Research on Chinese Public Policy Decision Model Based on Decision Tree Algorithm

Shenghao Fang

Law & Social Sciences Building, University of Nottingham, NG7 2RD, Nottingham, United Kingdom

ShenghaoFang0212@163.com

Abstract. In this paper, the decision tree classification algorithm is used to establish the prediction model of Chinese public policy decision. This paper selects characteristic attributes based on the principle of policy making. Then, taking the result of public policy decision as the target label, this paper optimizes the model by adjusting the maximum depth of decision tree, the minimum number of leaf samples and the decision threshold. The test set verifies that the optimized decision tree model has a good predictive effect on the result prediction of the public policy decision model. The value of AUC was 0.848 and the model had strong generalization ability. The AUC difference between training set and test set is less than 0.04.

Keywords: decision tree algorithm; chinese public policy; decision model; machine learning.

1. Introduction

The implementation effect of public policy reflects the capability and level of modern national governance, so exploring the factors and internal mechanism that affect the policy effect is one of the important theoretical propositions of the modernization of national governance. Generally speaking, the realization of governance effectiveness in countries with super-large governance scope is highly dependent on the interaction in policy implementation. For example, scholars have pointed out that the large governance scope and governance level make it difficult for public decision-making to overcome the failure of information collection and information transmission, and the interaction and consultation between implementer and target group and the "re-decision" in the implementation are effective guarantees for the implementation of policies in big countries [1]. But the interaction in implementation makes explaining and predicting policy effects more complicated. Moreover, the implementation of consultation cannot always bring better policy effects, and there are still a variety of implementation failures in reality. A model is an abstraction and simulation of a prototype. It is a kind of idea system constructed by the cognition subject according to the similarity principle for the purpose of certain cognition, in order to represent the real system as the object of study, namely the actual existence of things. Models (mainly theoretical models) are widely used in public policy analysis. These models reflect different perspectives of public policy thinking and provide a variety of ways to understand and analyze public policy. After our entry into the WTO, both the government and the society will be in a complex and changeable external environment. Public policy is playing a more and more important role as one of the important means for the government to regulate the interest relationship between the members of the society, realize the rational distribution of public interests and concrete administrative goals. In the period of social economic transition, the public policy should choose the proper decision-making model so as to give full play to its guidance, control and distribution function. In this paper, the decision tree classification algorithm is used to build the prediction model of Chinese public policy decision.

2. Two-level decision tree prediction model

Decision tree is a kind of technology to classify the sample set based on the characteristics of attributes. In the process of classification, all attributes need to be traversed and the optimal split threshold should be determined according to the value of attributes [2]. Different decision tree methods have different split modes and optimal split point judgment indexes. In the process of phased

modeling, selecting the appropriate tree building method according to the attribute characteristics is helpful to improve the overall performance of the model. C4.5 and CART are widely used decision tree methods at present. In C4.5 method, the information gain rate is used to select the split points of attributes. In the tree building process, all attribute values are traversed and the information entropy generated by the partition set is calculated, which can better analyze the relationship between nominal attributes. However, the regression fitting degree of continuous numerical data is not high, so it is more suitable for data classification. CART method uses GINI coefficient as splitting standard and adopts binary recursive segmentation to generate binary tree, which is not easy to generate data fragments. It has the characteristics of simple structure, high efficiency and strong scalability, and can handle highly slanted numerical data. The prediction accuracy of numerical data is higher than that of multi-fork tree, but binary division method is adopted in tree building process. In classification problem, the relationship between nominal attributes is not enough, so it is more suitable to establish regression model of numerical data. Combining the advantages of these two decision tree methods, this paper designs a two-stage tree building method called two-level decision tree. Figure 1 shows the structure diagram of the TDT model. Specific modeling ideas are as follows:

(1) All attributes in the sample are divided into two attribute sets according to category: nominal type attribute set, represented by M; A set of numeric attributes, represented by M*. (2) Based on the principle of C4.5 method, all attributes in M are traversed to construct a classification tree. When the sample quantity or purity of a node reaches the lower limit, this node is taken as the leaf node in this stage, and P is used to represent the set of leaf nodes. (3) All sample subsets in P are traversed, and all attributes in M* are split by CART method. When the number of samples or purity of nodes reaches the lower limit, the split is stopped, and the sample set contained by the current node is fitted by support vector machine.

