Data Prediction Based on Support Vector Machine (SVM)—Taking Soil Quality Improvement Test Soil Organic Matter as an Example

NIU Yan¹,²,³,⁴, YE Shenglan¹,²,³,⁴

¹Institute of Land Engineering and Technology, Shaanxi Provincial Land Engineering Construction Group Co., Ltd. Xi'an 710075;  
²Shaanxi Provincial Land Engineering Construction Group Co., Ltd. Xi'an 710075;  
³Key Laboratory of Degraded and Unused Land Consolidation Engineering, the Ministry of Land and Resources Xi'an 710075;  
⁴Shaanxi Provincial Land Consolidation Engineering Technology Research Center Xi'an 710075.

Corresponding author’s e-mail address:2330517641@qq.com

Abstract: Support Vector Machine (SVM) is a machine learning language based on statistical learning theory, mainly used for data classification and regression analysis. Taking the soil quality improvement test soil sample organic matter data as an example, the support vector machine is used to train and predict the data, and the relative error between the predicted value and the actual sample value is analyzed to verify the support vector machine data prediction in the field of land engineering. Operability, pointing out the inadequacies, in order to provide reference for relevant data analysis.

1. Preface
One of the most cutting-edge researches in artificial intelligence that best represent intelligent features is machine learning. Different from the traditional statistics based on the gradual theory, the data-based statistical learning simulates the ability of humans to learn and generalize from practice. It mainly studies the methods from the observation data that can not be obtained through the principle analysis, and will get the law applies to the analysis of objective phenomena, predicting and judging unknown data or new data that cannot be observe⁴. In the 1990s, Russian mathematician Vapnik proposed the concept of Support Vector Machines (SVM): Support Vector Machines is based on statistical learning theory, and structural risk minimization principle to establish a data model⁵, to solve Statistical pattern recognition in the case of limited data samples lays a solid foundation. Compared with traditional machine learning methods, this method has many advantages such as simple structure, good adaptability, global optimization, fast training speed and strong generalization ability⁶.

In this paper, the support vector machine is applied to the soil quality improvement test data analysis, and the data is trained and predicted to verify the operability of the support vector machine in the field of land engineering, and provide reference for the land engineering scientific research data processing.
2. Support vector machine principle
Support Vector Machine is a data-based machine learning method developed according to statistical learning theory\cite{4}, which is based on the principle of structural risk minimization\cite{5}. The support vector machine can maximize the promotion ability of the learning machine. Even if the collected discriminant function is based on limited data, the prediction error of the independent test set can still be small. In addition, the support vector machine is a convex quadratic optimization problem, which can guarantee that the obtained extremum solution is a global optimal solution. These two characteristics make the support vector machine an excellent machine learning algorithm. Support vector machine is the latest and most practical part of statistical theory. Its main content was basically completed in 1992-1995, and it was still in the stage of continuous development\cite{1}. It can be said that the statistical learning theory has received more and more attention since the 1990s, largely because of the development of support vector machines\cite{6}.

The success of SVM includes two key technologies. On the one hand, it solves the problem of data classification, and uses the support vector machine to obtain the most classified surface, so that the most classified surface has the most interval. On the other hand, to solve the regression problem, the kernel function method of support vector machine is used to obtain the linear learning algorithm which can replace the nonlinear transformation. The key technology of the support vector machine used in this paper is the application of regression problem\cite{7}.

3. Support vector machine implementation
LIBSVM is a simple, easy-to-use and fast and effective SVM pattern recognition and regression software package developed by Professor Lin Zhiren of Taiwan University. It not only provides compiled executable files for Windows series systems, but also provides source code. It is easy to improve, modify and apply on other operating systems; the software adjusts the parameters involved in SVM relatively, provides a lot of default parameters, can solve many problems by using these default parameters; and provides interactive check function. This paper uses Matlab to implement the computer of support vector machine.

