A Crop Disease Recognition Algorithm Based on Machine Learning

Yuchao Zhou, Kailiang Zhang(✉), Yi Shi, and Ping Cui

Jiangsu Province Key Laboratory of Intelligent Industry Control Technology, Xuzhou University of Technology, Xuzhou 221018, China
zhangkailiang@xzit.edu.cn

Abstract. There are many related diseases in the process of crop planting, which reduces the quality and yield of crops. Faced with such a situation, the prevention of crop diseases has become a hot spot and has broad application prospects. This experiment uses the image recognition technology of machine vision to analyze and recognize crop diseases. Based on the features of machine vision that can capture details that cannot be observed by the human eye, with high accuracy and high efficiency, it provides accurate image recognition of crop diseases. In accordance with. In the process of selecting the SVM classifier for image classification, the kernel function and gamma parameters in the classifier were adjusted, and the kernel function and high accuracy parameter interval suitable for crop disease analysis were found.

Keywords: Machine learning · Crop disease recognition · Support vector machine

1 Introduction

Our country is the country with the largest population in the world and a large agricultural country. The quality and output of crops are related to my country's economic lifeline and the people's living foundation. However, crop diseases will lead to a decline in plant viability, thereby affecting yield. Traditional crop diseases mainly rely on manual field identification, which not only takes time and effort, but also has low accuracy. Without accurate and quantitative standards, manual identification requires accumulation of experience, and often cannot be accurate in time when crops have problems. To judge and deal with different situations locally, these shortcomings have seriously affected the development of agricultural modernization [1, 2]. As a relatively mature advanced technology, machine learning technology, combined with agricultural knowledge, uses image recognition technology to make accurate judgments of crop diseases, which can then be controlled and processed to increase crop yields [3–5].

SVM classifier is used as the main classification technology for image recognition. The investigation found that due to the huge difference between the image features automatically extracted by the computer and the semantics understood by humans, the

D. Jiang and H. Song (Eds.): SIMUtools 2021, LNICST 424, pp. 513–522, 2022.
https://doi.org/10.1007/978-3-030-97124-3_38

results of image retrieval are unsatisfactory. Relevant feedback methods have appeared in recent years [6]. Using SVM as the classifier, the positive and negative samples marked by the user are learned in each feedback, and retrieval is performed according to the learned model. Relevant researchers used an image library composed of 9918 images to conduct experiments, and the results showed that this method has a good generalization function in the case of limited training samples [7]. Because SVM is analyzed for the linearly separable situation. For the case of linear inseparability, the non-linearly inseparable samples in the low-dimensional input space are converted into high-dimensional feature spaces by using the nonlinear mapping algorithm to make them linearly separable, so that the linear algorithm is nonlinear to the samples in the high-dimensional feature space [8, 9]. It is possible to perform linear analysis of features. And based on the theory of structural risk minimization, it constructs the optimal classification surface in the feature space, so that the learner can get the global optimization, and the expected risk of the entire sample space meets a certain upper bound with a certain probability [10]. From the above two basic ideas, SVM does not use the traditional derivation process, which simplifies the usual classification and regression problems: a small number of support vectors determine the final decision function of the SVM, and the complexity of the calculation depends on the support vector, and Instead of the entire sample space, this can avoid the "curse of dimensionality". A small number of support vectors determine the final result, which not only helps us to grasp the key samples, but also destined that the method is not only simple in algorithm, but also has good "robustness".

