

Study of plant communities based on supervised learning model SVM

Mingzhong Du^{1,*}, Chenghao Zhang², Xujie Mao³, Xun Tang⁴
and Yaowen Cao⁵

¹ School of Energy and Power Engineering, Nanjing Institute of Technology, Nanjing, China

² School of Power Engineering, Nanjing Institute of Technology, Nanjing, China

³ College of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing, China

⁴ School of Automotive and Rail Transportation, Nanjing Institute of Technology, Nanjing, China

⁵ School of Mechanical Engineering, Nanjing Institute of Technology, Nanjing, China

* Corresponding Author Email: 2589169957@qq.com

Abstract. Numerous studies have shown that the number of different species plays an important role in how the plant community adapts during successive multi-generational drought cycles. In this paper, we first collected relevant data and completed the pre-processing of the data. Then the community ecological environment index was quantitatively analyzed through indicators such as biodiversity and birth mortality in the region, and the correlation between the characteristic X indicators was obtained through correlation analysis, and the supervised learning relationship between the characteristic indicators, indicators and ecological environment index was established, and the best SVM model was obtained through Bayesian optimization and network parameter synchronization. The final model accuracy is 91.6%, and the root mean square error is less than 0.382, which is adaptable and strong scientific. Based on this model, the number of species in the index, the correlation between species and the value of the environmental quality index kept changing, the data changed in the model to the corresponding changes in the ecological environmental index, in order to see the impact of different environments and species in the ecosystem, a qualitative analysis was performed, and finally the impact of the number and type of species on the plant community was studied.

Keywords: SVM; Grey relational analysis; Plant communities; CRITIC method.

1. Introduction

In some plant communities with only a single species, offspring do not adapt to drought conditions as readily as individual plants in communities with four or more species. So, for a plant community to benefit from this local biodiversity, what is the minimum number of species needed? How does this phenomenon develop as the number of species increases?

In this paper, we first reduced, cleaned and normalized the data. Secondly, we quantitatively evaluated and analyzed the ecological and environmental indicators of the species in the area, which were evaluated and elaborated by CRITIC. Then, supervised learning relationships between species changes, relationships between species and environmental indicators and population ecological indicators were established, and SVM support vector machine models were built [1]. Finally, based on our SVM-optimized monitoring model, we adjusted the indicators separately, constantly changing the number of species, ecological diversity and environmental indicators, and compared, analyzed and processed the obtained results.

2. Plant community change model

2.1. Data pre-processing

2.1.1 Data Noise Reduction

Anomalies are extreme manifestations of overall partial variability either due to accidental deviations from experimental conditions and methods, or errors in observation, recording, calculation, or data transmission errors. By turning abnormalities into normal points through data adaptation, the model curve is smoother, the results are more accurate and the results are more convincing [4].

We decided to use a normality test for each eigenvalue and the test of using the data observed to determine whether the population obeys a normal distribution became a normal test, which is a particular test of the well ness hypothesis that is important in statistical judgments.

It represents an eigenvalue defined by the general pleasure of people, represents the type of sample, and represents the moment of the center of the sample order of the order. Skewness and the peak value of normal distribution are 0, where skewness and curtosis are determined respectively $Y_1, Y_2, \dots, Y_n, \bar{X}$ $m_i = \frac{1}{n} \sum_{j=1}^n (\frac{Y_j - \bar{Y}}{Y})^i$.

2.1.2 Data Cleaning

The missing values do not comply with the laws of mathematical models and big data statistics, and are most affected by external factors. Some outliers will lose valuable information during data mining modeling, resulting in more significant uncertainties in the model, and the laws contained in the model are harder to control, which can also cause confusion in the modeling process and form unreliable exits. The addition of missing data points will make the overall range of change the model more accurate, reduce error, so that it is more perfect, more optimized, and the results are more accurate, and these additional data missing points will not have much impact on the overall change trend.

2.1.3 Standardization and follow-up

Methods for standardizing data are Min-max normalization, z-dot normalization, and standardization of decimal scaling. In this question, we use the Z-Score normalization method. First, mathematical hope and standard deviation of each variable are found, then the formula $Z_{ij} = (X_{ij} - X_i)/S_i$ is used to standardize the data, and ultimately the positive and negative signs before the reverse indicator is reversed.