The general steps used by LIBSVM are:\(1\) Prepare the data set in the format required by the LIBSVM software package;\(2\) Perform a simple scaling operation on the data;\(3\) Consider the selection of the RBF kernel function;\(4\) Using cross-validation to select the best parameters C and g;\(5\) Using the best parameters C and g to train the entire training set to obtain the support vector machine model;\(6\) Test and predict using the acquired model.

4. Data Pre-learning Training and Prediction Based on Support Vector Machine

4.1 Data Sources
The soil quality improvement test was carried out in the experimental field of Fuping pilot test base in 2017. In order to study the practical significance of implementing straw returning, adding organic fertilizer and less tillage and no-tillage measures, the organic fertilizer fertilization mode was mainly used for treatment. Sub-treatment design test to study the changes of soil quality under different tillage treatments and different fertilization levels and the impact on crop production performance, so as to provide scientific basis for screening optimal fertilization and tillage combination, and achieve soil quality improvement and crop production performance. Upgrade. At the end of the experiment, soil samples were collected at different depths of the soil, and the organic matter content of the soil samples was determined by oil bath method. The content of organic matter indicators in each test treatment is shown in Table 2.

4.2 Learning sample and prediction sample determination
According to the soil quality improvement test setup, the tillage method, fertilization level and sampling depth are determined as variables. The farming methods include tillage, deep pine, no-tillage, tillage-deep pine, and the fertilization level includes high (1000 kg•Hm\(^{-2}\)), medium (625 kg•hm\(^{-2}\)), low
(375 kg·hm⁻²) three treatments, sampling depth including 0–20 cm, 20–40 cm, 40–60 cm. The Matlab implementation of SVM training for data needs to encode the learning samples. The specific coding method is shown in Table 1.

| Coding | Tillage | deep pine | no-till | tillage-deep pine |
|--------|---------|-----------|---------|-------------------|
| 1      |         |           |         |                   |
| 2      |         |           |         |                   |
| 3      |         |           |         |                   |
| 4      |         |           |         |                   |

Based on the coding method of Table 1, the test scheme coded as 1 1 1 is that the tillage method is tillage, the fertilization level is high, and the sampling depth is 0-20 cm. The corresponding coding of all 36 experimental schemes is shown in Table 2. Six matrices were selected using the Matlab random function as prediction samples, and the remaining 30 schemes were learning samples. In order to avoid the influence of dimension on data prediction, it is necessary to normalize the learning sample data. There are many methods of normalization. This paper uses linear function transformation. The formula is

\[ y = \frac{x - \text{MinValue}}{\text{MaxValue} - \text{MinValue}} \]

Where: x and y are the values before and after the conversion, and MaxValue and MinValue are the maximum and minimum values of the sample, respectively.

The learning samples, prediction samples and normalized data are shown in Table 2.

