Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070416
ISSN 2371-8838 Vol. 7 Num. 4

# *An Early Warning Model for Social Stability Based on the PSO-C4.5 Decision Tree Model*

**Hengfeng Shen[1], Henghua Shen[2], Ning Wang[1,*]**

*[1]Guangzhou Huashang College, Guangzhou, 522000, China*
*[2]Capital University of Economics and Business, Jieyang, 522000, China*
*[*]Corresponding author*

***Keywords:*** Social stability, Early warning model, Topsis, Decision tree

***Abstract:*** Since ancient times, social stability has always been an issue that people attach importance to, and changes in political, economic, cultural and other factors have brought new challenges and made the country fragile for social stability, so the study of maintaining social stability has become particularly important. This paper establishes an early warning model of social stability, first based on the topsis entropy weight method to calculate the comprehensive score of social stability in 215 countries, and at the same time encode the data based on the score to form three social stability grades of strong, medium and weak, and then establish the C4.5 decision tree model, after hyperparameter optimization through PSO, the optimal parameter social stability early warning model is established. The advantage of the decision tree model is that it plots the decision-making process for each metric. The decision tree chart can clearly display the threshold of each indicator when dividing the social stability level, which is conducive to early warning of the indicators of the social stability index system.

## 1. Topsis Entroropy Weight Method Social Stability Classification

TOPSIS Method is a commonly used comprehensive evaluation method within the group, which can make full use of the information of the original data, and its results can accurately reflect the gap between the evaluation schemes[1]. The basic process is based on the original data matrix after normalization, and the cosine method is used to find out the optimal scheme and the worst scheme in the finite scheme, and then calculate the distance between the evaluation object and the optimal scheme and the worst scheme, to obtain the relative proximity of the evaluation object and the optimal scheme, as the basis for evaluation. The method has no strict restrictions on the data distribution and sample content, and the data calculation is simple and easy[2].

TOPSIS The basic idea of the method is: build a normalized matrix for the original data, calculate the difference between the evaluation object and the optimal vector and the worst vector, so as to measure the difference between the object and the object. Suppose that there are n evaluation objects, m indicators, and the basic steps of the TOPSIS method are:

Step 1: Original data is trending[3]

Distinguthe index categories (high excellent or low excellent) in the index system, and make forward treatment according to different types of indicators need to be made according to different formulas.

Construct the matrix X ij of column n row m, where X represents the value of the j th index of the i th object.

Step 2: to construct a standardization matrix[4]

$$Z_{ij} = \frac{X_{ij}}{\sqrt{\sum_{k=1}^{n} (X_{ij})^2}} \tag{1}$$

Step 3: calculates the gap between each evaluation index and the optimal and worst vectors

$$D_i^+ = \sqrt{\sum_{j=1}^{m} w_j (Z_j^+ - z_{ij})^2}, D_i^- = \sqrt{\sum_{j=1}^{m} w_j (Z_j^- - z_{ij})^2} \tag{2}$$

Where wj is the weight (importance) of the j-th attribute.

Step 4: measures the proximity of the evaluation object to the optimal scheme

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{3}$$

A ger Ci value indicates better the evaluation object

Topsis Method is generally used together in combination with the entropy weight method. Entropy is a concept in information theory and is a measure of uncertainty. Larger information content means less entropy, less uncertainty and more entropy. According to the definition of information entropy, the entropy value of an index can be used to judge the dispersion degree of an index, the smaller the entropy value, the greater the dispersion degree of the index, and the greater the influence of the index (i. e. weight) on the comprehensive evaluation[5].

