Comparison of Methods for Forecasting Yellow Rust in Winter Wheat at Regional Scale

Chenwei Nie, Lin Yuan, Xiaodong Yang, Liguang Wei,
Guijun Yang, and Jingcheng Zhang*

Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China
{nie_chenwei,byl6690,weiliguang}@126.com,
{yangxd,yanggj}@nercita.org.cn, zhangjc_rs@163.com

Abstract. Yellow rust (YR) is one of the most destructive diseases of wheat. To prevent the prevalence of the disease more effectively, it is important to forecast it at an early stage. To date, most disease forecasting models were developed based on meteorological data at a specific site with a long-term record. Such models allow only local disease prediction, yet have a problem to be extended to a broader region. However, given the YR usually occurs in a vast area, it is necessary to develop a large-scale disease forecasting model for prevention. To answer this call, in this study, based on several disease sensitive meteorological factors, we attempted to use Bayesian network (BNT), BP neural network (BP), support vector machine (SVM), and fisher liner discriminant analysis (FLDA) to develop YR forecasting models. Within Gansu Province, an important disease epidemic region in China, a time series field survey data that collected on multiple years (2010-2012) were used to conduct effective calibration and validation for the model. The results showed that most methods are able to produce reasonable estimations except FLDA. In addition, the temporal dispersal process of YR can be successfully delineated by BNT, BP and SVM. The three methods of BNT, BP and SVM are of great potential in development of disease forecasting model at a regional scale. In future, to further improve the model performance in disease forecasting, it is important to include additional biological and geographical information that are important for disease spread in the model development.

Keywords: Yellow rust, Disease forecast, Bayesian network (BNT), BP neural network (BP), support vector machine (SVM), fisher liner discriminant analysis (FLDA).

1 Introduction

Yellow rust (YR), caused by Puccnia striiformisWestend f.sp.tritici Eriks, is one of the most important epidemic diseases of wheat. It can cause significant loss on wheat at a global scale [1, 2]. It is of great importance to predict the YR effectively at an early stage, since it can provide critical information to agriculture plant protection

* Corresponding author.

D. Li and Y. Chen (Eds.): CCTA 2014, IFIP AICT 452, pp. 444–451, 2015.
DOI: 10.1007/978-3-319-19620-6_50

departments to facilitate timely spray recommendation. So far, a series of studies have been conducted to forecast YR over a long time based on meteorological and agronomy data around the world. Hu et al (2000) constructed a BP model to predict YR in Hanzhong city, Shaanxi Province. The forecast results were highly consistent with actual situation of disease occurrence [3]. Chen et al (2006) predicted YR severities at a seasonal time step in both Maerkang county and Tianshui city using a discriminant analysis, with rewind accuracy and cross-validation accuracy greater than 78% [4]. Coakley et al (2006) developed an improved method to predict YR severity [5]. Wang et al (2012) conducted a study to develop a stable neutral network for predicting YR prevalence degree [6].

To date, it should be noted that there were few attempts made in regular YR forecasting (time step = 7 days) at a regional scale. Instead, efforts were made on forecasting seasonal severities of YR which rely on spores counting data and meteorological observations. Those models can achieve high accuracy at a local site, whereas it is difficult to apply those models in vast areas where the spores counting data are not available. In addition, given the distribution of the YR pathogen over large area is driven by oversummer and overwinter process at a regional scale which is closely associated with weather conditions, it is thereby necessary to develop a model that can be applied at a large spatial scale. However, such forecasting models are lacking recently.

Several critical weather factors associated with the occurrence of YR on winter wheat had been reported, they are air temperature, humidity, precipitation and sunshine duration [7]. It is important to relate YR occurrence with meteorological factors in the development of YR forecasting model. Several machine-learning techniques have been widely used for classification purpose since the intelligent learning feature allowing them to take advantages from large amount of data [8]. In this study, the Gansu province, which is an important YR prevalence region in China, was selected as our study area. Based on continuous YR field survey data over multiple years (2010-2012) and corresponding meteorological data, the potential of BNT, BP, SVM and FLDA in disease forecasting were examined and compared. The YR forecasting model was thus established to facilitate regular disease management at a regional scale.