| Test treatment | coding | Organic matter (g/kg) | Normalized data | Test treatment | coding | Organic matter (g/kg) | Normalized data |
|---------------|--------|-----------------------|-----------------|---------------|--------|-----------------------|-----------------|
| D-H 0-20 cm   | 1 1 1  | 9.13                  | 0.57            | T-H 0-20 cm   | 3 1 1  | 12.54                 | 1.00            |
| D-H 20-40 cm  | 1 1 2  | 9.47                  | 0.62            | T-H 40-60 cm  | 3 1 3  | 4.52                  | 0.00            |
| D-H 40-60 cm  | 1 1 3  | 6.05                  | 0.19            | T-M 0-20 cm   | 3 2 1  | 10.97                 | 0.80            |
| D-M 0-20 cm   | 1 2 1  | 9.48                  | 0.62            | T-M 20-40 cm  | 3 2 2  | 8.93                  | 0.55            |
| D-M 20-40 cm  | 1 2 2  | 9.73                  | 0.65            | T-L 0-20 cm   | 3 3 1  | 10.46                 | 0.74            |
| D-L 0-20 cm   | 1 3 1  | 8.93                  | 0.55            | T-L 20-40 cm  | 3 3 2  | 8.91                  | 0.55            |
| D-L 40-60 cm  | 1 3 3  | 10.12                 | 0.70            | T-L 40-60 cm  | 3 3 3  | 7.27                  | 0.34            |
| N-H 0-20 cm   | 2 1 1  | 11.85                 | 0.91            | DT-H 0-20 cm  | 4 1 1  | 9.49                  | 0.62            |
| N-H 40-60 cm  | 2 1 2  | 8.50                  | 0.50            | DT-H 20-40 cm | 4 1 2  | 9.41                  | 0.61            |
| N-H 40-60 cm  | 2 1 3  | 6.64                  | 0.26            | DT-M 0-20 cm  | 4 2 1  | 9.51                  | 0.62            |
| N-M 0-20 cm   | 2 2 1  | 11.00                 | 0.81            | DT-M 20-40 cm | 4 2 2  | 8.78                  | 0.53            |
| N-M 40-60 cm  | 2 2 2  | 7.46                  | 0.37            | DT-M 40-60 cm | 4 2 3  | 6.41                  | 0.24            |
| N-L 0-20 cm   | 2 3 1  | 7.86                  | 0.42            | DT-L 0-20 cm  | 4 3 1  | 7.99                  | 0.43            |
| N-L 20-40 cm  | 2 3 2  | 7.94                  | 0.43            | DT-L 20-40 cm | 4 3 2  | 7.95                  | 0.43            |
| N-L 40-60 cm  | 2 3 3  | 9.00                  | 0.56            | DT-L 40-60 cm | 4 3 3  | 9.85                  | 0.66            |

| Forecast sample | coding | Organic matter (g/kg) | Normalized data | Test treatment | coding | Organic matter (g/kg) | Normalized data |
|-----------------|--------|-----------------------|-----------------|---------------|--------|-----------------------|-----------------|
| D-M 40-60 cm    | 1 2 3  | 8.50                  | /               | T-H 20-40 cm  | 3 1 2  | 7.92                  | /              |
| D-L 20-40 cm    | 1 3 2  | 9.43                  | /               | T-M 40-60 cm  | 3 2 3  | 5.72                  | /              |
| N-M 20-40 cm    | 2 2 2  | 11.62                 | /               | DT-H 40-60 cm | 4 1 3  | 13.60                 | /              |

Note: D stands for Deep pine; N stands for No-till; T stands for Tillage; DT stands for Deep pine Tillage; H stands for high fertilization; M stands for medium fertilization; L stands for low fertilization.
4.3 Data training learning and prediction

Before applying the support vector machine to train the learning samples, the particle swarm optimization algorithm is used to search and optimize the parameters of the support vector machine, and the optimal parameters of the support vector machine are obtained. The optimization process is implemented in Matlab. Figure 1 shows the optimization. The fitness curve of the process, the optimal parameters obtained are \( C=4.7682, g=1.4134 \).

![Particle swarm optimization algorithm parameter fitness curve](image)

Using the optimal support vector machine parameters obtained by particle swarm optimization, the learning samples are trained and learned, and the relationship model between farming mode, fertilization level, soil sample depth and organic matter content normalized value is obtained. The predicted sample is predicted, and the obtained predicted value is inverse normalized to obtain a predicted value of the organic matter content, and the predicted value is compared with the value obtained by the actual sample. The data comparison is shown in Table 3.