The steps are:

(1) Each factor was normalized according to the number of each option:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} or x'_{ij} = \frac{\max(x_j) - x_{ij}}{\max(x_j) - \min(x_j)} \tag{4}$$

(2) Calculate the proportion of the j th indicator for the i th user:

$$y_{ij} = \frac{x'_{ij}}{\sum_{i=1}^{m} x'_{ij}} \tag{5}$$

(3) Calculate the information entropy of the j-th index:

$$e_j = -K \sum_{i=1}^{m} y_{ij} \ln y_{ij} \tag{6}$$

Where, K is a constant

$$K = \frac{1}{\ln m} \tag{7}$$

(4) Calculate the weight of the j th index

$$w_j = \frac{1-e_j}{\sum_j 1-e_j} \tag{8}$$

The following table 1 shows the weight calculation results of the entropy weight method[6], and the weight of each index is analyzed according to the results:

Table 1: Weight Calculation Results Table entropy weight method

| Item | Information entropy value e | Information utility value d | weight (%) |
|---|---|---|---|
| population | 0.973 | 0.027 | 4.398 |
| finance | 0.962 | 0.038 | 6.168 |
| economy | 0.923 | 0.077 | 12.542 |
| Workforce population (person) | 0.779 | 0.221 | 35.88 |
| education | 0.991 | 0.009 | 1.492 |
| hygienism | 0.886 | 0.114 | 18.571 |
| Energy production and use | 0.976 | 0.024 | 3.853 |
| investment climate | 0.944 | 0.056 | 9.09 |
| trade | 0.978 | 0.022 | 3.647 |
| government finance | 0.985 | 0.015 | 2.466 |
| Monetary stability | 0.993 | 0.007 | 1.095 |
| Poverty and income | 0.995 | 0.005 | 0.798 |

Table 1 showcase the weight calculation results of the entropy weight method, Population weight of 4.398% and finance of 6.618%, The weight of economy is 12.542%, the weight of labor population (people) 35.88%, that of education 1.492%, of health 18.571%, of energy production and use 3.853%, of investment environment 9.09%, of trade 3.647%, of government finance 2.466%, of monetary stability 1.095%, of poverty and income 0.798%, the maximum index weight of the labor population (people) (35.88%), The minimum value was poverty and income (0.798%)[7].

Table 2: Comprehensive score table of entropy right topsis method

| index of matrix | Positive ideal solution distance (D +) | Negative ideal distance (D-) | Comprehensive score index | sort |
|---|---|---|---|---|
| Algeria | 0.912 | 0.158 | 0.147 | 195 |
| Angola | 0.851 | 0.276 | 0.244 | 39 |
| Benin | 0.901 | 0.237 | 0.208 | 71 |
| Botswana | 0.864 | 0.221 | 0.204 | 77 |
| Burundi | 0.899 | 0.208 | 0.187 | 100 |
| Cameroon | 0.902 | 0.205 | 0.185 | 107 |
| Cape Verde | 0.868 | 0.230 | 0.209 | 69 |
| Central African | 0.804 | 0.460 | 0.364 | 5 |
| Chad | 0.827 | 0.303 | 0.268 | 28 |
| Comoros | 0.890 | 0.224 | 0.200 | 81 |
| Congolese (cloth) | 0.841 | 0.308 | 0.268 | 29 |

The following table 2 shows the comprehensive score calculated by the entropy weight topsis method. Where D + and D-value, the two points to represent the evaluation object and the optimal or worst solution (A + or A-) distance (European distance), the practical meaning of the two values, the evaluation object and the optimal or worst solution, the greater the value of the distance, the farther, the larger the object D + value, the farther and the optimal solution distance, the greater the D-value, the farther the distance with the worst solution[8]. The most understood subject is that the smaller the

D + value and the larger the D-value. Comprehensive degree score C value, C= (D-) / (D + + D-), the calculation formula, the molecule is D-value, the denominator is the sum of D + and D-: the larger the D-value,

It means that the further the object is, the better the object is; the greater the C value, the better the object is.