2 Materials and Methods

2.1 Yellow Rust Survey Data

The YR survey data is collected by Gansu Provincial Protection Station. During 2010 to 2012, a weekly field surveys were conducted across southern area of Gansu province (Fig.1). In detail, the surveyed data include the initial date of disease occurrence, the prevalence status and the infected area. The climate of the study region is characterized by high humidity and rainfall, and YR disease occurs almost every year. A total of 22, 9, 30 counties were surveyed in 2010, 2011 and 2012, respectively. The distribution of surveyed counties is demonstrated in Fig.1. The investigation ranged from the beginning of March to the end of July in each year. There were 16 weeks field survey data totally in each year. For model calibration and validation, the surveyed data were randomly split into 60% versus 40% in each year.

In this study, prevalence status of YR disease in the week before forecasting day was chosen as one of the input variables and the occurrence status of YR disease

during the week next the forecasting day was regard as the forecasting target. The occurrence status of YR disease was divided into four classes represented by D1, D2, D3 and D4, respectively. D1 meant there was no YR disease was found in the county, D2 indicate the YR disease was firstly found since the survey was conducted in the county, D3 indicate that the YR disease has been found in the county, but there was no development compared with last week, and D4 indicate the YR disease has been found before and it developed during the week after forecast day.

2.2 Meteorological Data

In this study, according to the research results of Cooke(2006) and Li & Zeng(2002)[2,9], four types of meteorological factors were chosen as the primary data, include air temperature(maximum air temperature, minimum air temperature and average air temperature), average humidity, precipitation and sunshine duration. The daily data of these meteorological factors from a total of 27 weather stations around the study area was acquired from Chinese Meteorological Data Sharing Service System. The time range of the data is from a week before YR occurrence (based on the investigation data) in spring to its mature stage in each year. There are 4 steps to process meteorological data, including removal of abnormal value, computing meteorological factors, choosing meteorological factors and interpolation of each factor to a resolution of 250m*250m. Considering some meteorological data have a strong relationship with altitude, the DEM (Digital Elevation Model) data was used the adjust the spatial maps of meteorological factors by interpolating the fitted residue across the region [10,11].

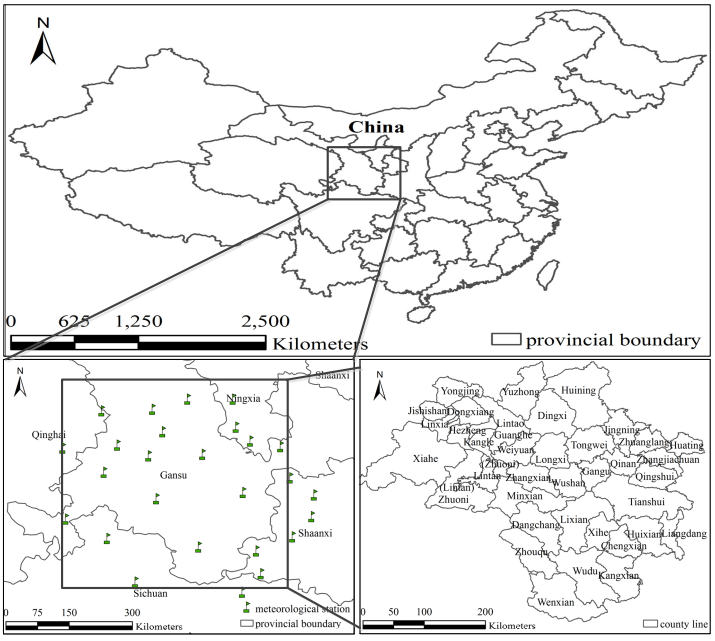

Fig. 1. Study area and the distribution of meteorological stations

Table 1. Meteorological factors used for predicting

Meteorological factors	Time range
Average minimum air temperature (aveminT)	From 7th day to 14th day before forecasting day
Average maximum air temperature(avemaxT)	From 7th day to 14th day before forecasting day
Days of Precipitation more than 0.25mm(Pdays)	From 7th day to 14th day before forecasting day
Average precipitation(aveP)	From march to the 7th day before forecasting day
Days of humidity lower than 50%(Hdays)	From 7th day to 14th day before forecasting day
Minimum of average humidity(minH)	From 7th day to 14th day before forecasting day
Average sunshine duration(aveS)	From march to the 7th day before forecasting day

As for select methods, the variance analysis between meteorological factors and disease prevalence classes and the correlation analysis among the same class meteorological factors were conducted by Tukey-Kramer method and Pearson method, respectively. For those meteorological factors have a *p-value*<0.05 were selected primarily, then if the R > 0.8 between two factors, the factor has a smaller *p-value* was chosen. And the meteorological factors were last chosen as shown in table 1. As for interpolation methods, the normality of the distribution of each meteorological factor was examined by Kolmogorov-Smirnov method. For those meteorological factors have a *p-value*<0.05, a kriging method is used to conduct interpolation. Otherwise, an inverse distance weighted method is adopted. Each meteorological factor was calculated on a county scale after the interpolation.

