Application of Decision Tree Algorithm Based on Clustering and Entropy Method Level Division for Regional Economic Index Selection

Yi Zhang(✉) and Gang Yang

Department of Information and Software Engineering,
Chengdu Neusoft University,
Dujianyan, Chengdu 611844, Sichuan, China
zy1044911641@gmail.com

Abstract. The economy of a region is affected by many factors. The purpose of this study is to use the entropy method clustering and decision tree model fusion to find the main factors affecting the regional economy with the support of big data and empirical evidence. First extract some important indicators that affect the regional economy, and use the entropy method to find the relative weights and scores of these indicators. Then use K-means to divide these indicators into several intervals. Based on the entropy fusion model, obtain the ranking of each category of indicators, use these rankings as the objective value of the decision tree, and finally establish an economic indicator screening model. Participate in optimization and build a decision tree model that affects regional economic indicators. Through the visualization of the tree and the analysis of feature importance, you can intuitively see the main indicators that affect the regional economy, thereby achieving the research goals.

Keywords: Clustering · Entropy evaluation method · Decision tree · Regional enconomy · Random forest

1 Introduction

The development of a regional economy is related to many factors, which will positively or negatively affect the development of the regional economy. The entropy method is an objective weighting method. The entropy value is used to judge the degree of discreteness of an indicator The greater the degree, the greater the impact (weight) of the indicator on the comprehensive evaluation, and the smaller its entropy value. In the improved entropy method and its application in the evaluation of economic benefits [4], the power factor method mentioned in the article solves some extreme data and indicators but essentially uses the entropy method to evaluate regional economic benefits. Objective It is relatively strong, and lacks certain persuasion for larger data samples and more

© Springer Nature Singapore Pte Ltd. 2020
Y. Tan et al. (Eds.): DMBD 2020, CCIS 1234, pp. 45–56, 2020.
https://doi.org/10.1007/978-981-15-7205-0_5

economic indicators. At the same time, the basic use of the CART decision tree was proposed in [5], but the application of index screening in the evaluation of the regional economy was still lacking. In this paper, we propose the concept of model fusion that combines CART decision tree and clustering machine learning algorithms for regional economic level evaluation and division based on the use of entropy method, which can more accurately extract the indicators that affect economic effects. As described below.

The entropy method objectively calculates the corresponding weight score based on the degree of discreteness of each feature data, and the distribution of the data is basically discrete, and the data is huge. Here we choose the K-means clustering algorithm based on the division to classify, on large data sets The calculation efficiency is also very high, and then the weight score is used to calculate the average score of each category, and the classification is divided. Because most of the data types are continuous, the CART decision tree is selected as the classification model. However, on the huge data, although the corresponding parameter adjustment and optimization operations are performed, it is inevitable that there will be overfitting and generalization capabilities. Worse. The optimization problem uses the random forest in ensemble learning for optimization, which can not only effectively run on big data but also solve high-dimensional data problems without reducing the dimension, and the estimated model is an unbiased model. To this end, we use several types of algorithm model fusion and use k-fold cross-validation to evaluate and tune the model. Our experiments on the economic data set opened by the Singapore government show that the accuracy rate of the final optimized model fusion is as high as 94%, Analyzed the main indicators affecting regional economic development, and put forward some suggestions to help regional economic sustainable development.

The rest of this paper is organized as follows: Sect. 2 discusses the main research methods, In Sect. 3 discussed the establishment of models, how to carry out model fusion, Corresponding experiments are carried out in Sect. 4, and the cross-validation results are given at the end. Finally, the work is concluded in Sect. 5.