Due to space constraints, only some of the results are presented here:

On the base of table 2, In order to facilitate the model solution of the social early warning model, we divided the comprehensive score into three categories through the third level. The higher the value, the higher the comprehensive evaluation score of social stability. Some of the data are as follows:

Table 3: National Stability Classification Table

| country | population | finance | economy | ... | Poverty and income | Monetary stability | Comprehensive score index | bracket |
|---|---|---|---|---|---|---|---|---|
| Central African | 1.0 | 0.04 | 0.100 | ... | 0.395 | 0.282 | 0.364 | 3 |
| Chile | 0.241 | 0.323 | 0.246 | ... | 0.395 | 0.282 | 0.195 | 2 |
| Gibraltar | 0.323 | 0.203 | 0.211 | ... | 0.395 | 0.282 | 0.181 | 2 |
| Chad | 0.555 | 0.030 | 0.103 | ... | 0.395 | 0.282 | 0.268 | 3 |
| Zambia | 0.471 | 0.053 | 0.028 | ... | 0.395 | 0.282 | 0.181 | 2 |
| Vietnam | 0.291 | 0.446 | 0.097 | ... | 0.395 | 0.282 | 0.172 | 2 |
| The West Bank and Gaza | 0.376 | 0.027 | 0.064 | ... | 0.395 | 0.282 | 0.141 | 1 |
| Jordan | 0.230 | 0.276 | 0.099 | ... | 0.395 | 0.282 | 0.16 | 1 |
| British Virgin Is | 0.324 | 0.203 | 0.211 | ... | 0.395 | 0.282 | 0.176 | 2 |
| Britain | 0.226 | 0.536 | 0.493 | ... | 0.395 | 0.282 | 0.249 | 3 |
| Indonesia | 0.317 | 0.124 | 0.116 | ... | 0.481 | 0.282 | 0.217 | 2 |

Table 3 shows the overall scores and bands of each country.

## 2. Establishment and Solving of the Early Warning Model of Social Stability

### 2.1. Define Indicators to Express the Classification Ability of the Model

In the process of model building we need to adjust the model training results to make it gradually close to the real situation, so need some indicators to measure the execution ability of the model, and data mining classification the two most basic index is the field of recall and accuracy, recall rate also called recall, accuracy also called accuracy, concept formula [9]:

Recall rate = number of correct transaction information identified by the model / total number of transaction information identified by the model

Accuracy = number of correct transaction information identified by the model / total number of transaction information identified by the model

It is worth noting that the accuracy and recall rate affect each other. Ideally, both must be high, but generally the accuracy is high and the recall rate is low, while the recall rate is high and the accuracy is low. To weigh the effects of the two, we introduce the F-Measure integrated model. F_Measure is the weighted harmonic average of accuracy and recall. Where P represents accuracy and R represents recall:

$$F = \frac{(\alpha^2 + 1)P * R}{\alpha^2 (P + R)} \tag{9}$$

When the parameter α =1. that is, the common F1, that is:

$$F = \frac{2 * P * R}{P + R} \tag{10}$$

At the same time, in the evaluation criteria of classification effect evaluation, the misjudgment rate M is also an important indicator, while the misjudgment rate M can be expressed as follows:

$$M = \frac{FP + FN}{Allcount} \tag{11}$$

Where FP is the count predicted 1 actually 0, FN is the count predicted 0 actually 1, and Allcount is the total number of samples.

## 2.2. Establishment of the Decision Tree Model

The decision tree is a tree-like structure, which starts from the root node and tests the data sample, divides the data sample into different data sample subsets according to different results, and each data sample subset constitutes the —— subnode. It is a process of classifying the data through a series of rules. It provides a method of a similar rule of what values are obtained under what conditions. Decision tree is divided into classification tree and regression tree. Classification tree makes decision tree for discrete variables, and regression tree makes decision tree for continuous variables.

When C4.5 algorithm selects attributes at the nodes of the decision tree, the gain ratio (gain ratio) is used as the selection criterion for attributes.

$$SplitInfomation(A, S) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{12}$$

$$GainRatio(A, S) = \frac{Gain(S, A)}{SplitInfomation(S, A)} \tag{13}$$

In SLIQ, SPRINT, PUBLIC algorithm, use gini index (gini index) instead of information quantity (Information) as the standard for attribute selection. The gini metric performs better than informational quantities and is easy to calculate. For a dataset containing n classes, dataset S, gini (S) is defined as:

$$gini(S) = 1 - \sum p_j * p_j \tag{14}$$

Where p i is the frequency of class j data in S. The smaller the gini is, the larger the Information Gain is. The figure 1 is workflow diagram for a decision tree.