2.2. Model Building

2.2.1 Selection of indicators affecting the community

We chose the temperature, rainfall, evaporation and pressure of water vapor data monthly data from the Hetao irrigation area (typical dry area) of Inner Mongolia from 2010 to 2020, including specific quantities of the four main crops and after the first step of data laundering, noise reduction and preprocessing, We received a set of standard data and first visualized it separately for change analysis:

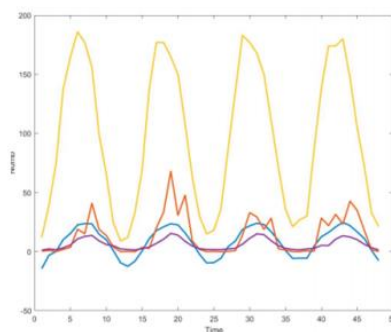


Fig. 1 10-year temperature data

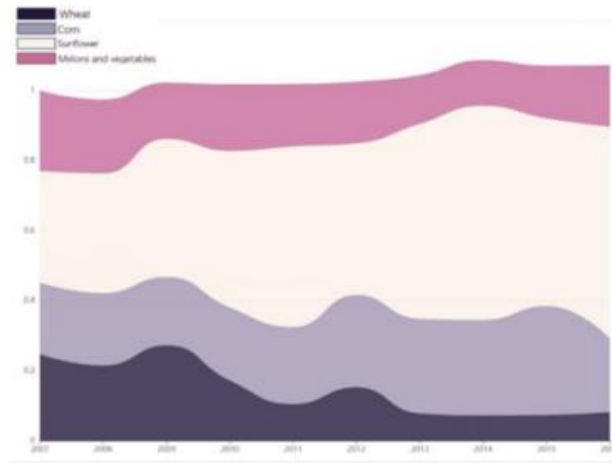


Fig. 2 10 years change in crop number

2.2.2 Quantitative analysis of ecological and environmental indicators

We start with a quality analysis of climate, environment and crop quantity. Second, for community ecosystem changes in the decade from 2010 to 2020, we used ecological diversity, species number and birth/mortality rate as the evaluation criteria to create a comprehensive assessment system for the region's ecological environment, in which the assessment indicators were the three standards of the above assessment.

The CRITIC algorithm synthesizes the index's objective weight based on the contrast strength of the rating index and conflict between indicators. Considering the variability of indicators while taking into account correlation between indicators, it is not that the higher the number, the more important it is and the objective the attributes of the data themselves are fully used for scientific evaluation [3].

For the CRITIC method, when the standard deviation is fixed, the smaller the conflict between indicators, the smaller the weight; The greater the conflict, the greater the weight; In addition, the greater the positive correlation between the two indicators (the closer the correlation coefficient to 1), the less conflict, indicating that the two indicators reflect greater similarities in the information reflected in the assessment of the advantages and disadvantages of the scheme.

(1) CRITIC ALGORITHM CLAIMS INDICATIVE WEIGHT

The CIRTIC algorithm synthesizes the index's objective weight based on the contrast strength of the rating index and conflict between indicators. Considering the variability of indicators while taking into account correlation between indicators, it is not that the higher the number, the more important it is and the objective the attributes of the data themselves are fully used for scientific evaluation.

CRITIC calculation process and steps.

$$\begin{cases} \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \\ S_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}} \end{cases} \quad (1)$$

The greater the variability of the index, the more information reflects, The stronger the intensity of the index's own assessment, the more weight should be assigned to the index.

$$R_j = \sum_{i=1}^p (1 - r_{ij}) \quad (2)$$

R_{ij} represents the coefficient of correlation between the indicators of evaluation and j . The more the same information is reflected, the more duplication of the rating content can be reflected. To some extent, this weakens the intensity of the index assessment and the weight assigned to the index should be reduced.

$$C_j = S_j \sum_{i=1}^p (1 - r_{ij}) = S_j \times R_j \quad (3)$$

The larger the C_j , the greater the role of the j th rating indicator in the overall rating index system and it should be given more weight.

$$W_j = \frac{C_j}{\sum_{j=1}^p C_j} \quad (4)$$

(2) Quantify the population ecological environment index through the assessment model.

$$Q = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \quad (5)$$

The quantitative Q index matrix of the population ecological environment obtained from the data in step 5 is all very large data, from which the ideal optimal solution and the worst solution can be extracted from you.