2 Research Methods

2.1 Introduction to K-Means Clustering Algorithm

The k-means clustering algorithm is an iterative solution based partitioning cluster analysis algorithm. It uses distance as a standard for measuring the similarity between data objects. Euclidean distance is usually used to calculate the distance between data objects. The formula for calculating the Euclidean distance is given below 1:

$$\operatorname{dist}\left(x_i, x_j\right)=\sqrt{\sum_{d=1}^{D}\left(x_{i,d}-x_{j,d}\right)^2} \tag{1}$$

Complete k-means algorithm flow, such as Algorithm 1

Algorithm 1: K-means clustering algorithm

```
Seeding: $k$ initial centers $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$
repeat
    foreach $i \in \{1, \ldots, k\}$ do
        $C_i \leftarrow \{x \in \mathcal{X} \mid \|x - c_i\| < \|x - c_j\| \,\forall j \neq i\}$ ;    / * $C_i$ is assigned
        the set of all points in $\mathcal{X}$ having $c_i$ as their closest center */
    end
    foreach $i \in \{1, \ldots, k\}$ do
        $c_i \leftarrow \frac{1}{|C_i|} \sum_{x \in C_i} x$;    /* modify $c_i$ to be the center of mass of $C_i$
        */
    end
until no more change of $C$;
return $\mathcal{C}$;
```

2.2 An Introduction to the Entropy Method

The entropy method refers to a mathematical method used to judge the degree of dispersion of a certain index. The greater the degree of dispersion, the greater the impact of this indicator on comprehensive evaluation. You can use the entropy value to judge the degree of dispersion of a certain index because, the information entropy can be used to calculate the weight of each indicator according to the degree of variation of each indicator, which provides a basis for comprehensive evaluation of multiple indicators.

Information entropy refers to the concept of entropy in thermodynamics and describes the average information of the source, as shown in formula 2

$$H(x) = E\left[\log \frac{1}{p\left(a_i\right)}\right] = -\sum_{i=1}^{q} P\left(a_i\right) \log P\left(a_i\right) \tag{2}$$

2.3 Overview of Decision Tree Generation Algorithms

The entire decision tree is based on the tree structure to make decisions, which can be divided into three progressive processes: optimal feature selection, decision tree generation, and pruning. The internal node corresponds to a test on an attribute, each branch corresponds to a possible result of the test, that is, a certain value of the attribute, and each leaf node corresponds to a prediction result. The main information gain calculation method is as follows:

$$IG(S|T) = \text{Entropy}(S) - \sum_{value(T)} \frac{(S_v)}{S} \text{Entropy}\left(S_v\right) \tag{3}$$

The term before the minus sign in the formula is the entropy of the training set classification, S represents the sample set, T is the set of all feature values, and Sv is the feature equal to v in the feature; the second half of the minus sign is the entropy for classification with v. The subtraction of the two entropies has the following meaning: using this feature classification, it can reduce how much uncertainty and how much information is carried.

2.4 Random Forest Algorithm Based on CART Tree

CART can be used for both regression analysis and classification analysis, and some integrated algorithms based on CART have been extended. To solve the problem of large data size and data volume in the context of big data, this study chose CART as the Basic random forest algorithm.

The CART decision [3] tree has the advantages of being easy to understand and having certain non-linear classification capabilities, but a single decision tree has some disadvantages.

The above-mentioned defects can be improved by the random forest integration method in integrated learning bagging. Randomforest is composed of many decision trees, and there is no correlation between different decision trees. First, randomly sample the samples, train the decision tree, and then classify the nodes according to the corresponding attributes until they can no longer split the position, and build a large number of decision trees to form a forest.

3 Establishing Model

Before using the decision tree algorithm to build a tree of regional economic indicators, it is necessary to determine the target value level of the decision tree, which is also the focus of this model. First, use the entropy method to calculate the corresponding weights for the indicators of the regional economy. Select n indicators and m periods So X_{ij} represents the y-th value of the i-th index $(i = 1, 2, ..., n; j = 1, 2, ..., m)$.

3.1 Indicator Normalization

The homogeneous indicators are homogeneous. Because the measurement units of the indicators are not uniform before they are used to calculate the comprehensive indicators, they must be standardized, that is, the absolute values of the indicators are converted into relative values, and $X_{ij} = |X_{ij}|$, so as to solve the problem of homogeneity of various qualitative index values. The specific method is as follows:

$$x'_{ij} = \frac{x_{ij} - \min\{x_{1j}, \ldots, x_{nj}\}}{\max\{x_{1j}, \ldots, x_{nj}\} - \min\{x_{1j}, \ldots, x_{nj}\}} \tag{4}$$

Then X'_{ij} is the value of the j time period of the i index. For convenience, the normalized data are recorded as X_{ij}.