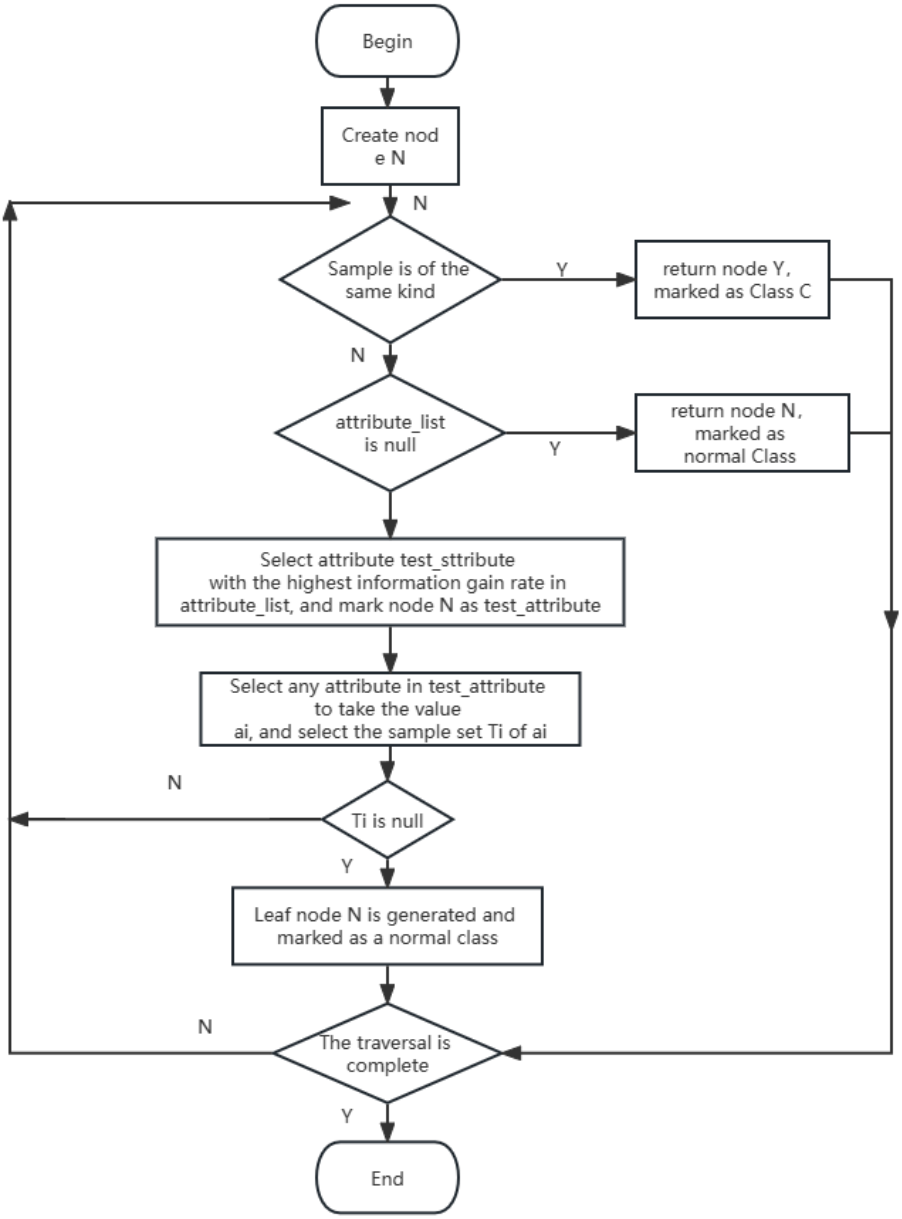


Figure 1: Flow chart of the decision tree

## 2.3. PSO Hyperparameter Model Tuning

By constructing a good decision tree, it has a good effect on solving the problem, but usually there are a large number of hyperparameters in the model. Hyperparameters refer to the parameters that cannot be learned from the data during the training process. The setting of hyperparameters will have a direct impact on the performance and performance of the model[10]. When determining the model hyperparameters, we often seek optimization according to the figure 2.