The optimal solution and the worst solution are

$$q_{min} = [m(q_{12} \dots q_{m1}), m(q_{1n} \dots q_{mn})] \quad (6)$$

The distance between the i -th component and optimal and worse solutions can be calculated.

$$d_i^+ = \sqrt{\sum_{j=1}^m [w_j (q_{ij}^+ - q_{ij})]^2} \quad (7)$$

The result of the i -th part can be counted.

$$d_i^- = \sqrt{\sum_{j=1}^m [w_j (q_{ij}^- - q_{ij})]^2} \quad (8)$$

$0 \leq S \leq 1$, when d , the smaller the distance, the higher the result; The smaller the d , the smaller the distance from the worst solution, and the lower the result. In this way, we can analyze the social stability of the region according to the level of points and we can get the corresponding result of adaptation according to the model.

We take the final trend of the quality of the ecological environment of the population in the area, as shown in Figure 3.

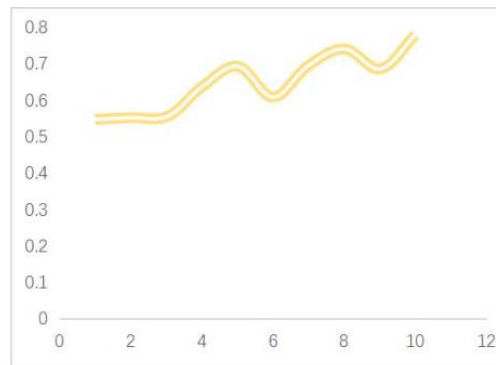


Fig. 3 Changing the quality of the ecological environment

2.2.3 Relationship between indicators and environmental indicators

Grey correlation analysis is suitable for analysis of interrelationship between multiple factors in the system, and its fundamental idea is to judge the narrowness between parent factor and multiple subfactors according to the geometric similarity of parent sequence turns and subsequence [2]. Based on comprehensive consideration, we will use the grey correlation analysis method to judge the relative relationship between the eight selected indicators and the ecological population of the environmental index by taking the grey correlation rate between the indicators.

Through grey correlation analysis, we received several high-correlation indicators and excluded indicators with low correlation with the ecological environment, as shown in figure 4 below.

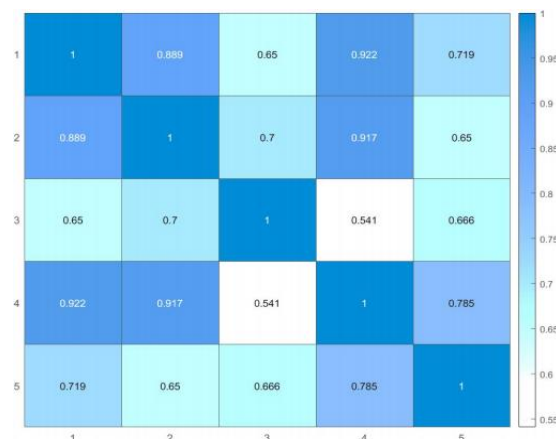


Fig. 4 Trends in the quality of the ecological environment

The high-correlation indicators obtained from our analysis here are considered to have strong correlations and are almost combined with each other in the build-up to the following model.

2.3. Supervised learning model SVM

2.3.1 Selection of community models

Through quantifying the region's ecological environment index and indicator correlation analysis, what we must do now must be to establish the functional relationship between the population ecological environment index in the region and correlation between different species, species and environmental impact factors. So we decided here to create a supervised learning system to achieve the implicit function link between XY by establishing the iteration of dark boxes.

Therefore, we must first determine how correlation between indicators is replaced in the learning system, and how to choose the learning model, so we first consider iterative the selection of the learning model using the decorrelated indicative data, and iterate through synchronizing network parameters to get the best model.

2.3.2 Iteration of supervised learning models

The supervised effect of model learning was obtained by teaching the usual algebraic classification of some models on the data through parallel learning method, and two highly monitored models were taken for Bayesian network study parameters and optimization respectively.

