Remote Sensing Based Monitoring of Winter Wheat Powdery Mildew at a Regional Scale Using Random Forest Model

Linsheng Huang, Wenjuan Ding, Jing Jiang, Zhaochuan Wu and Jinling Zhao

National Engineering Research Center for Analysis and Application of Agro-Ecological Big Data, Anhui University, Hefei 230601, China
aling0123@163.com

Abstract: Wheat powdery mildew is one of the main diseases of wheat. The emergence and spread have seriously affected the yield and quality of wheat. It is of great significance to accurately monitor the occurrence of such a plant disease at a regional scale. Satellite remote sensing imagery has been extensively utilized in monitoring and assessing various plant diseases, due to its own particular advantages of real-time performance, wide coverage, high tempo-spatial resolution, etc. It is therefore suitable as a data source for monitoring the occurrence of wheat powdery mildew. Gaocheng District, Zhao County and Jinzhou City of Hebei Province, China were used as the study area. A total of eleven features were extracted from Landsat 8 OLI (Operational Land Imager) remote sensing data. Due to the existence of inter-band correlation, the Relief-F algorithm was used to carry out the feature selection. According to the weight value of each feature, three features were determined as the input into the created monitoring model. Consequently, the monitoring model was respectively established using the random forest (RF) and support vector machine (SVM). The analysis results show that the overall accuracy (OA) of RF model reaches 84% with the Kappa coefficient of 0.67, which are better than the SVM model of 73% and 0.47.

1. Introduction

Wheat is one of the main crops in China, however, its yield and quality are always affected by various diseases and insect pests. Wheat diseases and insect pests are characterized by multiple species, significant effect and local occurrence [1]. As one of the main diseases, powdery mildew is particularly serious and can occur at all the growth periods of wheat. In recent years, the disease also has a more serious tendency in the northeastern, northern and northwestern regions of China. After being damaged by the disease, wheat leaves become dry and the yield can be reduced by more than 20% [2]. Therefore, it is of great significance to accurately and timely monitor wheat powdery mildew for preventing and controlling the spread of this pathogen.

During the period of infection, the cell structure, pigment, water content, nitrogen content and external shape of wheat leaves will accordingly change, which will cause the significant changes in spectrum, thus it provides a basis for remote sensing monitoring of diseases and insect pests [3, 4]. At present, due to the increasing number of airborne and spaceborne sensors, various types of remote sensing data have been provided for the monitoring and forecasting of crop diseases and insect pests. Previous researchers have proposed different remote sensing based reverse parameters related to wheat pests and diseases, such as vegetation index (VI), land surface temperature (LST), and humidity by remote sensing data. It is obvious that the emergency of remote sensing has greatly facilitated the monitoring of wheat pests and diseases [5, 6].
There is much information in remote sensing data, but redundant information also exits. More and more data mining algorithms have extensively used in the analysis and identification of diseases and insect pests [7-10]. Nevertheless, the accuracy of retrieval models is usually affected by the number of ground truth. When the number of samples is small, the sample labels may be inappropriate, e.g. over fitting, which will affect the usefulness of the model to a certain degree. Random forest algorithm is a new type of high-efficient combination classification method, which has been widely used in medicine, ecological environment and other fields [11, 12]. It is rarely used to differentiate the plant diseases. Landsat 8 OLI (Operational Land Imager) were used as the remote sensing data. A monitoring model or wheat powdery mildew was constructed based on random forest on a regional scale. Support Vector Machine (SVM) based model was used to validate the efficiency of our proposed method.

2. Materials and Methods

2.1. Study Area

The study area consists of Gaocheng District, Zhaoxian and Jinzhou City of Hebei Province, China (114°36´ E ~ 115°12´ E, 37°37´ N ~ 38°18´ N) (Figure 1). It is located in the warm temperate zone with a semi-humid continental monsoon climate. Wheat is the dominant crop in the study area due to a flat topography. According to agricultural statistical data of this region, powdery mildew frequently occurs and has caused great losses. It is relatively an ideal pilot for performing the remote sense-based disease monitoring experiments.

2.2. Data Collection and Preprocessing

Collection of Field Data. The ground truth data were obtained on May 23, 25, 27, and 28, 2014. A total of 75 available data were obtained, and the disease severity were classified into 5 levels: 0 (healthy), 1 (mild), 2 (moderate), 3 (severe) and 4 (extremely severe). Since the primary objective of this study is to identify the occurrence of disease, five levels were reclassified into two levels: health...
and disease. The mild level and healthy level were considered as the healthy, and the moderate, the severe, and the extremely severe level were used as the diseased.

**Remote Sensing Data.** According to the data collection time of ground survey, three Landsat 8 satellite images were acquired on May 6, May 22, and June 7, 2014. Preprocessing of the Landsat 8 OLI data were performed including radiometric calibration, atmospheric correction, image masking, etc. All the preprocessing procedures were carried out in the ENVI (The Environment for Visualizing Images) software.