In recent years, there have been some researches on crop disease identification. Literature [11] uses a variety of plant diseases and pests for analysis and research, and expands the practical field by modifying model parameters, saving time. However, the accuracy needs to be improved. Literature [12] studies image processing techniques for identifying and classifying symptoms of fungal diseases on different agricultural and horticultural crops. However, the early detection and classification of fungal diseases and their related symptoms need to be improved. Literature [13] studies, trains and tests a deep learning model with high accuracy. However, object-oriented is relatively large, and its wide popularization needs to be studied. In reference [14], a mechanism for dynamic analysis of disease images is provided, which is fast. However, the actual accuracy needs to be further tested. In literature [15], a customized efficient memory convolution neural network is proposed for automatic detection of rice grain diseases, with good classification and accuracy. However, the applicability of other crops needs to be studied. Some authors have also studied the expert system for identifying crop diseases [16] and the grading system for identifying diseases on a given image set [17], and achieved some results. However, the relevant model mechanism needs to be further improved. Some authors have proposed algorithms and mechanisms in the application of convolutional neural network to automatic identification of crop diseases [18, 19], but the accuracy and classification level need to be improved. During the image recognition process, it was found that different kernel function selections in SVM have an impact on feature description, and the relevant gamma parameters in the kernel function were experimentally studied. It was found that the RBF kernel function training obtained the highest matching degree in crop disease recognition [20]. And find out the parameter interval with high matching degree among the gamma parameters.

2 System Model

The experiment chooses the SVM classifier, which can segment the positive and negative examples in the sample set by finding the hyperplane of the sample set. SVM is developed from the optimal classification surface in the case of linear separability. The problems that can be solved are linear separable, approximately linear separable and nonlinear separable. Linear separability means that two types of samples can be completely separated with a linear function; in the case of non-linear separability, an adjustable error penalty coefficient c will be introduced to find the best classification hyperplane. However, in many classification situations, linear classifiers have limitations. Relaxation of constraints for nonlinear problems is prone to a large number of sample misclassification errors. Therefore, it is necessary to map the nonlinear problem to a linear in a high-dimensional space through nonlinear mapping. Improvement of separable problems. Therefore, the SVM with kernel function mapping is introduced SVM [21]. The four kernel function formulas of the classifier are as follows.

Linear function formula:

$$\mathrm{K(x, y) = x \cdot y} \tag{1}$$

Polynomial kernel function:

$$\mathrm{K(x, y) = (\gamma(x \cdot y) + r)^d, d = 1, 2, \cdots, N} \tag{2}$$

Radial basis (RBF) kernel function formula:

$$\mathrm{K(x, y) = exp\left(-\gamma\|x - y\|^2\right)} \tag{3}$$

Sigmoid kernel function formula:

$$\mathrm{K(x, y) = tanh(\gamma(x \cdot y) + r)} \tag{4}$$

Since the image recognition problem in this experiment is a two-classification problem, briefly describe the two-class SVM algorithm. The image sample training set is $D = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_n, y_n)\}$, where $x_i \in \chi = R^d$, $y_i \in \gamma = \{-1, +1\}$, n are sample dimensions, d is the sample dimension, $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$ is the i-th sample, and y_i is the corresponding label:

Define hyperplane formula:

$$\omega^T + b = 0 \tag{5}$$

Where $\omega = (\omega_1, \omega_2, \ldots, \omega_d)^T$ is the normal vector of the hyperplane, b is the intercept. The classification decision function:

$$f(x_i) = sign\left(\omega^T x_i + b\right) \tag{6}$$

Where *sign*() is the symbolic function:

Support vector machines are divided into two categories: linear and nonlinear. When linearly separable, the algorithm is as follows:

Normalize, set constraints:

$$y_i\left(\omega^T x_i + b\right) \geq 1, i = 1, \ldots, n \tag{7}$$

Maximize the distance between classes is $\frac{2}{||\omega||}$:

$$\begin{aligned} &\underset{\omega,b}{min} \qquad \frac{1}{2}||\omega||^2 \\ &subject\ to\ y_i\left(\omega^T x_i + b\right) \geq 1, i = 1, \ldots, n \end{aligned} \tag{8}$$

Construct Lagrangian function, construct and solve convex quadratic programming problem:

$$L(\omega, b, a) = \frac{1}{2}\|\omega\|^2 + \sum\nolimits_{i=1}^{n} \alpha_i\left[1 - y_i\left(\omega^T x_i + b\right)\right] \tag{9}$$

Where $\alpha_i \geq 0$ is the Lagrange multiplier.