In order to evaluate the prediction error and accuracy of the models, the root mean square error (RMSE) error (RMSE), root mean square error (MSE), and goodness of fit (R^2) are used to evaluate the model prediction results The RMSE, MSE, and R^2 are used to evaluate the model predictions.

2.3.3 Iteration of supervised learning models

The team adjusted the following parameters of the SVM vector car with the method of adjusting network parameters to achieve the goal of optimizing the RMSE result when the highest is achieved.

Iteration parameters and the iteration process are presented in Figure 5 below.

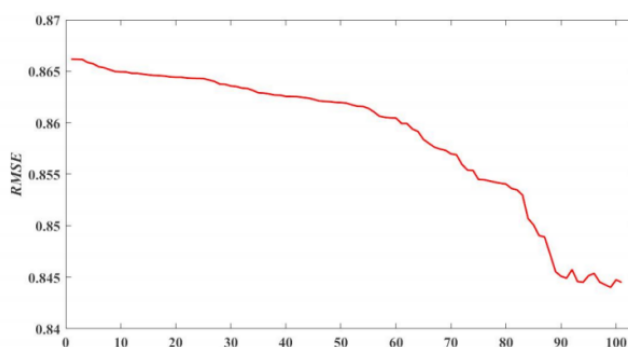


Fig. 5 Iterative parameter changes

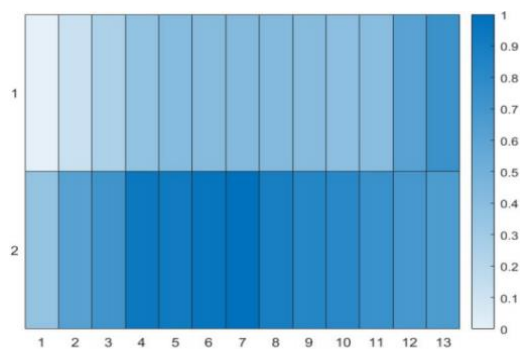


Fig. 6 RMSE and heat map R^2

The RMSE obtained and the heat map, as shown in Figure 6 below, and the importance of our team, we prefer to follow the smallest value for RMSE, but there are still some reference values, so we think RMSE and R^2 that the smaller the RMSE, the better, the better, so we follow the two values, conduct weighted analysis processing

The final confidence of histogram prediction is obtained through the cross-validity method of generation K=10:

Axis X presents the confidence interval for each point of raw data, and Axis Y presents the proportion of data points under that confidence interval. The team found that when the confidence interval is 90%~95%, the proportion of data is up to about 90%, meaning that most of the predicted data is within 1 point of the original data, and 81% of the data is within 0.5 points. By replacing data on the number of species and environmental impact factors on the appendix in the model, the following forecast data will be taken, first we show the difference between the predicted data distribution map and the exact data:

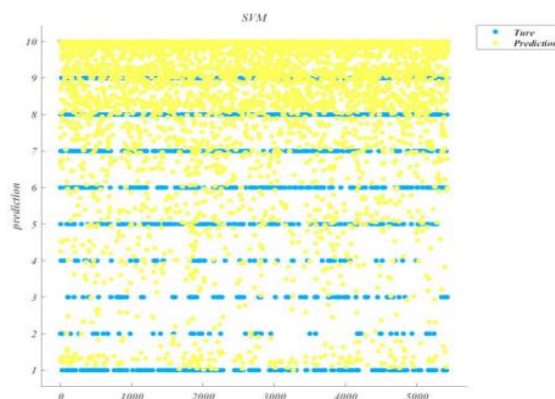


Fig. 7 Forecast Data Comparison

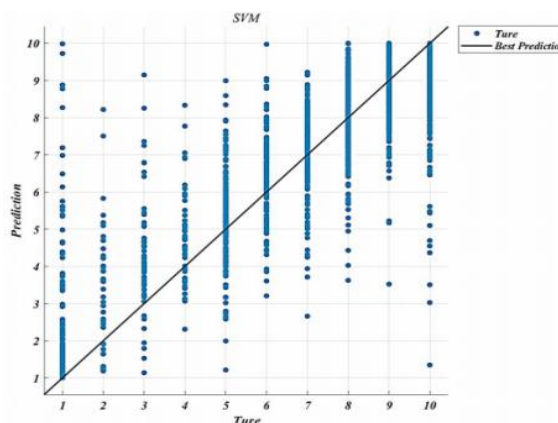


Fig. 8 Error data graph

At this time, the accuracy of the model reaches 91.3% and the root means square error RSME <0.2, which has strong data accuracy and explainability, and it can be clearly concluded that our team's

overall data laundering, noise reduction preprocessing, index selection, ecological index quantification and other work is better.