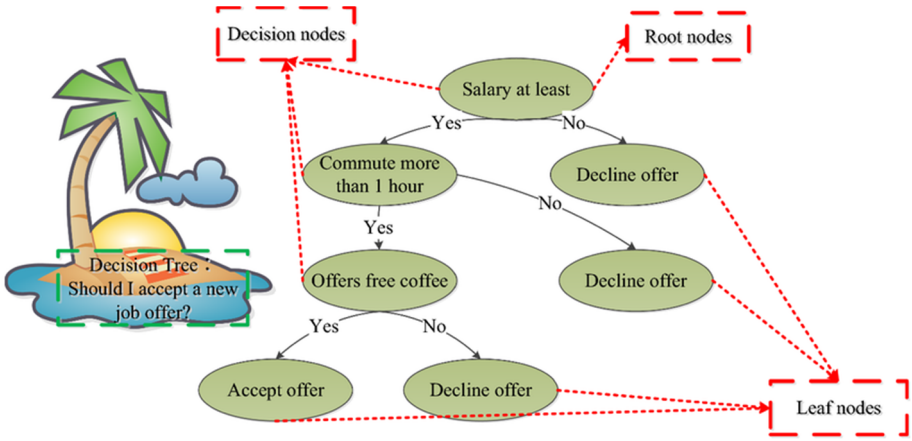

Figure 1. Structure diagram of two-level decision tree model

In order to reduce redundant branching of TDT tree structure and avoid model overfitting, a post-pruning strategy is used to guide the two-stage tree building process. Post-pruning is to traverse non-leaf nodes from bottom to top after the entire decision tree is constructed. If replacing the subtree of this node with a leaf node can improve the prediction accuracy of the decision tree, the subtree will be replaced with a leaf node. The construction process of TDT is divided into two stages [3]. After the end of the first stage, the tree structure is simplified by the post-pruning strategy, and on this basis, the tree is constructed in the second stage. After the whole tree is constructed, the post-pruning operation is carried out again on the non-leaf nodes generated in the second stage.

(1) Conditions for pruning in the first stage: if the node point tree can be replaced by the leaf node to improve the overall information entropy, then pruning will be performed; (2) Conditions for the second stage of pruning: the node point tree is replaced with the leaf node, and a new regression model is established based on this leaf node. If the prediction accuracy of the new regression model is improved, pruning will be performed.

2.1 Classification tree construction based on improved C4.5

Information gain rate is the basis for C4.5 to divide data sets. During the splitting process, the current splitting attribute is determined according to the information gain rate. Given sample set S, it can be divided into class K according to class attributes. The proportion of class k samples in set S is $p_k(k=1,2,\ldots,K)$, then the information entropy of S can be expressed as:

$$H(S)=-\sum_{k\in K}(p_k\times\log_2 p_k) \tag{1}$$

Suppose E represents all the attributes to be split, and attribute $A\in E, A$ has n possible values $\{A_1,A_2,\ldots,A_n\}$. If the data set S is divided into n subset $S_1,S_2,\ldots,S_j,\ldots,S_n$ according to the value of A, where S_j represents all the samples in the data set S whose value is A_j on attribute A, $p(A_j)$ represents the probability that attribute A is A_j, $p(A_j)=|S_j|/|S|$, The information gain after selecting attribute A to classify data set FS is:

$$Gain(S,A)=H(S)-\sum_{j\in n}(p(A_j)\times H(S_j)) \tag{2}$$

$H(A_j)$ represents the information entropy of data set S_j. The information gain rate of attribute A is:

$$GainRatio=Gain(S,A)/SplitI(A) \tag{3}$$

$$SplitI(A)=-\sum_{j=1}^{n}(p(A_j)\times\log_2 p(A_j)) \tag{4}$$

The C4.5 algorithm based on information gain rate classification can overcome the multi-value bias problem, and carry out traversal of all attributes and their attribute values in the classification process, so as to comprehensively explore the relationship between attributes. Considering only the information gain rate of a single attribute will lead to low prediction accuracy [4]. In order to make up for the shortcomings of the algorithm, the following two improvements were made in the process of generating classification trees: In order to take the dependence between attribute *A* and other non-class attributes to be split into consideration in the selection process of split attributes, *F* was assumed to represent the set of non-class attributes except attribute *A*, and the information gain and the mean value of information gain between attribute *A* and other attributes were expressed as

$$Gain(A_F)=\sum_{f\in F}(H(A)-H(A|f) \tag{5}$$

$$Gain(A_F)=Gain(A_F)/|F| \tag{6}$$

Where f represents any attribute in F, $H(A)=\sum_{j\in n}p(A_j)H(S_j), H(A|f)$ represents the conditional entropy of attribute A under the condition that the value of attribute f is known. If f is $\{1,\ldots,m,\ldots,M\}, f_m$, it represents the f value of attribute f, and the value of f_m can divide the data set S into m subset $Q_1,Q_2,\ldots,Q_m,\ldots,Q_M, p(f_m)=|Q_m|/|S|$. Similarly, if A has n possible values $\{A_1,A_2,\ldots,A_n\}$, the data set Q_m is divided into n subset $S_1^m,S_2^m,\ldots,S_j^m,\ldots,S_n^m$ according to the value of A, where S_j^m represents all the samples in the data set Q_m whose value is A_j on attribute A, and $p(A_j|f_m)=|S_j^m|/|Q_m|$ is then:

$$H(A|f)=\sum_{m\in M}p(f_m)\times\sum_{j\in n}(p(A_j|f_m)\times H(S_j^m)) \tag{7}$$

The difference between the entropy of attribute *A* and the conditional entropy of known attribute *f* is called information gain. The calculation formula is shown in Equation (8):

$$Gain(A_F) = \sum_{f \in F} (H(A) - H(A \mid f)) \tag{8}$$

Formula (8) is used to improve the information gain rate of attribute A. The calculation formula is shown in Formula (9):

$$GainRatio = Gain(A) / \left(SplitI(A) + \overline{Gain(A_F)} \right) \tag{9}$$

2.2 Regression model construction based on improved CART

Binary segmentation method is introduced in CART to build trees, which can process continuous data and is suitable for modeling complex data with multi-characteristic variables. CART method uses GINI coefficient as the selection criterion for split attributes, and the purity of data set S is measured by GINI coefficient:

$$Gini(S) = \sum_{k \in K} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k \in K} p_k^{\,2} \tag{10}$$

The GINI index of attribute A can be defined as:

From the candidate attribute set E, the attribute with the minimum GINI index after partitioning is selected as the optimal attribute during each splitting:

$$A^* = \arg\min_{A \in E} Gini\_index(S, A) \tag{11}$$

When there are many continuous attributes and many decision tree nodes, the calculation amount of the optimal split point selection process will become very large, so the Fayyad boundary point decision theorem is introduced to improve the segmentation threshold selection process.

3. Political flow in the Internet era

Independent of the problem flow and policy flow is the political flow, which is composed of the change of public opinion, the adjustment of the key personnel of local government, the division of functions and powers among various departments, the competition among various interest groups and other factors [5]. The essence of public policy agenda setting is the process of coordination, competition, game and compromise among multiple interest subjects around specific social issues in order to maximize their own interests. The policy agenda setting in the network era involves many subjects: government decision-makers, institutional media, and relevant interest groups. Figure 2 shows the multi-source flow model in the Internet era.