Take the derivative of $L(\omega, b, \alpha)$ with respect to ω and b and assign it to 0:

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \\ \frac{\partial L}{\partial b} = 0 \end{cases} \rightarrow \begin{cases} \omega = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{10}$$

Finally, organize (4) into:

$$\begin{aligned} &\underset{\alpha}{max} \quad \sum\nolimits_{i=1}^{n} \alpha_i - \frac{1}{2}\sum\nolimits_{i=1}^{n} \alpha_i\alpha_j y_i y_j x_i^T x_j \\ &subject\ to\ \sum\nolimits_{i=1}^{n} \alpha_i y_i = 0 \\ &\alpha_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{11}$$

Formula (7) has a unique solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$, get $\omega^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$ and $b^* = y_i - \sum_{i=1}^{n} \alpha_i^* y_i x_i^T x_j$

Get the final classification function:

$$f(x) = sign\left[\left(\left(\omega^*\right)^T x + b^*\right)\right] = sign\left(\sum\nolimits_{i=1}^{*} \alpha_i^* y_i x_i^T x + b^*\right) \tag{12}$$

For nonlinear problems, nonlinear mapping is introduced to solve them. At this time, the sample does not satisfy the constraint conditions in the case of linear separability. The new constraint conditions are as follows:

$$y_i\left(\omega^T x_i + b\right) \geq 1 - \xi_i, i = 1, \ldots, n \tag{13}$$

Constrained optimization:

$$\begin{aligned} &\underset{\phi,b}{min} \quad \frac{1}{2}||\omega||^2 + C\sum\nolimits_{i=1}^{n} \xi_i \\ &subject\ to\ \ y_i\left(\omega^T x_i + b\right) \geq 1 - \xi_i, \\ &\xi_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{14}$$

Among them, $C > 0$ is the penalty parameter, the larger the C value, the greater the cost.

Construct Lagrangian function, construct and solve convex quadratic programming problem, after introducing nonlinear mapping, the final classification function is:

$$\begin{aligned} \max_{\alpha} \quad & \sum\nolimits_{i=1}^{n} \alpha_i - \frac{1}{2}\sum\sum \alpha_i\alpha_j y_i y_j K\left(x_i, x_j\right) \\ \text{subject to} \quad & \sum\nolimits_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C, i = 1, 2, \ldots, n \end{aligned} \tag{15}$$

In the low-dimensional space, the inner product operation is completed through the kernel function, and the nonlinear classification function is obtained:

$$f(x) = sign\left[\left(\left(\omega^*\right)^T \phi(x) + b^*\right)\right] = sign\left(\sum\nolimits_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\right) \tag{16}$$

3 Parameter Optimization

3.1 Key Parameter Analysis

The Gamma parameter is a parameter that comes with the function after selecting the RBF function as the kernel. It implicitly determines the distribution of the data after mapping to the new feature space. The larger the gamma, the fewer the support vectors, and the smaller the gamma value, the more support vectors. The number of support vectors affects the speed of training and prediction.

3.2 Optimization Experiment

We selected the diseases that tomatoes in the crops may suffer from during the cultivation process and tested them. Using the collected tomato diseases as a data set, we conducted training tests on different kernel functions, predicted the same disease map, and recorded the matching of each category. The greater the matching degree, the more consistent the feature points are (Tables 1, 2, 3, 4 and 5).

Table 1. Umbilical rot test data sheet

Serial number	Kernel function	Hollow fruit	Umbilical rot	Botrytis	Soft rot	Leaf mold	Result
1	LINEAR	−1.03413	−0.986507	−0.999999	−1.00081	−1	Correct
2	POLY	−1.0127	−0.99561	−1.00193	−1	−1.0021	Correct
3	RBF	−1.06445	−0.973202	−1	−1.00163	−1	Correct
4	SIGMOID	−0.97682	−1.02455	−0.985664	−1.01534	−0.987585	Wrong