3.2 Calculate the Weight of Each Indicator

$$p_{ij} = \frac{x_{ij}}{\sum_{n=1}^{n} x_{ij}}, i = 1, \ldots, n, j = 1, \ldots, n \tag{5}$$

$$e_j = -k \sum_{i=1}^{n} p_{ij} \ln\left(p_{ij}\right), j = 1, \ldots, m \tag{6}$$

Calculate weights for various years:

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, j = 1, 2, \ldots, m \tag{7}$$

Calculate the comprehensive score obtained by each indicator:

$$s_i = \sum_{j=1}^{m} w_j \cdot p_{ij}, i = 1, \ldots, n \tag{8}$$

In this way, the corresponding weights and scores of each index can be obtained. This score is convenient for later determination of the characteristic value of the decision tree.

3.3 Clustering for Rank

This article clusters the indicators that affect the economy. According to the "Elbow Rule" and the actual economic stage in Singapore history, the categories are finally classified into 4 categories, and then the average score of each category is calculated. It is obtained by adding the scores obtained by the entropy method of each type of index and averaging. These four scores are then ranked and divided into four intervals, and then each sample is compared to which interval the corresponding score is calculated according to the weight. Finally, each sample can obtain a corresponding rank. The rank is the target value to be evaluated by the decision tree.

3.4 Decision Tree Selection of Economic Indicators

(1) For each feature A and all possible values a, divide the data set into two subsets A = a and A! = A, and calculate the Gini index of set D:

$$\text{Gini}(D, A) = \frac{D_1}{D}\text{Gini}\left(D_1\right) + \frac{D_2}{D}\text{Gini}\left(D_2\right) \tag{9}$$

(2) Iterate through all the features A, calculate all the Gini indexes of its possible values a, select the feature corresponding to the minimum Gini index of D and the cut point as the optimal division, and divide the data into two subsets.

(3) Recursively call steps (1) (2) on the above two child nodes until the stop condition is satisfied.

4 Experiment

Before the experiment, we downloaded and organized the data from the Singapore government statistics website https://www.singstat.gov.sg/.

4.1 Data Source Analysis

The data preparation stage includes data acquisition and data preprocessing, which is the foundation of data mining. In order to make the analysis results real and effective, the data comes from the official website of the relevant region. Due to the variety of data obtained, we need to select the data appropriately according to the research goals set in advance. With reference to the existing research results, we analyzed more than 20 relevant indicators (GDP), government operating income, employment opportunities, private consumption expenditures, and output investment that affect economic development. Following the principles of scientificity, representativeness, availability, and operability of selected indicators, this article takes Singapore as an example to select 19 representative indicators, and obtains data for each quarter from 1975 to 2017 from the department of statistics of Singapore Make up the data set. The relevant indicators are shown in Table 1 below.

Table 1. The indicator system

GDP	Government spending
Aggregate demand index	Gross Fixed Capital Formation
Employee wages	Total operating surplus
Tax cuts subsidies	Residential price index
The birth rate	The company number
Gas Sales	Electricity Generation
Retail consumption index	Inbound Tourism numbers
Tourism Receipts	Air Cargo Loaded
Motor Vehicle Population	Vessel Arrivals number
Vessel Total Cargo (Thousand Tonnes)	

4.2 Processing Data Missing Values

From Singapore for the presence of some data of 19 indicators data missing value, according to the usual data before operation is to take the average and median to fill the missing value, but to fill the lack of scientific data and accuracy, so in this paper, the way is through the correlation between indicators and indicators to fill the missing value, through index correlation analysis between the first, find a missing value indicators and other indicators of relevance, through strong correlation to fill in the missing data. The correlation analysis of these 19 indicators is shown in Fig. 1 below.