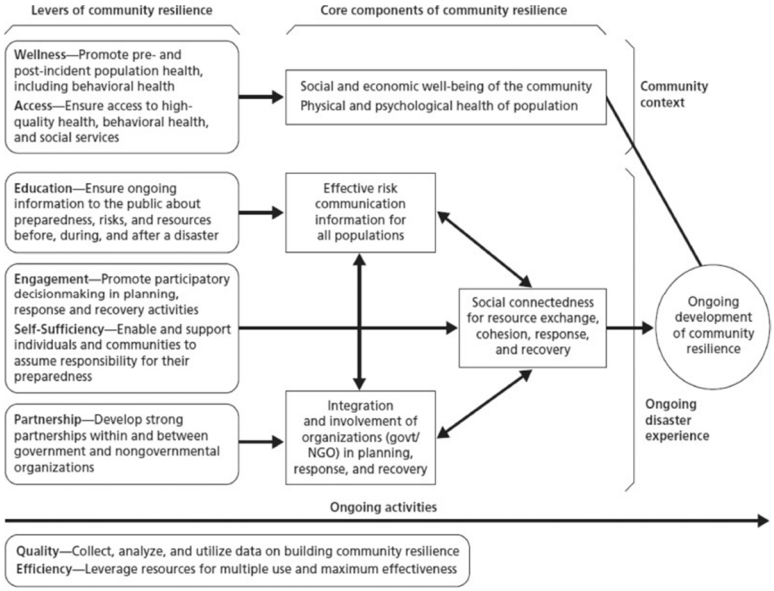

Figure 2. Multi-source flow model in the Internet era

3.1 Government decision-makers

By guiding, adjusting, combining and bonding the fragmented network will, government decision-makers will remove the false and seek the truth, remove the coarse and extract the essence from the diversified policy appeals, and judge whether the public issues will enter the policy issues through the definition of problems, the selection and demonstration of alternative schemes, and the demarcation of management authority [6]. With the development of The Times, strategic issues such as "common prosperity", "sustainable development concept", "building a harmonious socialist society", "letting the people share the fruits of reform", "ecological civilization construction" and "inclusive growth" have been put forward, which means that the social and economic issues entering the agenda are more and more prominent with the public characteristics of "people-oriented". This is a policy agenda-setting approach with Chinese characteristics.

3.2 Organization Media

When there is a deviation in the direction of public opinion, the mainstream media plays the role of information filtering and strengthening. By timely and in-depth reporting of authoritative information, the mainstream media can correct the deviation of public opinion and put the real public issues in the eye-catching place. As Bernard puts it, "It's hard for the media to tell the audience how to think, but it's easy for it to control what the audience thinks."

3.3 Policy window opens

The discourse on the Internet presents the characteristics of fragmentation, disorder and scattered. However, once the focus event occurs, the media's report on the event and the netizens' discussion on the Internet will make these discourses meet with each other on the online platform, and jointly promote their evolution to policy issues and the coupling of problem flow, policy flow and political flow. Nowadays, Chinese society is in a transition period, and the reform has entered a deep-water zone and a period of overcoming difficulties [7]. Traditional values and new values are colliding with each other, and negative emotions such as confusion and loss, impetuous fear, hatred of wealth and worship of money are pervasive. Although these negative emotions are not the mainstream emotions of social development, they have had a certain negative impact on the reform process of China, so the relevant departments should face up to the appeals of these interest groups. Public agenda setting is actually a game process between multiple interest subjects and multiple values, which should ultimately point to the pursuit, realization and maintenance of public interests. With the advent of the era of Internet "we media", netizens express their appeals and pay attention to the generation of policy schemes through the Internet, and the government absorbs public opinions timely through the Internet, interacts positively and resolves conflicts. Such bottom-up network agenda and top-down national will are integrated in the interaction of information flow, and finally transform social problems into policy issues.