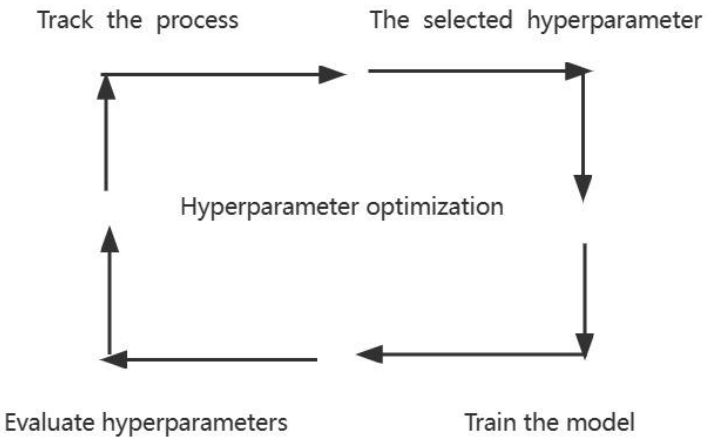


Figure 2: Flow chart of automatic optimization of PSO hyperparameters

The standard PSO algorithm is a global optimization algorithm that concentrates the "group" and "optimization" principles to optimize through the adaptive value of the particles. The Pso algorithm retains the population-based global search strategy, viewing each individual as a weight-free and bulk particle in the n-dimensional search space and flying at a certain speed that is dynamically adjusted by the individual flight experience and the flight experience of the population. In each iteration, each particle adjusts its flight speed and position according to the following formula:

$$\begin{aligned} v_{ij}(t+1) &= wv_{ij}(t) + c_1 r_{1j}[p_{vj}(t) - x_{ij}(t)] + c_2 r_{2j}[p_{gj}(t) - x_{ij}(t)] \\ x_{ij}(t+1) &= x_{ij}(t) + v_{ij}(t+1) \end{aligned} \tag{15}$$

Where j represents the particle j th dimension; i represents the i th particle; t represents the t th generation; c1, c2 represents the acceleration constant;

The steps based on the PSO optimization parameters are shown in the figure 3:

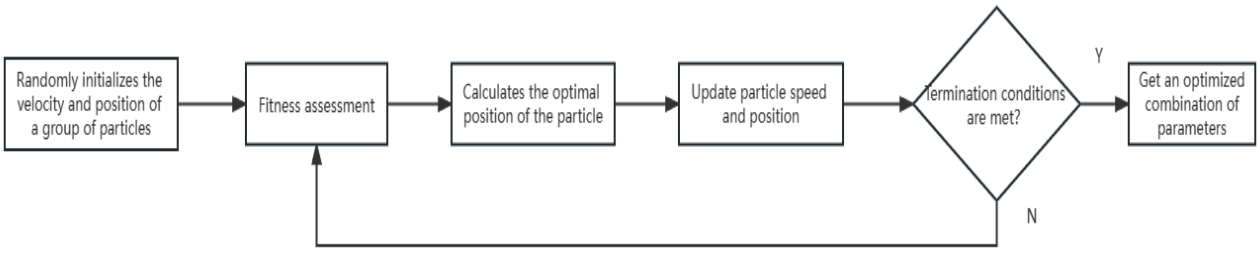


Figure 3: Step diagram of PSO optimization parameters

## 3. Model Solution

Based on the comprehensive score as Y, the data are divided into training set and test set, in which the proportion of test set is 20%. After training the model, the model performance parameters calculated based on the prediction results of the classification model are obtained in the following table 4. It can be seen that the effect of the model is excellent.