Table 2. Leaf mold test data sheet

Serial number	Kernel function	Hollow fruit	Umbilical rot	Botrytis	Soft rot	Leaf mold	Result
1	LINEAR	−1.00965	−1.00383	−0.997917	−0.999992	−0.99505	Correct
2	POLY	−1.00345	−1.00166	−0.998995	−1	−0.998132	Correct
3	RBF	−1.01885	−1.00755	−0.996097	−0.999981	−0.990068	Correct
4	SIGMOID	−1.06562	−1.00243	−1.03528	−1.03178	−1.0364	Wrong

Table 3. Gray mold test data sheet

Serial number	Kernel function	Hollow fruit	Umbilical rot	Botrytis	Soft rot	Leaf mold	Result
1	LINEAR	−1.02671	−1.00132	−0.993519	−0.999949	−1.00033	Correct
2	POLY	−1.00978	−1.00077	−0.998458	−1	−1.00023	Correct
3	RBF	−1.0516	−1.00252	−0.986959	−0.999988	−1.00067	Correct
4	SIGMOID	−1.00465	−0.998737	−1.01102	−1.01893	−1	Wrong

Table 4. Hollow fruit test data sheet

Serial number	Kernel function	Hollow fruit	Umbilical rot	Botrytis	Soft rot	Leaf mold	Result
1	LINEAR	−0.993842	−0.997809	−1.00205	−1.00083	−1.00017	Correct
2	POLY	−0.999627	−0.998101	−1.00121	−1.00096	−1.00084	Wrong
3	RBF	−0.988151	−0.995578	−1.00411	−1.00196	−1.00027	Correct
4	SIGMOID	−1.26598	−1.04431	−1.04837	−1.08427	−1.069	Wrong

Table 5. Soft rot test data sheet

Serial number	Kernel function	Hollow fruit	Umbilical rot	Botrytis	Soft rot	Leaf mold	Result
1	LINEAR	−1	−1.00003	−0.999999	−0.998154	−1.00038	Correct
2	POLY	−1	−1	−1	−0.999891	−1.00006	Correct
3	RBF	−1	−0.999993	−1	−0.996123	−1.00043	Correct
4	SIGMOID	−1.1057	−1.03209	−1.03451	−1.08342	−1.04455	Wrong

From the results of testing and comparing the records of five different diseases, it is found that the kernel function LINEAR and the kernel function SIGMOID are still lacking in the accuracy rate, and the kernel function POLY and the kernel function RBF have a certain degree of accuracy. By comparing the matching degree of the kernel function POLY and the kernel function RBF in each disease, it is found that the matching degree obtained by using the RBF kernel function will be relatively accurate.

Then, in the research, it was found that the gamma parameter in SVM has a certain influence on the matching degree of RBF, and the parameter was tested and researched. The result of the research is shown in Fig. 1.

