

Ecological Security Classification of Regionally Sustainable Utilization of Land Resources Based on SVM

ZHAO Jing^{1, a}, JIN Zhen^{2, b}

¹ School of Economics and Management, Xi'an University of Technology, Xi'an, China

² School of Economics and Management, Xi'an University of Technology, Xi'an, China

^a zj00000000@126.com, ^b z28j@hotmail.com

Keywords: ecological security; land resource; classification; support vector machine

Abstract. Ecological security classification of land resources play an important role in sustainable utilization of land resources and improve benefit of healthy development of urbanization in China. According to the county level of ecological security classification of land resources data which is large scale and imbalance, this paper presented a support vector machine (SVM) model to classify the county level of ecological security of land resources. The method was compared with artificial neural network, decision tree, logistic regression, and naive Bayesian classifier regarding the county level of ecological security of land resources classification for Guanzhong urban agglomeration. It is found that the method has the best accuracy rate, hit rate, covering rate and lift coefficient, and provides an effective measurement for county level of ecological security of land resources classification and prediction.

Introduction

Ecological security is the basis of sustainable development. The county level of ecological security of land resources is an essential factor for regional sustainable development as well as for promoting the healthy development of urbanization. It is becoming clear that ecological security of land resources is the core and basic support of urbanization. Based on the development trends over the past decade or so, we can well expect that ecological security will make a positive effect to human life, health and regional sustainable development in the future [1]. A small improvement in county level of ecological security would yield a considerable increase in overall level of ecosystem integrity and healthy development of urbanization. Therefore, different promoting measures and supporting policies could be carried out for different county level of ecological security of land resources through classifying counties based on the predicted level of ecological security of land resources [2].

Data mining technology could detect important information behind large amounts of data by building county level of ecological security classification of land resources model and then work more effectively. In order to classify the county level of ecological security of land resources more effectively, domestic and foreign experts did numerous predictive modeling and have achieved a certain effect [3], but the accuracy rate of these models is not high enough.

According to regional ecological security of land resources dataset characteristic, this paper presented an support vector machine (SVM) model [4] and utilized county dataset provided by Guanzhong urban agglomeration of Shaanxi province China to make a county level of ecological security of land resources classification.

Research Methodology

The support vector machine (SVM) first proposed by Cortes and Vapnik (1995) is gaining popularity because of its excellent properties of high generalization performance and global optimal solution [5]. Not only its structure is simple, but also its various technical capabilities is obviously boosted, especially the generalization ability.

For given training set:

$$T = \{(x_1, y_1), \dots, (x_n, y_n)\} \quad (X \in Y)^n \quad x_i \in X = R^d, y_i \in Y = R, i=1, \dots, n$$

First, through nonlinear transform of $x \otimes j(x)$, the paper mapped the input space into Hilbert space [6], to construct the optimal linear function

$$f(x) = w \cdot j(x) + b. \quad (1)$$

Thus it could get the linear approximation in feature space, Vapnik advanced that could take ϵ -insensitive loss function as measurement of approximation.

$$L(x, y, f(x)) = |y - f(x)|_e = \max\{0, |y - f(x)| - \epsilon\}. \quad (2)$$

ϵ is a pre-selected permitted error and is positive. In interval $[-\epsilon, +\epsilon]$, there is not error between the classification function $f(x)$ and actual value y , the function loss “sensitivity”, so called as ϵ -insensitive loss function. According to the principle of structure risk minimization, $f(x)$ should minimize following formula (3) [7].

$$\min_{w \in R^n, b \in R} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n L(x_i, y_i, f(x_i)). \quad (3)$$

In order to measure the deviation degree of samples outside ϵ -insensitive area, the paper introduced two nonnegative relaxation variables x_i and x_i^* , and the support vector machine used in function approximation should be:

$$\min_{w \in R^n, x_i, x_i^* \in R^{2n}, b \in R} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (x_i + x_i^*). \quad (4)$$

$$\begin{aligned} & \sum_{i=1}^n y_i - f(x_i) \leq \epsilon + x_i \\ \text{s.t. } & \sum_{i=1}^n f(x_i) - y_i \leq \epsilon + x_i^* \\ & x_i, x_i^* \geq 0 \quad i = 1, 2, \dots, n \end{aligned}$$