2.3 Methods

To find an appropriate method for predicting YR disease, in this study, four YR disease forecasting models were established with four classical methods, respectively, include BNT, BP, SVM and FLDA. The characters of all 4 methods are shown in table 2[12]:

Table 2. Characters of each method

Methods	Description
Bayesian network(BNT)	Based on traditional statistical theory. Has been used effectively to model those problems with characters uncertainty and non-linearity by incorporating prior knowledge extracted from selected sample datasets.
BP neural network(BP)	Based on traditional statistical theory. Has strongly adaptive and learning capability. Has been effectively to model those objects with characters uncertainty and non-linearity.
Support vector machine(SVM)	Based on statistical learning theory, has been widely introduced in disease recognition. Has been used effectively to model those problems with characters small sample, uncertainty and non-linearity.
Fisher linear discriminant analysis(FLDA)	An important branch of Statistical pattern recognition. Has been widely used in pattern recognition field, since it is a simple, rapid and efficient method.

BNT is a Directed Acyclic Graph(DAG), the nodes in the DAG structure representing domain variables, and the arcs between nodes represent probabilistic dependencies. In this research, the nodes were used to represent meteorological factors, prevalence status in last week before the forecasting day and the occurrence

status during the week next the forecasting day. Each meteorological factor was graded according to the clinical characters of YR disease before the BNT model was established. Next, a BNT structure was constructed with the MCMC method and corresponding expertise. Then, the parameters were optimized using maximum likelihood estimation against the calibration data set. According to the chain rule of probability, the probability of an event occurs is calculated as:

$$p(x) = \prod_{i=1}^{9} p(x_i | x_1, \cdots, x_{i-1}) \tag{1}$$

One-hidden-layers BP forecasting model was constructed in this study with training function *trainlm* and adaption learning function *learngdm* under MATLAB environment. *Tansig* was used as the transfer function of both hidden layer and output layer. The structure of BP that is used in this study is shown in Fig.2. As a binary output, the y_i will be marked as '1' if it gets the maximum, which means the occurrence status of YR disease is D*i*. Otherwise, the y_i will be marked as '0', which means the occurrence status of YR disease is not D*i*.

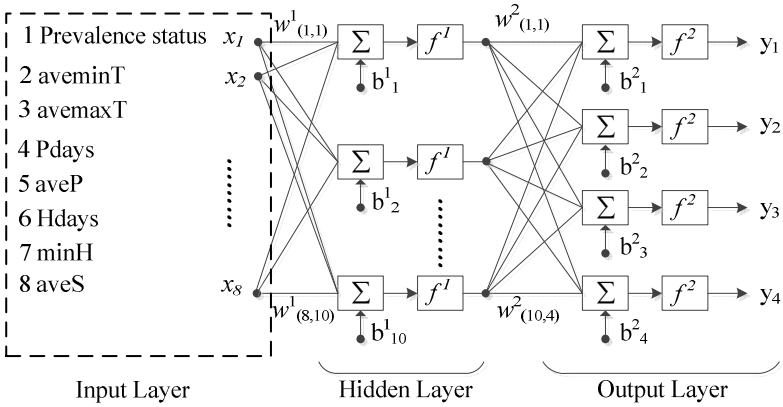

Fig. 2. Feed forward back propagation neural network

Note: **y**=Output vector, **X**=Input vector, f^1,f^2=Transfer functions on hidden layer and Output layer, b^1, b^2=bias on hidden layer and Output layer, w^1, w^2=weights in Input layer and hidden layer.

The output of the BP model is calculated as:

$$\mathbf{y} = f^2(w^2 f^1(w^1 \mathbf{X} + \mathrm{b}^1) + \mathrm{b}^2) \tag{2}$$

In the third part of the experiment design, a SVM model was constructed under MATLAB environment with help of Libsvm tool and SVM_GUI tool. The RBF function was used as a kernel function. Grid search method was used as the parameters optimizing algorithm to search for the best penalty coefficient *c* and *gamma* with a step of 0.5. Both *c* and *gamma* have a range of [-8,8]. The structure of SVM model in this study is shown in Fig.3:

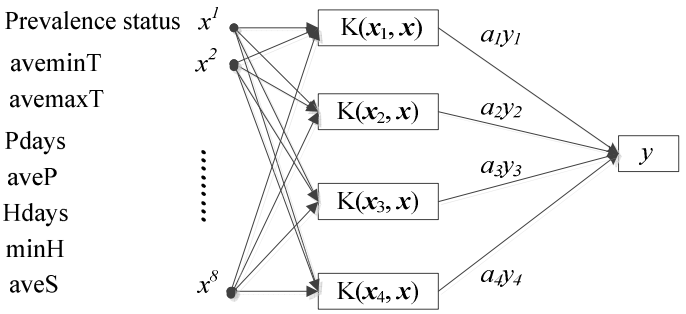

Fig. 3. Support vector machine design

Note: y is the output, $K(\boldsymbol{x}_i, \boldsymbol{x})$ is nonlinear transformation(inner product computation) function, $\boldsymbol{x}_i$ is support vector, $\boldsymbol{x} = (x^1, x^2, \ldots, x^8)$ **is** Input vector, a_i is Lagrangian coefficient.

The output of SVM model is calculated as:

$$y = \text{sgn}(\sum_{i=1}^{4} a_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b) \tag{3}$$

For FLDA forecasting model, likelihood ratio was used to assign observations to groups. It was run under MATLAB environment.

2.4 Evaluation of Disease Forecast Models

For the all 4 trained examples of the classifiers, a separate validation data set was used to evaluate the model accuracy. Different from the BP, SVM and FLDA, which classify a sample into several infection status directly, the BNT forecasting model produced result in probability, which would be converted to infection status by applying a certain threshold. A sample will be marked as D_i when the occurrence status D_i gets the maximal probability. The performances of four different forecasting models were evaluated by overall accuracy as comparing the forecasting results against validation data.

3 Results and Discussion

Table 3 summarized the forecasting results of BNT, BP, SVM and FLDA versus the validation data set. The results suggested that BNT, BP and SVM produced more accurate forecasts than FLDA in general. The SVM, BNT and BP produced approximately similar accuracies with SVM model having relatively high accuracy. The actual occurrence (Fig.4 (a)) and forecasting results (Fig.4 (b-e)) of disease distribution pattern of the 4 methods were demonstrated in Fig.4. As shown in Fig.4(a), the actual occurrence, there were only two counties infected by the disease in 9th week, however, the all 9 validation counties were infected in 13th week. It is obvious that the infection patterns estimated by BNT, BP and SVM are highly consistent with field survey records across multiple dates, with BNT outperformed BP and SVM.