3. In-depth analysis of model results

3.1. The impact of the number of species

We decided to observe the trend of the population ecological environment index obtained from the supervised learning model by increasing data for increasing the number of species or reducing the number of species, as well as increasing correlation and decreasing correlation.

The iteration diagram of each parameter and the schematic diagram of parameter error are as follows:

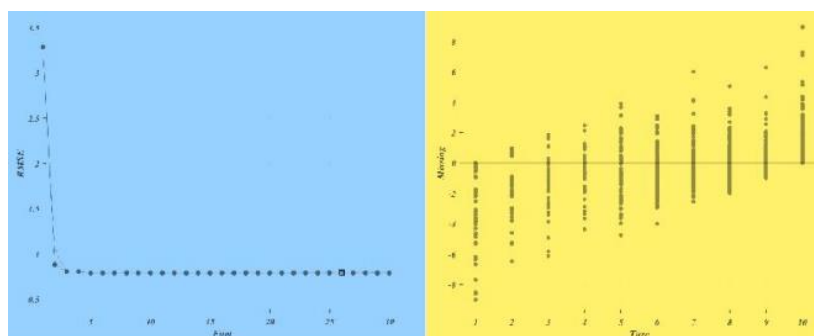


Fig. 9 Strong link between species

We can still clearly see that the stability, accuracy and suitability of the model remain almost unchanged after the increase or decrease in the number of indicators, so we can make many changes in indicators, species changes and correlation changes vary from 0-20 species and (0.0)-(30,30!)), these results are irritated in separate models, And some classic turns of population ecological index changes stem from the results:

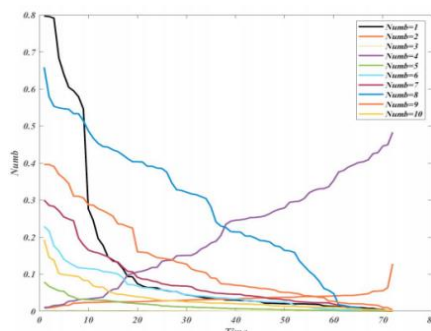


Fig. 10 Strong link between species

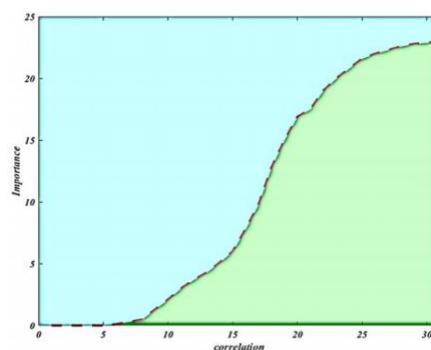


Fig. 11 Strong link between species

Through analysis, we can obtain that when the number of species is greater than 8 and the correlation of the indicators is greater than (3.8), the population ecological environment index shows

a slowdown in the state of recovery, and it can be seen that when the two conditions above are met, the mutual stimulating effect between species is higher than the effect of inhibition and destruction of the arid environment.

Through analysis, we can obtain that when the number of species is greater than 3 and the correlation of the index is greater than (2.2), the population ecological environment index shows a slowing state and it can be seen that when the two conditions above are met, mutual stimulation between species will play a role, which will produce some drought resistance.

Through analysis, we can obtain that when the number of species is greater than 15 and the correlation of indicators is greater than (10.78), the population ecological environment index shows a state where the population ecological environment index shows a state almost without decline at first and increases and the growing trend gradually slows down in a stable state, at which time you can be seen that when the two conditions above are met, reciprocal incitement between species has an absolute effect and the negative effect of drought will be reduced almost to 0.