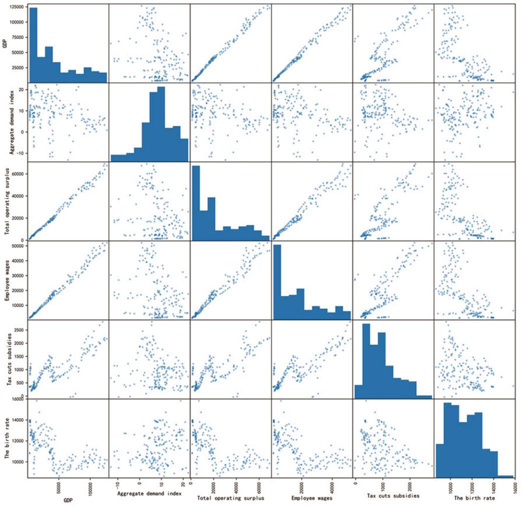

Fig. 1. Correlation scatter plots between indicators

The missing values in the 19 indicators are Total operating surplus, Employee wages, Tax cuts subsidies, The birth rate It can be seen from Fig. 1 that the linear correlation between features is very high. You can use GDP to predict Total operating surplus, use GDP to predict Employee wages, use Government spending and Total Merchandise Trade, and The company number to predict The birth rate, See Table 2 in order.

Table 2. Prediction equation

f(x)
$f(x) = 0.5284713 * x - 597.95273615$
$f(x) = 0.0365264104 * x + 289.10634$
$f(x) = 2.2291718.925 * 10^{-5} * x^2 - 1.6824388 * x + 15161.526483621026$

4.3 Ranking by Clustering and Weighting

The following is the weight of each index obtained by the entropy method, as shown in Table 3 below.

Table 3. Weight of each indicator

Index	Weight	Index	Weight
GDP	0.091025	Government spending	0.093909
Aggregate demand index	0.013786	Gross Fixed Capital Formation	0.076256
Employee wages	0.091720	Total operating surplus	0.090709
Tax cuts subsidies	0.036622	Residential price index	0.060454
The birth rate	0.035269	The company number	0.043344
Gas Sales	0.042019	Electricity Generation	0.063323
Retail consumption index	0.044275	Inbound Tourism numbers	0.069598
Tourism Receipts	0.046299	Air Cargo Loaded	0.045678
Motor Vehicle Population	0.046014	Vessel Arrivals number	0.009698

Since the data distribution is partitioned, we use the k-means algorithm to cluster the economic data sets accordingly. Before clustering, we first use the PCA dimensionality reduction algorithm to reduce the data set to 3 dimensions because there are many features and the dimensions are inconsistent, which affects the clustering too much. As shown in Fig. 2.

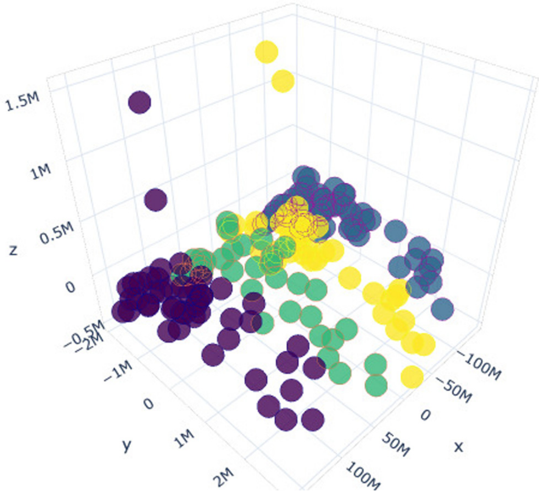

Fig. 2. Clustering graph

Based on the above clustering effects, we divide the original indicators into four broad categories. Then calculate the average for each category. The data for each category are shown in Table 4 below.

Table 4. Composite scores for four types of samples

Level	Interval score
First	$307162.4 \leq x \leq 439842.85$
Second	$222486.69 \leq x \leq 307162.4$
Third	$116022.01 \leq x \leq 222486.69$
Fourth	$x \leq 116022.01$

From the Table 4 above, we can know the score interval corresponding to each category. We use the weight score and each row of data to calculate the score and determine the level.