Where, C is penalty parameter to control the penalty degree of samples which go beyond permitted error [8]. Minimizing formula (4) is a down convex quadratic optimization problem. Introducing Lagrange function, according to the meanings of Wolfe duality, we can get corresponding dual problem:

$$\min_{a, a^* \in R^{2n}} \frac{1}{2} \sum_{i,j=1}^n (a_i^* - a_i)(a_j^* - a_j)(x_i \cdot x_j) + \epsilon \sum_{i=1}^n (a_i^* + a_i) - \sum_{i=1}^n y_i(a_i^* - a_i). \quad (5)$$

$$\begin{aligned} & \sum_{i=1}^n (a_i - a_i^*) = 0 \\ \text{s.t. } & \sum_{i=1}^n 0 \leq a_i, a_i^* \leq \frac{C}{n}, \quad i = 1, 2, \dots, n \end{aligned}$$

Resolving formula (5) to get unique solution, i.e. optimal solution $\bar{a} = (\bar{a}_1, \bar{a}_1^*, \bar{a}_2, \bar{a}_2^*, \dots, \bar{a}_n, \bar{a}_n^*)^T$, so the solution of original problem formula (4) about w is:

$$\bar{w} = \sum_{i=1}^n (\bar{a}_i^* - \bar{a}_i) x_i. \quad (6)$$

Introducing Kernel function $K(x, x')$ into support vector machine to predigest nonlinear approximation, we can get decision function:

$$f(x) = \sum_{i=1}^n (\bar{a}_i^* - \bar{a}_i) K(x_i, x) + \bar{b}. \quad (7)$$

Kernel function should not only satisfy the Mercer condition, but also reflect the distribution characteristics of training sample sets in practical application. The selection of kernel function is a core issue of SVM theory. However, many SVM regression application studies are performed by ‘expert’ users having good understanding of SVM methodology. Since the quality of SVM models depends on a proper setting of SVM meta-parameters, the main issue for practitioners trying to apply SVM regression is how to set these parameter values (to ensure good generalization performance) for a given data set. Whereas existing sources on SVM regression give some recommendations on appropriate setting of SVM parameters, there is clearly no consensus and (plenty of) contradictory opinions [9]. Hence, resampling remains the method of choice for many applications. Unfortunately,

using resampling for (simultaneously) tuning several SVM regression parameters is very expensive in terms of computational costs and data requirements.

Empirical Study

Data Source In this paper, we take county dataset provided by Guanzhong urban agglomeration of Shaanxi province China as the core data, including 26 counties. The county level of ecological security of land resources indicators in this paper measure sixteen things: 1) per capita cultivated land; 2) cultivated land proportion; 3) forest cover rate; 4) the proportion of woodland area; 5) land grain yield; 6) land reserve resource rate; 7) per capita GDP; 8) rural per capita net income; 9) per capita disposable income of urban residents; 10) fertilizer consumption of per hectare cultivated area; 11) pesticide consumption of per hectare cultivated area; 12) natural population growth rate; 13) population density; 14) urbanization level; 15) per capita living space in rural areas; 16) per capita living space in urban areas.

The data sets in this paper are collected from 1993 to 2015. We constructed the data window from 1993 to 2009, take 2010 as the delay time, take data sets of per capita GDP, rural per capita net income, per capita disposable income of urban residents, fertilizer consumption of per hectare cultivated area, pesticide consumption of per hectare cultivated area, natural population growth rate, population density and urbanization level during the data window as input data, and take per capita cultivated land, cultivated land proportion, forest cover rate, the proportion of woodland area, land grain yield, land reserve resource rate, per capita living space in rural areas and per capita living space in urban areas from 2011 to 2015 as the output data. We can extract the average quotient parameter as the evaluation criteria for ecological security of land resources parameter. If the indicator from a county was less than average quotient parameter, then we call the county level of ecological security of land resources was low. And if the indicator from a county was more than average quotient parameter, then we call the county level of ecological security of land resources was high.