| Test treatment | coding | Organic matter (g/kg) | Non-optimized parameter predictions | Anti-normalized value | Relative error | Optimized parameter prediction | Anti-normalized value | Relative error |
|----------------|--------|-----------------------|-------------------------------------|-----------------------|----------------|-------------------------------|-----------------------|----------------|
| D-M 40-60 cm   | 1 2 3  | 8.5                   | 0.5219                              | 8.705638              | 2.42%          | 0.4823                        | 8.388046              | 1.32%          |
| D-L 20-40 cm   | 1 3 2  | 9.43                  | 0.5456                              | 8.895712              | 5.67%          | 0.5671                        | 9.068142              | 3.84%          |
| N-M 20-40 cm   | 2 2 2  | 11.62                 | 0.5407                              | 8.856414              | 23.78%         | 0.7261                        | 10.34332              | 10.99%         |
| T-H 20-40 cm   | 3 1 2  | 7.92                  | 0.5328                              | 8.793056              | 11.02%         | 0.5147                        | 8.647894              | 9.19%          |
| T-M 40-60 cm   | 3 2 3  | 5.72                  | 0.4662                              | 8.258924              | 44.39%         | 0.2709                        | 6.692618              | 17.00%         |
| D-T 40-60 cm   | 4 1 3  | 13.6                  | 0.4906                              | 8.454612              | 37.83%         | 0.3559                        | 7.374318              | 45.78%         |

Note: D stands for Deep pine; N stands for No-till; T stands for Tillage; DT stands for Deep pine Tillage; H stands for high fertilization; M stands for medium fertilization; L stands for low fertilization.

It can be seen from Table 3 that among the six prediction schemes, the relative error between the predicted value of the scheme 1 2 3, 1 3 2 and the actual sample value is small, and the relative errors when the parameters are not optimized are 2.42% and 5.67%, respectively. The relative error after optimizing the parameters is 1.32% and 3.84%, respectively. The relative error between the predicted value and the actual sample after the optimization parameters of the scheme 2 2 2 and 3 1 2 is about...
10%. The relative error between the predicted value of the scheme 3 2 3 and 4 1 3 and the actual sample value is slightly larger. It shows that the application support vector machine can predict the data well after training samples, but the error of individual schemes is large. Comparing the relative error between the unoptimized parameters and the relative error of the optimized parameters, it is found that except for the scheme 4 1 3, the relative errors between the predicted values and the actual values of the optimized parameters are reduced, indicating that the particle swarm optimization algorithm is applied. Vector machine parameters are necessary for optimization.

5. Summary and discussion
This paper first introduces the principle of support vector machine and computer implementation method. Secondly, taking the soil quality improvement test soil sample organic matter data as an example, the training sample and the prediction sample are determined. The support vector machine is used to train the learning and prediction, and the prediction error is analyzed. The results show that, besides the individual schemes, the SVM can be used to predict the data after training the learning samples, and it is necessary to use the particle swarm optimization algorithm to optimize the parameters of the support vector machine.

The variables considered in this paper include farming mode, fertilization level and soil sample depth. However, the factors affecting the organic matter content of soil samples are not limited to these three variables. The factors affecting the prediction indicators should be considered as much as possible, and the training learning samples should be added, which can greatly improve the accuracy of data prediction.

References
[1] Bai Peng. Support vector machine theory and engineering application examples [M]. Xi'an: Xi'an University of Electronic Science and Technology Press. 2008.
[2] Yang Zhongwei. The ruin prediction of support vector machine based on particle swarm optimization and genetic algorithm[J]. Computer Engineering and Applications, 2013, 49(18): 265-270.
[3] Chen Tao. Selective Support Vector Machine Integration Algorithm[J]. Computer Engineering and Design, 2011, 32(05): 1807-1809+1819.
[4] Lei Jian. Algorithm and application research of RBF neural network based on support vector machine [D]. Jiangxi University of Technology, 2009.
[5] Gao Wei. Research on incremental learning algorithm of support vector machine [D]. Liaoning University of Science and Technology, 2018.
[6] Sun Yu. Support vector machine research and its application to electric field optimization of UHV transformers [D]. Hebei University of Technology, 2011.
[7] Wu Jiahui. Process support vector machine and its application in satellite thermal equilibrium temperature prediction [D]. Harbin Institute of Technology, 2008.