### 2.3. Feature Selection

A total of eleven features were selected in the study, including seven vegetation indices (Table 1) related to wheat powdery mildew, LST, the wetness reflecting the surface moisture, the greenness reflecting the ground vegetation coverage and leaf area index, and the Brightness reflecting soil brightness [13].

**Table 1.** Description of the selected vegetation indices.

| Vegetation index (VI) | Formula* | Reference |
|-----------------------|----------|-----------|
| Normalized difference vegetation index (NDVI) | \(\frac{(B_{\text{mir}}-B_{\text{red}})}{(B_{\text{mir}}+B_{\text{red}})}\) | [14] |
| Green band normalized vegetation index (GNDVI) | \(\frac{(B_{\text{mir}}-B_{\text{green}})}{(B_{\text{mir}}+B_{\text{green}})}\) | [15] |
| Enhanced vegetation index (EVI) | \(\frac{(B_{\text{mir}}-B_{\text{red}})}{(B_{\text{mir}}+6*B_{\text{red}}-7.5*B_{\text{green}}+1)}\) | [16] |
| Soil adjusted vegetation index (SAVI) | \(\frac{(B_{\text{mir}}-B_{\text{red}})}{(B_{\text{mir}}+B_{\text{red}}+0.5)}\) | [17] |
| Modified simple radio index (MSR) | \(\frac{B_{\text{mir}}/B_{\text{red}}}{1/[(B_{\text{mir}}/B_{\text{red}})^{*1/2}+1]}\) | [18] |
| Radio vegetation index (RVI) | \(B_{\text{mir}}/B_{\text{red}}\) | [19] |
| Re-normalized difference vegetation index (RDVI) | \(\frac{(B_{\text{mir}}-B_{\text{red}})}{(B_{\text{mir}}/B_{\text{red}})^{*1/2}}\) | [20] |

*blue, green, red and nir respectively refer to the blue band, green band, red band and near-infrared band of Landsat 8 OLI.

### 2.4. Reduction of Feature Dimensions

Due to the correlation of different features, to improve the operational efficiency and prediction accuracy of the classification model, the Relief-F algorithm was introduced to reduce the feature dimensions. As a filter selection method [21], Relief-F algorithm can provide different weights according to the relevance of each feature and category.

### 2.5. Random Forest (RF)

RF is a combination of decision tree classifiers and is widely used in recognition researches. It combines Breimans' "Bootstrap aggregating" idea [22] and Ho’s "random subspace" method [23]. Its essence is a classifier containing multiple decision trees \(\{h_i(X, \lambda_i)\}, i=1, 2, ..., N\). These decision trees are formed using a random approach, so a random decision tree is also made, and there is no correlation between the trees. When the test data enters the RF, it can actually classify each decision tree to form a classification model sequence \(\{h_1(X, \lambda_1), h_2(X, \lambda_2), ..., h_i(X, \lambda_i)\}\). Finally, the category with the most classification results in all decision trees is the final prediction result (Eq. 1).

\[
f(x_t) = \text{majority vote}\left[ h_1(X, \lambda_1), h_2(X, \lambda_2), ..., h_i(X, \lambda_i) \right]^{N}_{1} = 1
\]

(1)

where \(X\) is the input vector, \(\lambda_i\) is the same-distributed and independent random vector, and \(f(x_i)\) is the predicted result of the sample \(x_t\) to be tested.
2.6. Development of Monitoring Model

There are two important parameters in the RF model construction process, namely the number of decision trees in the forest (ntree) and the number of randomly selected attributes of the internal nodes (mtry). The ntree should be greater than 100, mtry should be less than the number of features used for model construction [24]. In this experiment, ntree was 500 and mtry was 1. SVM model has two important parameters c and g. In this paper, we used the grid parameter optimization GS [25] to find the optimal c and g, optimize the SVM model, and improve its classification accuracy.

A total of 75 ground survey points was obtained in this experiment. Due to the small number of ground truth points, the experiment was divided into two groups. The first group was to randomly select 45 points as the training samples, and the remaining were used for the accuracy verification. The second group was to randomly select 50 as the training samples, and the remaining were used for the accuracy verification.