We select triple standard difference method and double standard deviation to reject abnormal data. Ultimately, we obtain 378 examples for training sets by sampling (including 174 high county level of ecological security of land resources and 204 low county level of ecological security of land resources), and 220 examples for test sets during the same period with training sets (including 101 high county level of ecological security of land resources and 119 low county level of ecological security of land resources).

Results and Discussion According to the analysis above, we constructed the samples (x, y) , where x is the input vector and y is the type of sample properties, for high county level of ecological security of land resources, $y=1$, and for low county level of ecological security of land resources, $y=-1$.

Through empirical analysis on MATLAB 2013, and based on the established routine: 1) representation; 2) fitness evaluation; 3) selection, crossover, and mutation operation; 4) evolutionary cycle [10], it is found that, the SVM model constructed by data sets considered Laplace kernel where parameter $u=0.09$, $C=12$ could reach an excellent performance (as can be seen in Table 1).

TABLE I. THE SVM PREDICTION RESULTS BASED ON DIFFERENT KERNEL FUNCTIONS

Kernel Function	Model parameter		Accuracy rate	Hit rate	Covering rate	Lift coefficient
	C	u				
RBF Kernel	2.3	1.09	0.6378	0.7149	0.2062	1.6748
Polynomial kernel	12	2.1	0.6153	0.7178	0.0191	1.5547
Laplace kernel	12	0.09	0.7405	0.7709	0.2038	1.7670
Cauchy kernel	19	0.18	0.6927	0.7651	0.1937	1.7607

The evaluation standard of the model can be shown in Table 2, accuracy rate= (A+D)/(A+B+C+D), hit rate = A/(A + C), covering rate = A/(A + B), lift coefficient= accuracy rate/ county level of ecological security of land resources rate in the data sets.

TABLE II. CLASSIFICATION MATRIX

County level of ecological security of land resources state	Predicted low county level of ecological security of land resources	Predicted high county level of ecological security of land resources
Real low county level of ecological security of land resources	A	B
Real high county level of ecological security of land resources	C	D

In order to take a contrast test, we selected the best parameter value and highest accuracy of predictive model, and utilized models including SVM, decision tree, logistic regression, naive Bayesian classifier and artificial neural network. Table 3 shows the result of classification.

TABLE III. THE COMPARISON OF SVM, C4.5, LOGISTIC REGRESSION, NAIVE BAYESIAN CLASSIFIER AND ANN

Method	Accuracy rate	Hit rate	Covering rate	Lift coefficient
SVM	0.7405	0.7709	0.2038	1.7670
C4.5	0.5386	0.6245	0.4155	1.4936
Logistic regression	0.5939	0.6784	0.1861	1.6699
Naive Bayesian classifier	0.6370	0.7206	0.0134	1.5206
ANN	0.6369	0.7822	0.0304	1.6146

The decision tree model utilized the popularized C4.5; the decision surface of logistic regression is $\frac{1}{1 + \exp\left(-\beta_0 - \sum_{i=1}^d \beta_i x_i\right)} - \frac{1}{2} = 0$, where β_0 is the constant term, d is the dimension, β_i is the regression coefficient, indicates the contribution of x_i to decision surfaces. The discriminating surface of naive Bayesian classifier is $p(x|w_i)P(w_i) - p(x|w_j)P(w_j) = 0$, while $p(x/w)$ is conditional probability, and $P(w)$ is prior probability; ANN model used a nonlinear model with single hidden layer, the condition to stop is the error has arrived local minimum [11].

