.525.

[5] Donghoon Lee, Songhwai Oh, “Head Pose Estimation Using Random Forests,” The Korean Institute of Information Scientists and Engineers, Vol.40, No.8*, pp.435*-441, 2013.8.

DOI: http://kiise.or.kr/e_journal/2013/8/SA/pdf/02.pdf

[6] Jun Heon Lee and Jun Geol Baek, "Real-time control chart using a random forest-based multi-category classifier," Korean Institute Of Industrial Engineers*, pp. 673-682*, 2017.11
DOI http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07262460

[7] Kim, Pan Jun, "An Analytical Study on Automatic Classification of Domestic Journal articles Using Random Forest," Journal of the Korean Society for information Management , vol.36. no.2*, pp.57-77*, 2018
DOI: https://doi.org/10.3743/KOSIM.2018.35.2.037.

[8] Um, Nam-Kyoung , Woo, Sung-Hee and Lee, Sang-Ho, "The Hybrid Model using SVM and Decision Tree for Intrusion Detection," KIPS Transactions on Computer and Communication Systems, Vol.14, No.1, *pp.1-6*, 2007.
DOI: 10.3745/KIPSTC.2007.14.1.1, Full Text

[9] Ju, Sang-Lim, Kim, Nam-Il and Kim, Kyung-Seok, "Deep Learning-based Antenna Selection Scheme for Millimeter-wave Systems in Urban Micro Cell Scenario," The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 20, No. 5, *pp.57-62*, Oct. 31, 2020.
DOI: https://doi.org/10.7236/JIIBC.2020.20.5.57.

[10] Lee, Gye-Sung and Kim, In-Kook, "A Study on Simplification of Machine Learning Model," The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol.16, No.4, *pp.147-152*. Aug, 31, 2016.
DOI: http://dx.doi.org/10.7236/JIIBC.201616.4.147