3. Results and Discussions

3.1. Result of Relief-F

It is obvious that the top three weights are Wetness, MSR and SAVI (Table 2). The larger the weight values calculated by the relief-F algorithm, the greater the correlation between the feature and the category. Consequently, the three features were chosen for the construction of RF and SVM models.

|       | VI   | NDVI | GNDVI | EVI  | SAVI | MSR  | LST  |
|-------|------|------|-------|------|------|------|------|
| Weight | 0.0645 | 0.0582 | 0.0405 | 0.0967 | 0.1135 | 0.0604 |

|       | VI   | RVI  | RDVI | Brightness | Greenness | Wetness |
|-------|------|------|------|------------|-----------|---------|
| Weight | 0.0771 | 0.0752 | 0.0675 | 0.0716 | 0.1546 |

3.2. Spatial Distribution Map of Wheat Powdery Mildew

**Result of the First Set of Experiments.** It could be seen from the monitoring results that the RF monitoring results differed greatly from the SVM. The results of RF monitoring showed that there was a large area of powdery mildew in Zhaoxian, and powdery mildew occurred in the northern part of Gaocheng. The results of SVM showed that the incidence of powdery mildew in the three districts was relatively mild. The monitoring results of the two models showed that the greatest difference for the occurrence of powdery mildew was located in Zhaoxian (Figure 2).

According to the statistical analysis of the two monitoring results by ArcGIS software, the area ratio between the powdery mildew and total wheat in Zhaoxian was the largest: 61% (Figure. 2a) vs. 39% (Figure. 2b). According to the ground truth, 21 of the 28 sample points collected in Zhaoxian were disease samples. Accordingly, it could be also seen that the occurrence of powdery mildew in Zhaoxian was more serious. As shown in Figure. 2b, the distribution of powdery mildew in the southern part of Gaocheng and Jinzhou were relatively scattered. It was characterized by rapid propagation, wide spread area, and generally did not occur sporadically. Therefore, it could be found that the RF based monitoring result was more in line with the occurrence of powdery mildew in general.
Healthy Powdery Mildew
(a) Monitoring results of RF (b) Monitoring results of SVM

**Figure 2.** Spatial distribution map of wheat powdery mildew (First group).

**Result of the Second Set of Experiments.** In the second group of experimental monitoring results, the spatial distribution maps of the occurrence of powdery mildew monitored by the two models were extremely similar. Specifically, the occurrence of powdery mildew in the northern areas of both Zhaoxian and Gaocheng were more serious in the two monitoring results (Figure. 3).

(a) Monitoring results of RF (b) Monitoring results of SVM

**Figure 3.** Spatial distribution map of wheat powdery mildew (Second group).
3.3. Comparison of the Two Experiments
Comparing the two groups of experiments, the spatial distribution map of the occurrence of powdery mildew in RF monitoring was similar, and the results of SVM monitoring were very different, indicating that the number of samples selected had a significant influence on the SVM model.

3.4. Validation of the Two Models
To verify the accuracy of the model prediction, the ground truth data of powdery mildew on May 27 and 28, 2014 were used to evaluate the models. The overall accuracy (OA) of the two models was above 70%. In the first group of experiments, the OA of RF and SVM models were respectively 84% and 73%. In the second group of experiments, the overall accuracy of RF and SVM models were 84% and 92%, respectively. The SVM model constructed with 50 training samples had the highest accuracy, while the SVM model created with 45 training samples had the lowest accuracy with the Kappa coefficient of less than 0.5. The accuracy of RF in both groups of experiments reached 80%, and the Kappa coefficient exceeded 0.5, indicating that RF can be applied to wheat powdery mildew monitoring and insensitive to training samples. The SVM models in the two groups of experiments differ greatly, indicating that the SVM model relies on training samples (Table 3).

Table 3. Overall verification results.

| Models | Actual sample | Prediction sample Healthy PM | Sum | OA | Mean | Kappa |
|--------|---------------|------------------------------|-----|----|------|-------|
| | | Healthy PM | Sum | | | |
| | | | 11 | 4 | 4 | 11 | 15 | 15 | 30 | 73% | 73% |
| | | | 15 | 15 | 30 | 73% | 0.47 |
| SVM | Sum | | 13 | 2 | 3 | 12 | 16 | 14 | 15 | 15 | 30 | 87% | 80% |
| | | | 84% | 0.67 |
| | | | 7 | 1 | 1 | 16 | 8 | 17 | 8 | 17 | 25 | 85% | 94% |
| RF | | | 8 | 17 | 25 | 92% | 0.81 |
| | | | 7 | 1 | 3 | 14 | 10 | 15 | 8 | 17 | 25 | 87% | 82% |
| | | | 84% | 0.65 |

4. Conclusions
According to the above analysis, in both sets of experiments, RF achieved higher accuracy, but the two-group experimental results of SVM-based model differed greatly. According to the experiments, the reasons can be considered as follows: (1) the data used for model training is too small, (2) there are some sample label errors in the ground truth data, (3) the imbalance between the number of diseased and healthy samples, and (4) the selection of training samples. The over-fitting of SVM model leads to
the misclassification phenomenon. Conversely, RF-based model can handle the over-fitting problem well, and the balance of classification errors can still be maintained for the unbalanced data. Therefore, the prediction accuracy is improved as a whole. In the actual situation, the samples will inevitably lead to some problems such as misclassification, imbalance. When collecting samples, the number of samples tends to be less, and RF is more suitable than SVM to monitor the occurrence of wheat diseases on a regional scale.