As shown in Table 3, the accuracy rate, the hit rate, the covering rate and the lift coefficient of SVM model is 0.7405, 0.7709, 0.2038 and 1.7670 respectively. Except the hit rate is slightly lower than ANN and the covering rate is lower than decision tree (C4.5), other indices are all much higher than the predicted results of other models. The covering rate of ANN model is 0.0304, it indicates that the model is over-fitted; the unusual samples of data sets affect the decision tree (C4.5) model least. Higher accuracy shows a better integrated predictive capacity of SVM. The ideal hit rate, covering rate and lift coefficient indicate that, under different foundations of county sustainable development, SVM is capable of retaining safer counties on ecological security which are willing to have higher county level of ecological security of land resources. Consequently, when selecting appropriate kernel function and parameters, SVM would have an excellent performance in county level of ecological security of land resources classification.

Conclusion

In this paper, we applied SVM to the classification of county level of ecological security of land resources, by comparing with artificial neural network (ANN), decision tree (C4.5), logistic regression and naive Bayesian classifier as a benchmark. The results demonstrated that, from a methodological perspective, SVM had the characteristics of simple classification surface, high generalization

performance and high fitting accuracy, etc. From a perspective of conditions and structure of data, under conditions that has rich samples with plenty of support vectors, abundant indices, and larger probability of county level of ecological security of land resources in the samples, SVM had a high precision for predicting. Classifying and predicting the county level of ecological security of land resources according to SVM will provide a guide for the research on regionally sustainable utilization of land resources in China.

Acknowledgment

This work was financially supported, in part, by the National Social Science Foundation of China (Grant No. 16CJL051), the Soft Science Research Program of Shaanxi Province (Grant No. 2017KRM174), the Foundation of Shaanxi Educational Committee (Grant No. 17JZ053) and the Shaanxi provincial philosophy and social science characteristic discipline (Grant No. 107-5X1203).

References

- [1] Q. Lin, J. Mao, J. Wu, W. Li, and J. Yang, "Ecological Security Pattern Analysis Based on InVEST and Least-Cost Path Model: A Case Study of Dongguan Water Village," *Sustainability*, vol. 8, 2016, pp. 172-187.
- [2] H. S. Chen, "The Establishment and Application of Environment Sustainability Evaluation Indicators for Ecotourism Environments," *Sustainability*, vol. 7, 2015, pp. 4727-4746.
- [3] D. B. P. J Van, "Predictive Modeling of Ecological Risk from Agrochemicals to Non-target Organisms in Tropical Ecosystems," *Journal of Pesticide Science*, vol. 36, 2011, pp. 319-330.
- [4] S. Johnson, V. Shanmugam, "Feature Subset Selection for Hot Method Prediction using Genetic Algorithm wrapped with Support Vector Machines," *Journal of Computer Science*, vol. 7, 2011, pp. 707-714.
- [5] C. Cortes, V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, 1995, pp. 273-297.
- [6] S. J. Qin, T. A. Badgwell, "A Survey of Industrial Model Predictive Control Technology," *Control Engineering Practice*, vol. 11, 2003, pp. 733-764.
- [7] J. M. Zamarreño, P. Vega, "Neural Predictive Control: Application to a Highly Non-linear System," *Engineering Applications of Artificial Intelligence*, vol. 12, 1999, pp. 149-158.
- [8] A. Boukabou, A. Chebbah, N. Mansouri, "Predictive Control of Continuous Chaotic Systems," *International Journal of Bifurcation & Chaos*, vol. 18, 2008, pp. 587-592.
- [9] R. M. Balabin, E. I. Lomakina, "Support Vector Machine Regression (LS-SVM)– an Alternative to Artificial Neural Networks (ANNs) for the Analysis of Quantum Chemistry Data?" *Physical Chemistry Chemical Physics Pccp*, vol. 13, 2011, pp. 11710-11718.
- [10] G. Georgoulas, D. Gavrilis, I. G. Tsoulos, C. Stylios, J. Bernardes, and P. P. Groumpos, "Novel approach for fetal heart rate classification introducing grammatical evolution," *Biomedical Signal Processing & Control*, vol. 2, 2007, pp. 69-79.
- [11] R. Moraes, J. F. Valiati, W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications*, vol. 40, 2013, pp. 621-633.