4. Decision simulation

In this study, the same training set in the data samples was used to train the decision tree classifier model, and the maximum depth of the decision tree model was set as three and four layers respectively. The model evaluation level was investigated under the condition that the minimum number of leaf samples ranged from 10 to 220. The influence of the minimum number of leaf samples on the AUC value of the prediction model was shown in Figure 3.

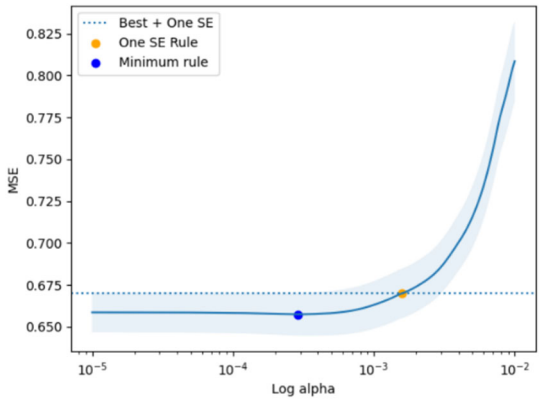

Figure 3. The effect of minimum leaf sample number on the AUC value of the prediction model

As can be seen from Figure 3, the training set AUC is higher than the test set on the whole, and the training set AUC slightly decreases with the increase of the change of the minimum leaf sample number, while the test set AUC jumps up when the minimum leaf sample number is 40, and then maintains a high level. The AUC of the test set decreased significantly until the minimum leaf sample size exceeded 140. When the minimum leaf sample number was 90, the maximum AUC value of the test set was 0.8339. When the minimum leaf sample number is relatively small, the model trained by the training set appears overfitting, the test set AUC value is low, and the model prediction level is low. When the minimum leaf sample number is set relatively large, the set model structure rules cannot describe and reflect the data characteristics to a large extent, and the AUC values of the training set and the test set are both at a low level, which indicates that the model is underfitting. Similarly, when the maximum depth of the decision tree model is set to four layers, the AUC value of the training set and the test set has the same trend with the minimum number of leaf samples.

5. Conclusion

The optimized decision tree model has a good predictive effect on the results of Chinese public policy decisions. The AUC value is 0.848 and the model generalization ability is strong. The difference between the training set and the test set AUC is less than 0.04. When the decision threshold is 0.4, the recall rate and accuracy rate predicted by the model for the test set data are both higher than 70%, which can realize the high-precision prediction of the results of Chinese public policy decisions.

References

[1] Li Yue. Evolution of public policy themes in public health emergencies: A case study of official wechat in national central cities. Journal of Information, vol. 39, pp.78-84, September 2020.

[2] Xu Chenxi, Chen Ying, Xie Baopeng, et al. Analysis of Obstacles to the implementation of the Policy of "Separation of Three Rights" in agricultural land: Based on Smith's policy implementation process model. China Land and Resources Economics, vol. 35, pp.79-84, June 2022.

[3] Huang Li, Huang Ansheng. An analysis of the change of forestry property rights Policy in China from the perspective of discontinuous equilibrium theory. World Forestry Research, vol. 34, pp.61-66, June 2021.

[4] Xu Guochong, Xu Yujun. Characteristic Change of Chinese Public policy in the New Era -- Discourse Analysis based on the Draft of "Five-Year Plan". Journal of CPC Tianjin Party School, vol. 24, pp.54-64, February 2022.

[5] Liu Xin, Wang Didian. Integration and Transformation: The Evolution of public policy on Intellectual property since the reform and opening up. China Soft Science, vol. 7, pp.12-21, October 2021.

[6] Han Yi, Liu Shasha. Multi-source model of public policy and its revision from the perspective of green development -- Taking the new National Standard of express package as an example. Journal of Sichuan Light Chemical Technology University: Social Science Edition, vol. 35, pp.170-174, January 2020.

[7] Xing Weibo, Tian Kun. Evaluating the potential benefits of public policy on tobacco control -- from the perspective of tobacco consumption and public health. Finance and Trade Economics, vol. 41, pp.16-25, November 2020.