5. Acknowledgements
This work was supported by the Anhui Provincial Major Scientific and Technological Special Project (16030701091).

6. References
[1] Hsam, S.L.K., Huang, X.Q., Ernst, F., Hartl, L., Zeller, F.J.: Chromosomal location of genes for resistance to powdery mildew in common wheat (tritici aestivum l. em thall.). 5. alleles at the pm1 locus. Theor. Appl. Genet. 96, 1129–1134 (1998)
[2] Ma, H., Huang, W., Jing, Y.: Wheat powdery mildew forecasting in filling stage based on remote sensing and meteorological data. Trans. CSAE 32, 165–172 (2016)
[3] Hatfield, J.L., Fjir, P.: Remote sensing for crop protection. Crop Prot. 12, 403–413 (1993)
[4] Newton, A.C., Hackett, C.A., Lowe, R., Wale, S.J.: Relationship between canopy reflectance and yield loss due to disease in barley. Ann. Appl. Biol. 145, 95–106 (2015)
[5] Graeff, S., Link, J., Claupein, W.: Identification of powdery mildew (Erysiphe graminis, sp. tritici) and take-all disease (Gaeumannomyces graminis, sp. tritici) in wheat (Triticum aestivum, L.) by means of leaf reflectance measurements. Cent. Eur. J. Biol. 1, 275–288 (2006)
[6] Zhao, C.J., Huang, M.Y., Huang, W.J., Liu, L.Y., Wang, J.H.: Analysis of winter wheat stripe rust characteristic spectrum and establishing of inversion models. In: 24th IEEE International Geoscience & Remote Sensing Symposium, pp. 4318–4320. IEEE Press, New York (2004)
[7] Rumpf, T., Mahlein, A.K., Steiner, U., Oerke, E.C., Plümer, L.: Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. Comput. Electron. Agr. 74, 91–99 (2010)
[8] Karimi, Y., Prasher, S.O., Patel, R.M., Kim, S.H.: Application of support vector machine technology for weed and nitrogen stress detection in corn. Comput. Electron. Agr. 51, 99–109 (2006)
[9] Moshou, D., Bravoa, C., Westb, J., Wahlen, S., McCartney, A., Ramon, H.: Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. Comput. Electron. Agr. 44, 173–188 (2004)
[10] Zhang, H., Hu, H., Zhang, X.B., Zhu, L.F., Zheng, K.F., Jin, Q.Y., Zeng, F.P.: Estimation of rice neck blast severity using spectral reflectance based on bp-neural network. Acta Physiol. Plant. 33, 2461–2466 (2011)
[11] Izmirlian, G.: Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. Ann. Ny. Acad. Sci. 1020, 154–174 (2010)
[12] Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. 67, 93–104 (2012)
[13] Rouse, B.T., Wells, R.J, Warner, N.L.: Proportion of T and B lymphocytes in lesions of Marek's disease: theoretical implications for pathogenesis. J. Immunol. 110, 534–539 (1973)
[14] Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N.: Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58, 289–298 (1996)
[15] Verstraete, M.M., Pinty, B., Myneni, R.B.: Potential and limitation of information extraction on the biosphere from satellite remote sensing. Remote Sens. Environ. 58, 201–214 (1996)
[16] Huete, A.R.: A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25, 295–309 (1988)
[17] Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Comput. Speech Lang. 13, 359–394 (1999)
[18] Jordan, C.F.: Derivation of leaf-area index from quality of light on the forest floor. Ecology 50, 663–666 (1969)

[19] Roujean, J.L., Breon, F.M.: Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. Remote Sens Environ, 51, 375–384 (1995)

[20] Liu, Q., Guo, Y., Liu, G., Zhao, J.: Classification of Landsat 8 OLI image using support vector machine with Tasseled Cap Transformation. In: 10th International Conference on Natural Computation, pp. 665–669. IEEE Press, New York (2014)

[21] Reyes, O., Morell, C., Ventura, S.: Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. Neurocomputing 161, 168–182 (2015)

[22] Breiman L.: Bagging predictors. Mach. Learn. 24, 123–140 (1996)

[23] Ho T.: The random subspace method for constructing decision forests. IEEE T. Pattern Anal. 20, 832–844 (1998)

[24] Huo, J., Shi, T., Chang, J.: Comparison of Random Forest and SVM for electrical short-term load forecast with different data sources. In: 7th IEEE International Conference on Software Engineering and Service Science, pp. 1077–1080. IEEE Press, New York (2016)

[25] Pal, M., Goel, A.: Prediction of the end-depth ratio and discharge in semi-circular and circular shaped channels using support vector machines. Flow. Meas. Instrum. 17, 49–57 (2006)