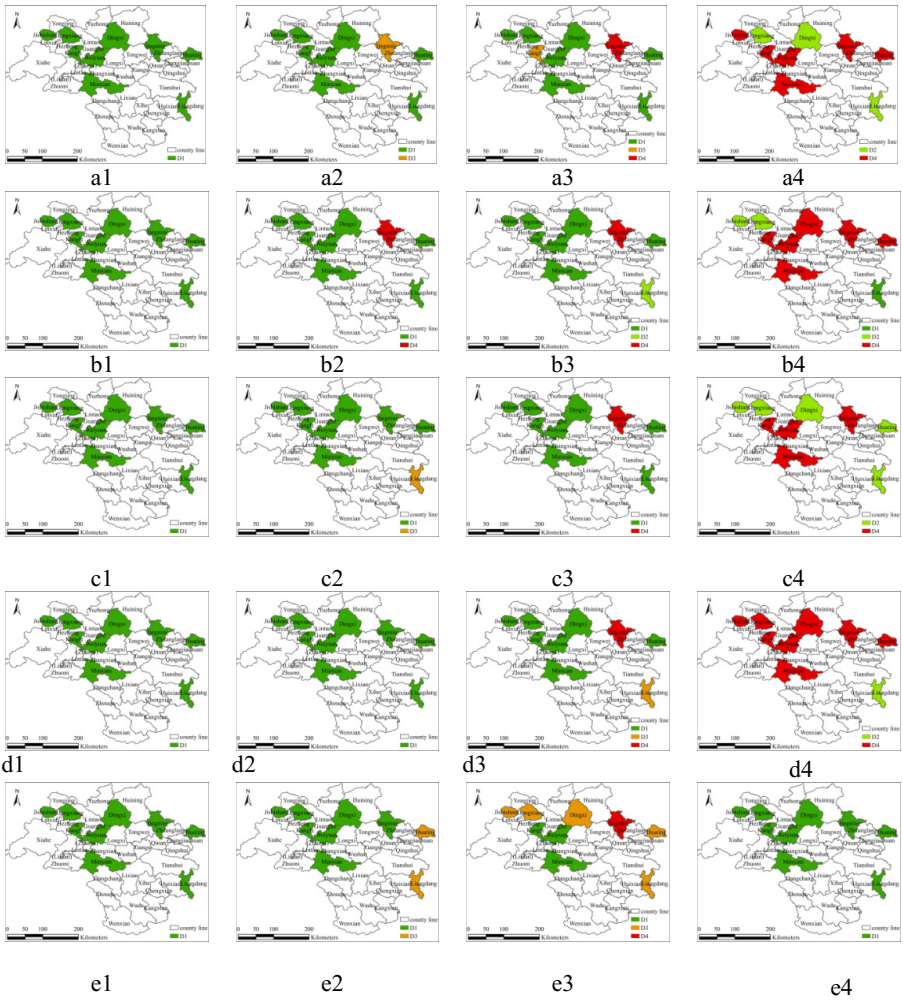

Fig. 4. The forecasting result of four methods and the true occurrence status

Note: a indicates the true occurrence status, b indicates the result of BNT, c indicates the result of BP network, d indicates the result of SVM, e indicates the result of FLDA. The number 1 indicates the first week, 2 indicates the fifth week, 3 indicates the ninth week and 4 indicates the thirteenth week.

Table 3. accuracy indices of tested methods

Methods	OAA	Kappa
BNT	82.29	0.73
BP	81.25	0.71
SVM	85.68	0.78
FLDA	68.75	0.56

To predict the YR disease more precisely, more information should be chosen as inputs of the forecasting model, since the YR disease is influenced by various factors, including meteorological conditions, growth vigor of wheat and number of fungus. As the remote sensing method can estimate the crop growth accurately and quickly in a large space scale, it would be used to forecast the YR disease integrating with meteorological information and fungus information in our future work.

4 Conclusions

A total of four methods including BNT, BP, SVM and FLDA were examined and compared in developing a forecasting model of yellow rust disease across vast area in this study. The performances of these models were evaluated against a weekly survey data during wheat's key growing stages from 2010 to 2012. The results confirmed that the disease forecasted results are able to reflect the spatio-temporal development and distribution pattern of YR except for FLDA. Further, a superior performance of BNT, BP and SVM also demonstrated that these nonlinearity methods are of great potential in forecasting yellow rust infection at a regional scale in weekly time step.

Acknowledgment. This work was supported in part by the Natural Science Foundations of China (Grant No. 41101395), Natural Science Foundations of Beijing (Grant No. 4122032) and Prior Sci-Tech Program for Scientific Activity of Overseas staff.

References

1. Shimai, Z.: Macro-phytopathology. Agriculture Press of China, Beijing (2005) (in Chinese)
2. Cooke, B.M., David, G.J., Bernard, K., et al.: The epidemiology of plant diseases. Springer, Dordrecht (2006)
3. Lei, Y., Shuqin, L.: Prediction of wheat stripe rust by wavelet neural network. Microcomputer Information 25(12-2), 42–43 (2009)
4. Chen, G., Wang, H., Ma, Z.: Forecasting wheat stripe rust by discrimination analysis. Plant Protection 32(4), 24–27 (2006)
5. Coakley, S.M., Line, R.F., McDaniel, L.R.: Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data. Phytopathology 78(5), 543–550 (1988)
6. Wang, H., Ma, Z.: Prediction of wheat stripe rust based on neural networks. In: Li, D., Chen, Y. (eds.) CCTA 2011, Part II. IFIP AICT, vol. 369, pp. 504–515. Springer, Heidelberg (2012)
7. Te Beest, D.E., Paveley, N.D., Shaw, M.W., et al.: Disease-weather relationships for powdery mildew and yellow rust on winter wheat. Phytopathology 98(5), 609–617 (2008)
8. Khan, M.N.A.: Performance analysis of Bayesian networks and neural networks in classification of file system activities. Computers & Security 31(4), 391–401 (2012)
9. Li, Z., Zeng, S.: Wheat rust in china. Agriculture Press of China, Beijing (2002)
10. Pan, Y.Z., Gong, D.Y., Deng, L., Li, J., et al.: Smart distance searching-based and DEM-informed interpolation of surface air temperature in China. Acta Geographica Sinaca 59(3), 366–374 (2004)
11. Wang, Z., Shi, Q.D., Chang, S.L., et al.: Study on spatial interpolation method of mean air temperature in Xinjiang. Plateau Meteorology 31(1), 201–208 (2012)
12. Bian, Z.Q.: Pattern Recognition. Tsinghua University Press, Beijing (2007)