Table 4: Model Performance Table

| class | Precision | Recall | F1-score |
|---|---|---|---|
| 1.00 | 0.43 | 1 | 0.6 |
| 2.00 | 1.00 | 0.78 | 0.88 |
| 3.00 | 0.51 | 1 | 0.67 |
| Wei0ghted avg | 0.91 | 0.81 | 0.84 |

The following table 5 is the parameter table of PSO-C4.5 decision tree model. The classification model can be set to obtain the optimal solution:

Table 5: Summary of the optimal solutions of the model parameters

| parameter name | optimal value | parameter name | optimal value |
|---|---|---|---|
| Max_depth | 5 | Min_impurity_decrease | 0.1 |
| Min_samples_split | 2 | Ccp_alpha | 0.2 |
| Min_samples_leaf | 1 | Learning_rate | 0.14 |
| Min_weight_fraction_leaf | 1 | Max_leaf_nodes | 0 |

Figure 4 shows the two-node division path of this decision tree model. Based on the threshold of these indicators, we can obtain the early warning value of social stability in each country.

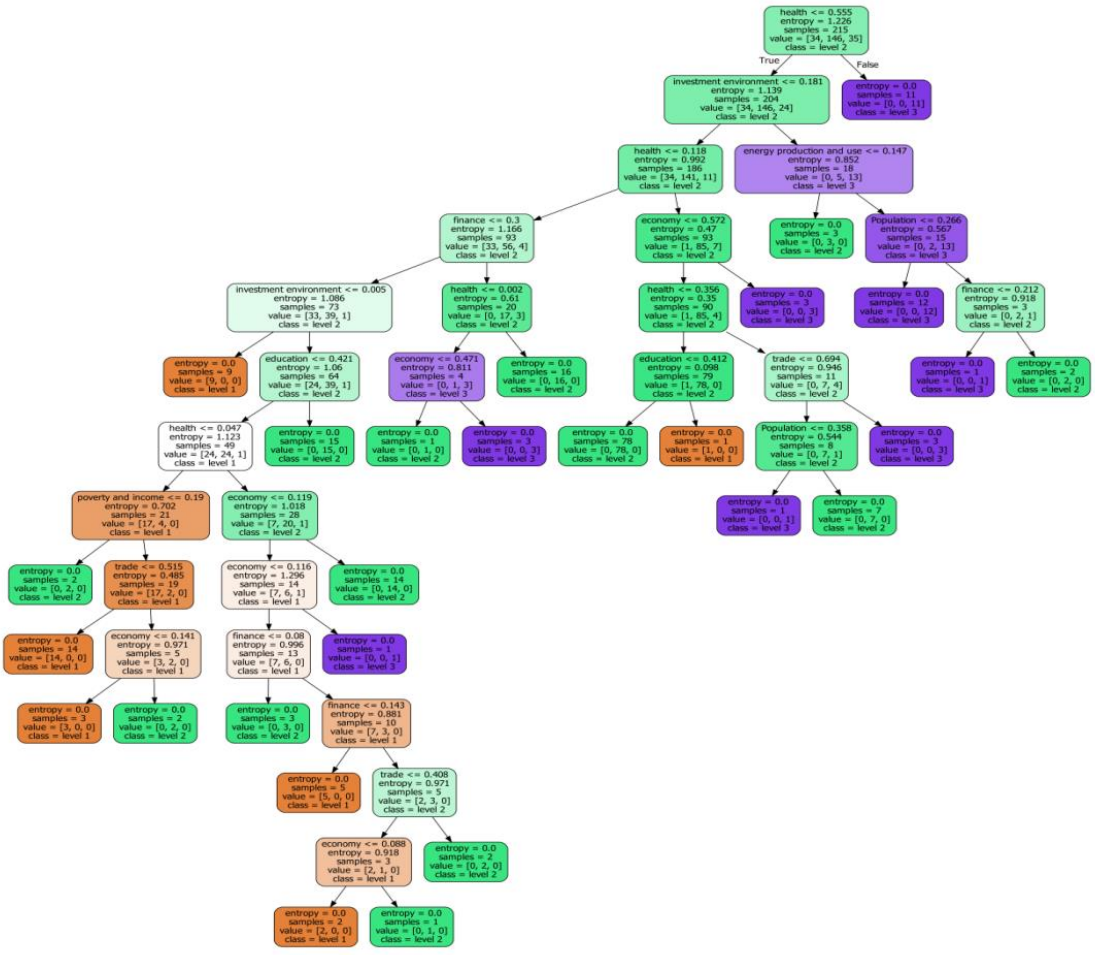


Figure 4: Decision tree model

## 4. Conclusion

Based on the PSO-C4.5 social stability early warning model in this paper, we can get the comprehensive score of various indicators in the analysis of topsis entropy weight method, clearly see the importance of the impact of various indicators on social stability, and find that the highest weight is the labor force, population, health and economy, and then use data coding to arrange the social stability of countries around the world, and divide the comprehensive score into three binches through the thirdary, the higher the value, the higher the comprehensive evaluation score of social stability. Then, by establishing a C4.5 decision tree model based on PSO algorithm hyperparameter automatic optimization, and listing the model performance table and model parameter optimal solution summary table, and listing the decision tree model diagram in this paper, which is convenient for us to better observe, the model diagram shows the decision-making division threshold of various indicators in the model, which provides strong support for the early warning of various indicators in society. Through the social stability early warning model established in this paper, it can play a certain warning and early warning role for the early situation of social stability in various countries, through which local governments can take timely measures to solve public dissatisfaction and concerns, such as corruption, lack of political freedom and economic inequality, focus on promoting economic growth and development, and strive to create employment opportunities, raise income levels, reduce poverty and inequality, etc. to maintain social stability.

