Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI:10.21474/IJAR01/10084
DOI URL: http://dx.doi.org/10.21474/IJAR01/10084

RESEARCH ARTICLE

URBAN CLASSIFICATION BASED ON RANDOM FOREST ALGORITHM.

Peng Lin and Lixin Yang.
School of Mathematics and Statistics ,Shandong University of Technology, Zibo, China.

Manuscript Info

Manuscript History
Received: 14 September 2019
Final Accepted: 16 October 2019
Published: November 2019

Key words:-
Random Forest Algorithm, City Classification, Index System, Objective Happiness.

Abstract

This paper discussed the research content from five parts. The first chapter mainly introduced the background and significance of the research and the research status at home and abroad. The second chapter introduced the theoretical knowledge and implementation of the research method, namely the random forest algorithm. The third chapter introduced the index system of city classification.The fourth chapter introduced the establishment of stochastic forest algorithm model and discusses the feasibility of the model from the model results. The fifth chapter made a summary of the whole thesis.

Introduction:-

Urban classification research is to classify cities according to some indicators so as to better treat urban development. Most of the analysis methods for city classification are comprehensive analysis method and function division method. The comprehensive analysis method adopts the common economic indicators, and the functional analysis rules are divided according to the urban functions, which adopt the urban function indicators.The research methods used include factor analysis, regression analysis or cluster analysis.However, in the new era of big data, it is more meaningful to use machine learning to analyze and process data.Random forest is a kind of machine learning algorithm, which integrates many decision trees into a forest and USES the combined forest to predict the final results for classification.Applying random forest to urban classification can get more accurate classification results.
Random Forest

As a newly emerging and highly flexible machine learning algorithm, the idea of random forest algorithm is to establish a forest. The establishment method of the forest is random, including numerous decision trees, all of which have high prediction accuracy and are almost irrelevant to each other. The combination of these decision trees is used to form a prediction model.Random forest algorithm is a kind of integrated algorithm (Ensemble Learning), belongs to the type of Bagging, its principle is weak by combining multiple classifiers, each weak classifier is given a vote, will all the results are combined to get a final result, in order to get a strong classifier, thus making the final the result of the random forest model has high accuracy and generalization performance.The high prediction accuracy of random forest algorithm is mainly attributed to the two factors of "random" and "forest".The meaning of "random" is that the selection and combination of decision trees are random, so that it has the ability to resist overfitting.The meaning of "forest" is that there are a large number of decision trees in the random forest model, thus ensuring the accuracy of its results.

Random forest is a kind of machine learning algorithm. Besides being flexible and easy to use, the most important point is that compared with other algorithms, random forest can also get good results in most cases without super-

Corresponding Author: Peng Lin
Address:- School of Mathematics and Statistics ,Shandong University of Technology, Zibo, China.

parameter tuning. In addition, randomness is insensitive to multivariate collinearity, which means that its results are not affected by abnormal data, so it can process as many as thousands of explanatory variables, which is regarded as one of the best algorithms at present.

The main factors affecting the performance of random forest classification:
Classification intensity of each tree in the random forest: The intensity of decision tree in the forest determines the classification intensity of the random forest, that is, the accuracy of the results of the random forest. The higher the intensity of decision trees, namely, the more luxuriant the branches and leaves of each tree, the more accurate the voting results will be, and the higher the classification performance of random forests will be.

The correlation between random tree in the forest: Between the decision tree in the forest also decided the classification performance of the random forest, the greater the correlation between the decision tree, the tree to tree branches and leaves of interlock, the more will lead to the decision tree is too high, the fitting degree between the classification of the random forest performance will be worse. Therefore, the correlation between decision trees is not high.

Two important parameters in random forest:
Number of variables selected by decision tree nodes (mtry): The value of this parameter determines the case of a single decision tree.In the random forest function, the number of variables selected by the default decision tree node is the root value of the number of variables in general. However, under normal circumstances, this default value cannot be the value with minimum error, that is, it is not the optimal value. Therefore, this value needs to be adjusted to obtain the number of variables with minimum error.

The number of decision Numbers in a random forest (ntree): this parameter determines the size of the entire random forest.The selection of the number of decision Numbers can be judged by drawing the relation graph between the number of decision trees and the error.

Urban Classification Index System
The classification of cities should be based on the index system. The index system is generally regarded as the classification basis of a certain phenomenon, which refers to the evaluation system composed of several relatively independent and interrelated statistical indicators that can reflect one or several characteristics of a phenomenon.So the most important part of grading cities is the selection of grading indicators,this paper uses the objective happiness index to divide.

In today's rapidly developing society, the economic development is not enough to judge the development level of a city, but more and more attention is paid to multiple development, so the term "happiness" is derived.There are two meanings of happiness. One is a judgment of the objective conditions and state of life, and the other is a value judgment of the subjective meaning and satisfaction of life, that is, happiness is divided into objective happiness and subjective happiness.Subjective happiness refers to the psychological experience of people's own perception of happiness, which is a subjective feeling and difficult to judge from the objective external environment. However, objective happiness can be judged based on external factors and measured by objective indicators. Therefore, the index adopted in this study is objective happiness.Therefore, the index adopted in this study is objective well-being.The factors influencing the objective happiness can be analyzed from four aspects: economy, social security, living convenience and environment.