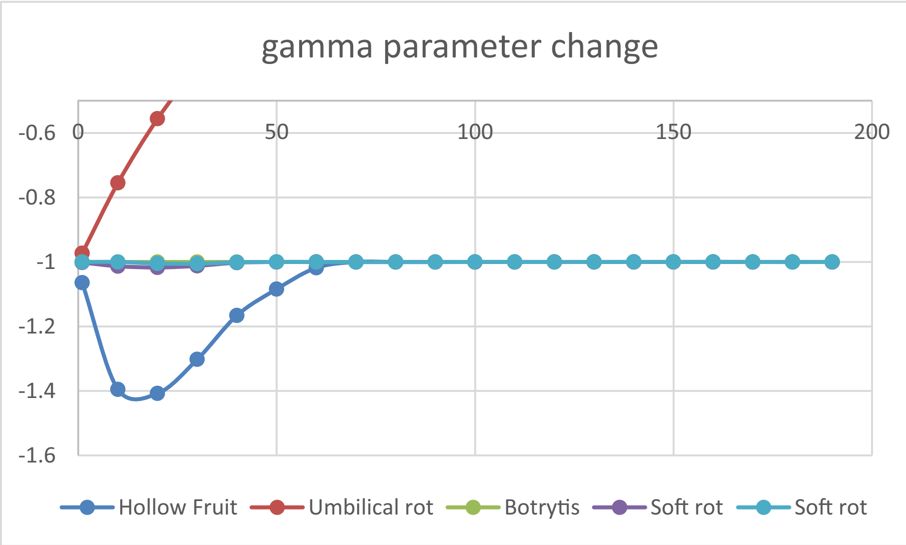

Fig. 1. The gamma parameter takes 1 as the growth gradient test chart

The selected research test picture is the tomato umbilical rot map. After a large number of data tests, it is found that the matching degree of this category is steadily increasing, and there is a significant gap with other categories. Through this test result, it is found that the greater the gamma parameter, the higher the matching degree. It can be inferred that when the gamma parameter approaches a certain value, the recognition result will be the most accurate. Furthermore, large-parameter research on gamma parameters is carried out, and the test data diagram is shown in Fig. 2.

Through the test, it is found that the larger the gamma parameter, the higher the matching degree, which is the same as the guess, but the improvement of the matching degree is not obvious when the parameter increases, and the gradual smoothing no longer has a significant improvement. The test found that the best parameter interval is 140 to 160. The gamma parameter in this interval will make the recognition of image recognition reach the best state, and it is found that there is a clear difference compared to other categories.

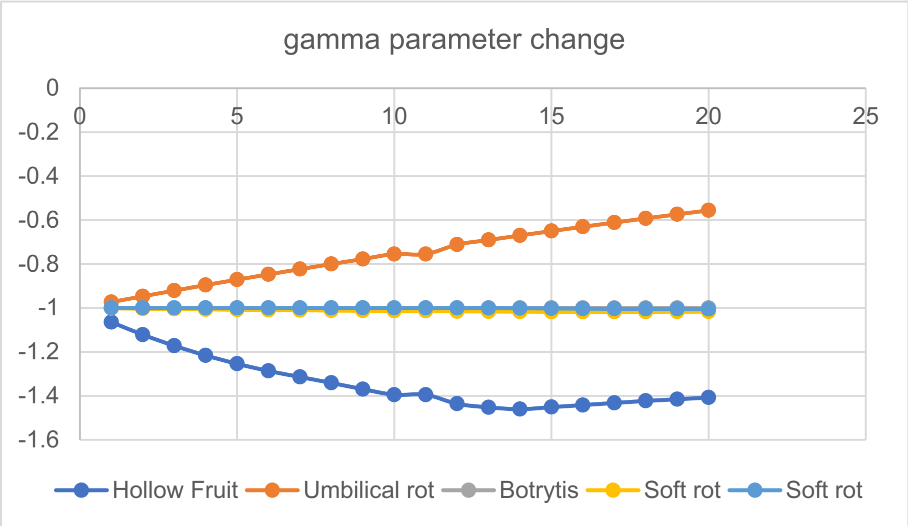

Fig. 2. The gamma parameter uses 10 as the growth gradient test chart

In the tomato disease studied, because most of the feature points are not linearly separable, the SVM classifier is not ideal for training and testing under the kernel function LINVAR, and when the kernel function RBF is selected Training, the matching degree has been improved to a large extent, and the accuracy of image recognition is also guaranteed.

4 Conclusion

In the paper, by adjusting the parameters and observing, we found that different kernel functions have an impact on the feature description of crop disease when training the model of the SVM classifier. After experiments, we found that the matching degree obtained by selecting the kernel function RBF is relatively accurate. However, in the process of understanding the kernel function RBF, it is found that the parameter gamma has a certain influence on the matching degree obtained by it, so the parameter is tested and studied.

After testing, it is found that under the training of the kernel function RBF, the effect of tomato disease recognition is the best, and it is found that when the gamma parameter is selected in the interval 140 to160, the accuracy of image recognition will be improved. In future research, we can also strengthen the preprocessing of disease images, taking into account the feature extraction of different attributes during image processing from multiple aspects, thereby improving the accuracy and efficiency of image recognition.