## Acknowledgments

This project is sponsored by the Youth Academic Program of Guangzhou Huashang University under grant 2022HSXS081, the Project Name is Study on Road Administration Lighting System of Smart City.

## References

*[1] Wang Yuanyuan, Bai Hongkun, Wang Shiqian, Bu Feifei, Wu Xiong, Li Haoyu. Power User Behavior Portrait Based on Information Gain and Spearman Correlation Coefficient. Electric Power Engineering Technology, 2022, 41 (04): 220-228. (in Chinese)*

*[2] Chen H. Analysis of water quality change characteristics in Shanghai based on Spearman rank correlation coefficient method. Environmental science and technology, 2020 (3): 28 and 33, DOI: 10.19824 / j.carol carroll nki cn32-1786 / x. 2020.0037.*

*[3] Li Ying, Ge Xizhen. Construction of cucumber downy mildew early warning system in greenhouse based on PCA and multiple regression algorithm. Journal of Anhui Agricultural Sciences, 2022, 50 (21): 232-234.*

*[4] Mo Yun, Guo Yan, Mo Hesheng, Lu Zhongwei, Zhang Shaorong. Hybrid feature selection method based on feature filtering and PCA dimension reduction. Journal of Guilin University of Astronautics Technology, 2022, 27 (02): 145-151. (in Chinese)*

*[5] Hou Meien, Wang Xiangbin. Extraction of biomechanical characteristics of gait in KOA patients by PCA. Medical Biomechanics, 2021, 36 (S1): 224.*

*[6] Wang Xueli. A study on the Effect of Online celebrities' Live streaming on Consumers' Purchase Intention. Gansu Science and Technology, 2020, 49 (11): 76-80.*

*[7] Cong Xiaoqi. Research on Entrepreneurial self-efficacy of accounting based on hierarchical regression. Journal of Science and Technology Entrepreneurship, 2020, 33 (08): 129-131.*

*[8] Liu Xiaojun. Study on Comprehensive Benefit Evaluation of Green construction Land saving Technology based on topsis entropy weight method. Mathematics practice and understanding: 1-10 [2023-02-05].*

*[9] Wang Zitao, Wang Jianping, Han Guang. Evaluation of groundwater richness based on entropy weight TOPSIS method in Qaidam Basin. Salt Lake Research, 2022, 30 (02): 42-51.*

*[10] Hu Mingqi, Zhang Senchang. Leaf node weighted random forest algorithm based on PSO optimization. Modern Computer, 2022, 28 (04): 1-4.*