4.4 Generate Decision Tree

Because of the type of data in the dataset, we use the CART algorithm. The Gini index is calculated as follows:

$$GINI(D) = \sum_{i=1}^{k} p_k \cdot (1 - p_k) = 1 - \sum_{i=1}^{k} p_k^2 \tag{10}$$

The decision tree is constructed without pruning. The decision tree is mainly constructed using python machine learning third-party library sklearn. The accuracy rate obtained by the test is 84%. In order to make the model more accurate and the error smaller, a random forest is used to optimize the decision tree. The CART decision tree visualization is shown in Fig. 3.

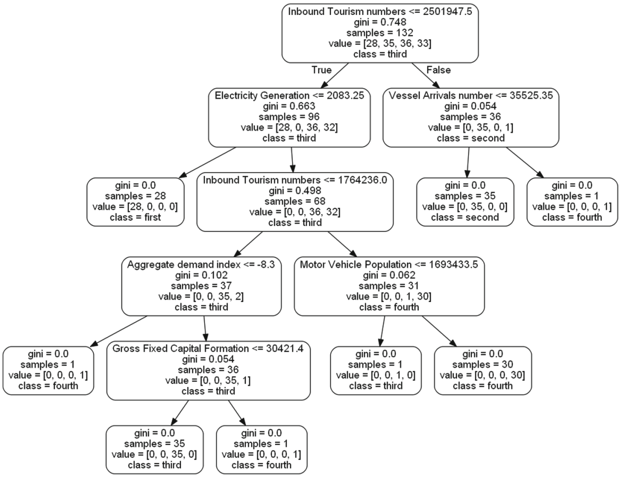

Fig. 3. Economic indicator decision tree

[2] here we use grid search k-fold cross-validation to adjust the parameters of the random forest parameters, and it is concluded that the effect is best when the estimator is 50. After getting the number of estimators, To further improve the accuracy, due to the limited data provided by the government, the dimension cannot reach several hundred dimensions. Here we mainly discuss the use of cross-validation to obtain the optimal *min_impurity_split* size for pruning to prevent excessive growth from causing poor generalization ability. The verification curve results are shown in Fig. 4:

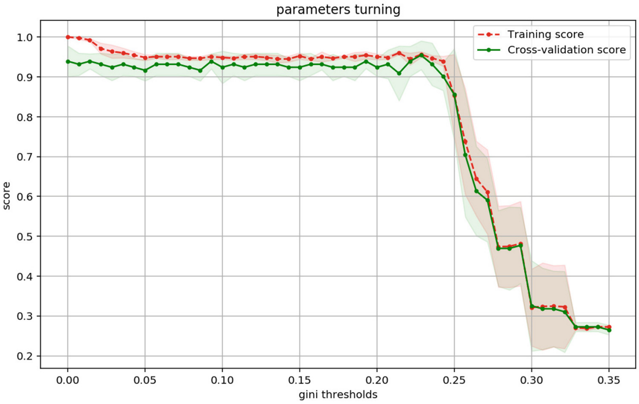

Fig. 4. Cross validation results

It can be seen from Fig. 4 that when the impurity index is 0.22857142857. Finally, the accuracy of the training set and the test set was 94.7%.

After the random forest model is obtained, each decision tree in the random forest is judged separately when a new sample is entered. The bagging set strategy is relatively simple. For classification problems, the voting method is usually used, and the most votes category or one of the categories is the final model output. The time complexity is $O(M(mnlogn))$. Some features need to be selected randomly during the calculation process, and additional time is required to process this process, so it may take more time. Where n represents n samples, m represents m features, and M represents the number of decision trees participating in the voting.

4.5 Model Importance Interpretation

This paper uses the common algorithm xgboost in boosting, and compares and verifies the feature importance obtained from the random forest algorithm in bagging. Since the GBDT algorithm only has a regression tree, it will not be discussed here. This adjusted random forest model consists of 50 lessons of decision trees. Each tree can get an impurity measure about each feature, and then the scores can be added according to the feature to get the relevant feature importance [1]. At the same time, we use the xgboost algorithm, which is also composed of 50 decision trees, to compare. After adjusting the parameters, the optimal subsample is 0.5204081. The best learning_rate is 0.3000012, the import_type is modified to weight, and the objective is modified to multisoftprob. After completion, we can get the feature importance corresponding to the two algorithms, see Fig. 5 below.