In order to study the objective happiness of 36 major cities in China, including provincial capitals and municipalities directly under the central government, this paper uses four level-one indexes to measure: Economic status; Environment; Social security degree; Living convenience. The corresponding secondary indicators are: The selection of GDP to reflect the economic situation; Urban green coverage rate and excellent rate of air quality as environmental indicators; Use the general public budget expenditure and the number of hospitals to reflect the level of social security; The number of cinema and per capita park green space occupancy to reflect the degree of living convenience.

Establishment and Prediction of Random Forest Algorithm Model:

The processing of datasets:

This process was to read in the data set in R, obtained the basic information of the data set, so as to have a preliminary understanding of the data set and name each variable.

When constructing the random forest, about 1/3 of the samples were excluded from the training data set of each tree in the forest.These samples are called "out of bags" (OOB);Each tree has a different set of OOB samples, OOB sample is not used to construct the decision tree and the tree form an independent test sample, namely when modeling the random forest, not using all of the sample data, of which the selection of 2/3 as training set, is used to forecast model, and the other was a third of the sample as the test set, used to test the accuracy of the model.Therefore, before modeling, divided the data set into training set and test set, and looked at its basic properties.

Specific modeling operations is as follows:

```
>ind=sample(2,nrow(city),replace=TRUE,prob=c(0.7,0.3))
                                        >set.seed(100)
>train=city[ind==1,]
>test=city[ind==2,]
```

Building model

In the process of modeling, two important parameters were selected, one was the number of decision tree and the other was the number of variables selected by decision tree nodes.

Firstly, the following described the number of variables selected by the nodes of the decision tree. In general, the default variable of the function is 2, but the corresponding error is not necessarily the smallest, so in order to choose the most appropriate decision tree node number of the selected variables, the corresponding error need to do all its variables select minimum error of the corresponding variables.

Specific modeling operations and results are as follows:

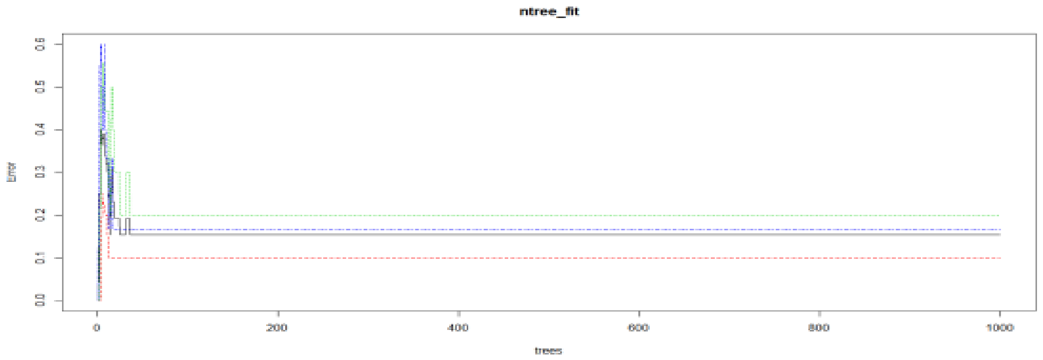
```
Process:
> n=length(names(train))
>set.seed(100)
>for(i in 1:(n-1)){
+    mtry_fit=randomForest(class~.,data=train,mtry=i)
+    err=mean(mtry_fit$err.rate)
+     print(err)
+ }
Results:
[1] 0.2231487
[1] 0.169527
[1] 0.1577842
[1] 0.1267847
[1] 0.1559124
[1] 0.1467675
[1] 0.1343926
```

As can be seen from the results, when the number of variables selected by the nodes of the decision tree is 4, the corresponding error is the smallest, that is, the variable selected by the most appropriate decision tree node is 4.

Then, the following described how to determine the number of decision trees. Specific modeling operations and results are as follows:

```
>ntree_fit=randomForest(class~.,data=train,mtry=4,ntree=1000)
>plot(ntree_fit)
```

Figure 1:- Relation of decision tree and error



According to the figure, when the number of decision trees is greater than 100, the model error tends to be stable, so the number of decision trees is selected as 100.

Finally, the following introduced how to establishing random forest model.The R language code is shown as follows:

```
>ntree_fit=randomForest(class~.,data=train,mtry=4,,ntree=1000)
>set.seed(100)
>rf=randomForest(class~.,data=train,mtry=4,ntree=100,importance=TRUE)
>rf
```

Results Analysis:-

The results obtained by the constructed model is as follows:

```
Call:
randomForest(formula = class ~ ., data = train, mtry = 4, ntree = 100, importance = TRUE)
              Type of random forest: classification
                    Number of trees: 100
No. of variables tried at each split: 4
OOB estimate of error rate: 7.69%
Confusion matrix:
first     second    third     class.error
first        12        1      0      0.07692308
second        0        8      0       0.00000000
third         0        1      4      0.20000000
```

1. The result “Type of random forest” shows that the model is the “classification” model.
2. The result “Number of trees” shows that the model contains 100 decision trees.
3. The result “No. of variables tried at each split” shows that the number of variables selected at each decision tree node is 4.