References

1. Kulkarni, O.: Crop disease detection using deep learning. In: International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–4 (2018)

2. Zhu, S., Zhang, J., Shuai, G., Hongli, L., Zhang, F., Dong, Z.: Autumn crop mapping based on deep learning method driven by historical labelled dataset. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 4669–4672 (2020)
3. Kalimuthu, M., Vaishnavi, P., Kishore, M.: Crop prediction using machine learning. In: International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 926–932 (2020)
4. Medar, R., Rajpurohit, V.S., Shweta, S.: Crop yield prediction using machine learning techniques. In: IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1–5 (2019)
5. Xu, Q., Zhang, J., Zhang, F., Zhu, S.: Develop large-area autumn crop type product using a deep learning strategy. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 4673–4676 (2020)
6. Singh, J., Mahapatra, A., Basu, S., Banerjee, B.: Assessment of Sentinel-1 and Sentinel-2 satellite imagery for crop classification in Indian region during Kharif and Rabi crop cycles. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 3720–3723 (2019)
7. Kumar, A., Sarkar, S., Pradhan, C.: Recommendation system for crop identification and pest control technique in agriculture. In: International Conference on Communication and Signal Processing (ICCSP), pp. 0185–0189 (2019)
8. Verma, G., Taluja, C., Saxena, A.K.: Vision based detection and classification of disease on rice crops using convolutional neural network. In: International Conference on Cutting-Edge Technologies in Engineering (ICon-CuTE), pp. 1–4 (2019)
9. Kang, J., Zhang, H., Yang, H., Zhang, L.: Support vector machine classification of crop lands using Sentinel-2 imagery. In: International Conference on Agro-geoinformatics (Agro-geoinformatics), pp. 1–6 (2018)
10. Chu, H., Zhang, D., Shao, Y., Chang, Z., Guo, Y., Zhang, N.: Using HOG descriptors and UAV for crop pest monitoring. In: Chinese Automation Congress (CAC), pp. 1516–1519 (2018)
11. Hu, H., Su, C., Yu, P.: Research on pest and disease recognition algorithms based on convolutional neural network. In: International Conference on Virtual Reality and Intelligent Systems (ICVRIS), pp. 166–168 (2019)
12. Pujari, J.D., Yakkundimath, R., Byadgi, A.S.: Identification and classification of fungal disease affected on agriculture/horticulture crops using image processing techniques. In: IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–4 (2014)
13. Militante, S.V., Gerardo, B.D., Medina, R.P.: Sugarcane disease recognition using deep learning. In: IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), pp. 575–578 (2019)
14. Park, H., Eun, J., Kim, S.: Image-based disease diagnosing and predicting of the crops through the deep learning mechanism. In: International Conference on Information and Communication Technology Convergence (ICTC), pp. 129–131 (2017)
15. Emon, S.H., Mridha, M.A.H., Shovon, M.: Automated recognition of rice grain diseases using deep learning. In: International Conference on Electrical and Computer Engineering (ICECE), pp. 230–233 (2020)
16. Prashar, K., Talwar, R., Kant, C.: CNN based on overlapping pooling method and multi-layered learning with SVM & KNN for American cotton leaf disease recognition. In: International Conference on Automation, Computational and Technology Management (ICACTM), pp. 330–333 (2019)
17. Nikhitha, M., Roopa Sri, S., Uma Maheswari, B.: Fruit recognition and grade of disease detection using inception V3 model. In: Communication and Aerospace Technology (ICECA), pp. 1040–1043 (2019)
18. Ai, Y., Sun, C., Tie, J., Cai, X.: Research on recognition model of crop diseases and insect pests based on deep learning in harsh environments. IEEE Access **8**, 171686–171693 (2020)

19. Genaev, M., Ekaterina, S., Afonnikov, D.: Application of neural networks to image recognition of wheat rust diseases. In: Genomics and Bioinformatics (CSGB), pp. 40–42 (2020)
20. Liu, J., Lv, F., Di, P.: Identification of sunflower leaf diseases based on random forest algorithm. In: Automation and Systems (ICICAS), pp. 459–463 (2019)
21. Zhu, S., Xu, C., Wang, J., Xiao, Y., Ma, F.: Research and application of combined kernel SVM in dynamic voiceprint password authentication system. In: IEEE 9th International Conference on Communication Software and Networks (ICCSN), pp. 1052–1055 (2017)