So we end up visualizing the number of species, the correlation between species and the impact of the dry environment, as shown in the following figure 11.

3.2. Effect of species type

We decided to observe the trend of changing the population ecological environment index obtained from the learning model monitored by constantly changing species with different changes in drought resistance, and first we modified the resistance of species drought, i.e. the apparent trend of species decline with the progression of drought time, and replaced it in our supervised learning model. The iteration diagram of each parameter and the schematic diagram of parameter error are as follows:

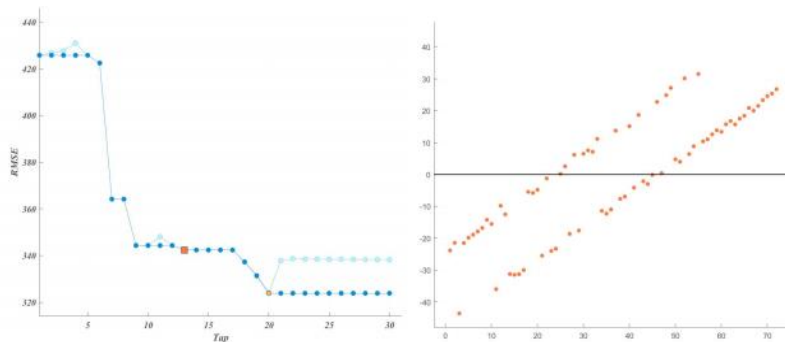


Fig. 12 Changes in population ecological index

We can still clearly see that the stability, accuracy and suitability of the model remain almost unchanged after the increase or decrease in the number of indicators, so we can make many changes in the indicators, we greatly regulate the resistance of the species' drought and their ability to decrease, respectively, these results are used for replacing models, from the results to obtain some classic population ecological index curves:

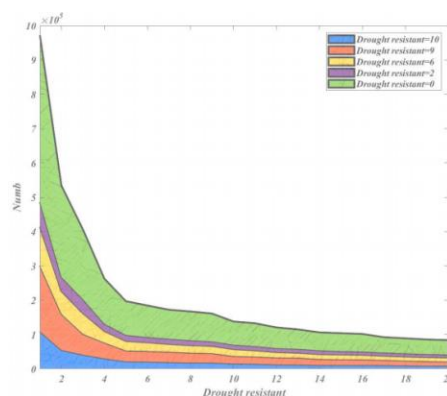


Fig. 13 Changes in population ecological index

The bottom line: As shown in the figure above, it is clear that when the species species changes, the overall stability of the respective ecosystem will have a falling trend and the lower the drought the resistance of the species, the sooner the combined result falls at the end.

4. Summary

This paper comprehensively examines the various characteristic factors affecting the environment of community ecosystems and incorporates the data into a normative study that requires comprehensive consideration. The lacking values and higher values in the original data are processed and normalized to make the data more refined and accurate. The SVM support vector machine model is built and comparatively analyzed with strong rationality and more accurate prediction results. SVM support vector machines use kernel functions to map large size spaces, which can solve nonlinear classification and classification problems with simple ideas and better classification results. In this paper, we use multiple heat maps for visual representation to facilitate the understanding and use of the model.

However, SVM support vector machines are difficult to solve large-scale training samples and multi-classification problems. It is sensitive to missing data and the choice of kernel parameters and functions. Also, we have estimated some values of the model, which can lead to some errors in the results.

References

- [1] D.B. Wei, N. Wei, L. Yang, et al. An mRMR-SVM-based data flow detection method for spatial information networks[J]. Computer Applications and Software,2022,39(8):111-118.
- [2] Zheng Zhen,Cha Bingting,Tan Yuran,et al. A three-dimensional point cloud recognition algorithm based on gray correlation analysis[J]. Journal of Nanjing University of Technology (Natural Science Edition),2022,46(6):679-687.
- [3] Zhang Chi, Yang Li, Zhu Junqi, et al. Evaluation of ecological and environmental carrying capacity of Yangtze River economic zone based on CRITIC-TOPSIS model [J]. Journal of Hebei Institute of Environmental Engineering,2022,32(5):4-9.
- [4] Xu Hui. Data preprocessing in data mining [J]. Computer Knowledge and Technology,2022,18(4):27-28,31.