The result “OOB estimate of error rate” shows that the total prediction error of the model is 7.69%.

The results “Confusion matrix” shows the difference between the prediction of the final model and the actual results of the training set.It can be seen that the final model correctly predicted the samples of 12 training sets in category first, incorrectly predicted one of the samples as second, with a misjudgment rate of7.69%. The final model correctly predicted the samples of 8 training sets in category second, with a misjudgment rate of 0.00%. The final modelcorrectly predicted the samples of 4 training sets in category third, and wrongly predicted one of them as second, with a false prediction rate of 20%.

Finally verify and predict, in order to make the results more credible, using the test set sample data validation model that was not used for modeling.

```
> pred1=predict(rf,data=test)
> Freq1=table(pred1,test$class)
>sum(diag(Freq1))/sum(Freq1)
[1] 0.9230769
```

According to the above results, the accuracy of the prediction model can reach 92.31%.

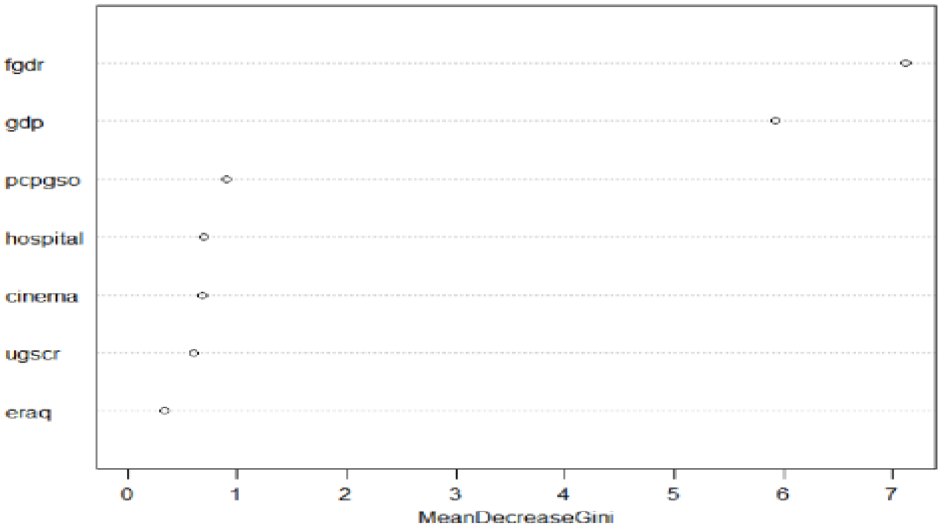
The importance of the independent variable:

The random forest model is different from the ordinary linear model. In the discriminant model, it is usually impossible to compare the importance of the variables between the models, and the variables cannot be significantly tested after the model is established. However, in the random forest model, the importance of the respective variables for the discriminant effect of the model can be calculated by the importance () function. The specific R language code and result are as follows:

```
Process：
>importance=importance(x=rf)
>importance
>set.seed(100)
>varImpPlot(rf)
Result:
MeanDecreaseGini
gdp 5.9271112
fgdr          7.1124700
hospital        0.6927750
cinema          0.6825423
pcpgso          0.9014552
ugscr          0.5933083
eraq           0.3272612
```

The results list the importance values of the corresponding independent variables calculated by all the independent variables under the Gini coefficient measurement standard. In the above results, the higher index values corresponding to the independent variables indicate that the independent variables have a greater influence on the discrimination of the model. That is, the importance of the independent variables are: local general public budget expenditure, GDP,per capita park space occupancy, number of hospital, number of cinema, urban green coverage rate, and excellent rate of air quality

Figure 2:-The importance of the independent variable



Conclusion:-

In the past, when grading cities, the grading indicators used were mostly traditional economic indicators, and the analysis methods used were mostly regression analysis or cluster analysis. The objective happiness index used in this paper is a relatively new grading index. The established random forest model also proved that the random forest algorithm can obtain a high-precision classification result. The advantage of random forests differs from other algorithms is that they can effectively reduce the amount of calculation without reducing the accuracy, and in the random forest model, the importance of the respective variables for the discriminant effect of the model can be calculated.

In this paper, the establishment of random forest algorithm model was introduced. Firstly, the data was analyzed and processed, and then the random forest modeling process was carried out. In the modeling process, two parameters were selected, one was the number of decision trees, and the other was the number of variables selected in the decision tree node. After the parameters were properly set, an accuracy result was obtained, and the feasibility of the model was obtained from the results. The total prediction error of the final model obtained is 7.69%. The accuracy of the prediction model obtained by using the test set can reach 92.31%. Therefore, the established random forest algorithm model is highly feasible.