As shown in Fig. 5, assuming that the sum of the importance of all features is 1, it can be seen that in two well-known algorithms, the importance of Inbound Tourism numbers is the largest, indicating that the largest factor affecting the regional economy is Inbound Tourism numbers during the entire classification This indicator is followed by GDP. Among them, aggregate demand index and tax cuts subsidies and Air Cargo Loaded and other corresponding features account for a small proportion, indicating that in the process of economic development, the impact of these factors is small. We can draw from this model that Singapore can vigorously develop the tourism industry and prompt the corresponding GDP, which needs to be strengthened on the characteristics of lower scores.

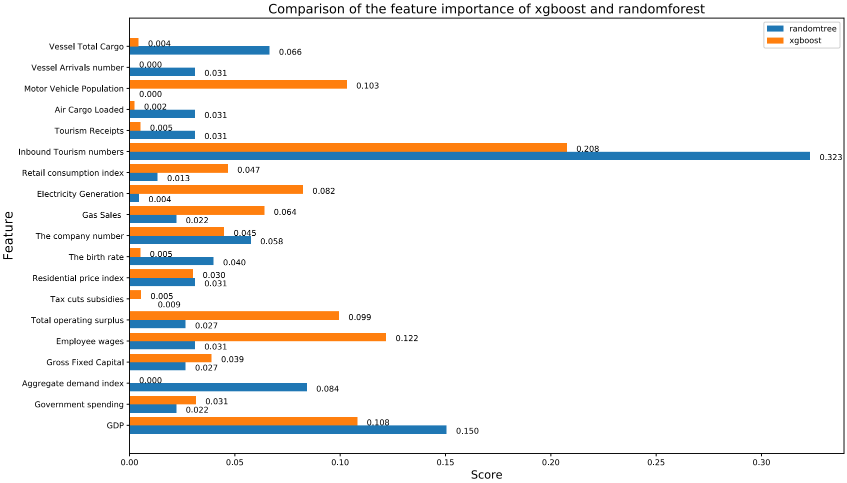

Fig. 5. Compare results

5 Conclusion

The weight of each index is determined by the entropy method, and K-means is used to cluster and divide the levels, and the target value is determined for each index. An economic index screening model based on the CART tree random forest algorithm. Through this model, it is possible to scientifically screen out those important indicators affecting the regional economy. The accuracy of the fusion of the above models is as high as 94%, and there will not be too much error when analyzing the factors affecting the regional economy. This method helps regional economic decision-makers to provide accurate positioning by providing decision support for regional economic development. The analysis of the above results shows that the Singapore government can vigorously develop the tourism industry because the importance of inbound tourism numbers is very high and the economic level development can be completed accurately.

Based on the research conclusions of this paper, the real-time analysis of larger data can be carried out based on the study of machine learning technology to screen the main factors affecting the regional economy, and an economic impact factor analysis system based on each region or country can be established. If the data supports, the number of indicators can be larger, so that the main factors affecting the regional economy can be analyzed more comprehensively and accurately.

References

1. Li, X.: Research on application of decision tree integration method in precision traffic safety publicity. In: Proceedings of the 13th China Intelligent Transportation Annual Conference, pp. 562–571. China Intelligent Transportation Association (2018)
2. Ying, Q.: Research and application of educational data mining based on decision tree technology. Zhejiang Normal University, Zhejiang (2018)
3. Lin, Z., Wang, S., Qiu, Z., Lulu, Y., Nan, M.: Application of decision tree algorithm to forecasting stock price trends. In: The 15th Network New Technology and Application Year of China Computer User Association Network Application Branch. 2011 Conference Proceedings of the 15th Annual Network New Technology and Application Conference of the Network Application Branch of China Computer Users Association in 2011, pp. 129–131. China Computer Users Association, Beijing (2011)
4. Guo, X.: Improved entropy method and its application in economic benefit evaluation. Syst. Eng. Theory Pract. (12), 99–103 (1998)
5. Chen, H., Xia, D.: Application research of data mining algorithm based on CART decision tree. Coal Technol. **30**(10), 164–166 (2011)