Filomat 32:5 (2018), 1971–1976
DOI (will be added later)

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Research on the Model for Tobacco Disease Prevention and Control Based on Case-based Reasoning and Knowledge Graph

Lichuan Gu[a], Yingchun Xia[a], Xiaohui Yuan[b], Chao Wang[a], Jun Jiao[a]

[a]*School of Information & Computer, Anhui Agricultural University,China*
[b]*Department of Computer Science and Engineering, University of North Texas,Texas,76203 US*

Abstract. Tobacco is one of the most important economic crops in China. The yield and quality of tobacco reduce severely because of long-time disease invasion. Currently, the main focus of researches on tobacco disease prevention and control is the diagnosis of disease that has occurred, which ignores to predict disease before it outbreaks. Therefore, in this paper, we follow the idea that prediction is used before disease prevention and control and study the model for tobacco disease prevention and control by using knowledge graph and case-based reasoning (CBR). In order to implement the model, we choose tobacco mosaic virus (TMV) as research object and follow the following methods to prevent occurrence of that. At first, a method to predicting environmental factors by using principal component analysis (PCA) and support vector machine (SVM) is proposed. According to the prediction result, knowledge graph and CBR are used to retrieve the most similarity case and finally determine the best solution. Experimental results demonstrate that our model can achieve high accuracy and give the most appropriate scheme for disease prevention and control.

1. Introduction

Tobacco is one of the most important economic crops in China, but the yield and quality of tobacco reduce seriously because of disease invasion every year. Therefore, how to prevent and control tobacco disease is a serious problem that needs to be solved urgently in agricultural production.

At present, researches on tobacco disease prevention and control are mainly concentrated on technologies of prevention and control and technologies of disease diagnosis. For example, Hunan Agricultural University identified virus-like diseases and had a comprehensive treatment about using pesticides. It improved comprehensive treatment level of the major tobacco disease [1]. Niu Xiaoyi proposed the scheme for preventing tobacco disease using fungi polysaccharides after analyzing the harm of using pesticide [2]. Besides, Zhang Yanling built genetic neural network model for tobacco disease with multiple feature fusion and developed automatic identification and diagnosis system for tobacco disease [3]. However,

2010 *Mathematics Subject Classification*. Primary 68T01; Secondary 62C25, 62H30

Keywords. Knowledge graph, case-based reasoning, tobacco disease, prediction, prevention and control

Received: 12 October 2017; Accepted: 17 November 2017

Communicated by Hari M. Srivastava

Research supported by the National Natural Science Foundation of China (Grant No.31771679,No.31371533), the Special Fund for Key Program of Science and Technology of Anhui Province, China (Grant No.15czz03131, 16030701092), the Natural Science Foundation of the Anhui Higher Education Institutions of China(Grant No.kJ2016A836)

Email addresses: glc@ahau.edu.cn (Lichuan Gu), yingchun.xia@ahau.edu.cn (Yingchun Xia), xiaohui.yuan@unt.edu (Xiaohui Yuan), 31727122@qq.com (Chao Wang), jiaojun2000@sina.com (Jun Jiao)

these researches only emphasize on control and diagnosis, but ignore to predict the occurrence of disease. Thus, this paper proposes a model for tobacco disease prevention and control based on CBR and knowledge graph.

2. Related Concepts

2.1. CBR

CBR is a common knowledge-based solution. It solves problem by retrieving solution in previous cases which is most similar with the target case [4]. Its handling process includes case representation, case retrieval, case reuse, case revise and case-retaining, where the first three are research cores [5]. The key point for accelerating retrieval is plausible representation of relevant knowledge in case. For this purpose, we propose framework-based case representation as follows:

$Condition(A, T, F)$
$Result(S, P)$

Here, *Condition* is the condition framework. *Result* is result case framework after matching. A is the location. T is the time. F is the property assemblage. S is the matching result. P is the final scheme.

2.2. Knowledge Graph

Knowledge graph [6-8], which describes location, person, city and other entities, is a relationship network. It is a graph-based data structure. Points in graph are entities in actual world and edges indicates relationship. Unlike traditional keyword search, knowledge graphs grasp user intention at semantic level and enable complex information to be better queried. Our model exploits knowledge graph to assist knowledge expression in case representation, enrich condition content and enlarge the width and depth of keywords retrieval.

3. The Algorithm for Tobacco Disease Prevention and Control

Many Influencing factors are likely to cause tobacco disease infection during whole growth cycle. Each Influencing factor is indispensable. After taking full account of environmental factors for tobacco growth, we propose the algorithm for tobacco disease prevention and control by using PCA and SVM to achieve tobacco disease prediction and prevention. The following are the details of our algorithm.

3.1. Extracting Influencing Factors

Actually, environmental data always have large overlapping and high complexity. PCA is a common method for data analysis. It reduces dimensions of data with little information loss by compressing multiple features to less comprehensive indicators [9]. Thus, we use PCA to simplify indicator data set.

Assume p is the number of observation indicators about each sample and X is raw data matrix.

Step 1: Importing matrix (*dataset*) and obtaining the standard deviation (*stdr*) of each variables.

Step 2: Standardizing row data matrix.

Step 3: Establishing the correlation coefficient matrix R for matrix X.

Step 4: Solving characteristic root and corresponding unit eigenvector.

Step 5: Calculating contribution rate and cumulative contribution rate. When cumulative contribution rate of k principal components reaches 80%, p original variables can be replaced by them.

3.2. Predicting Influencing Factors

The key points for building a model is mastering environmental factors for tobacco growth and prediction of environmental factors. In this paper, we exploit the idea of classification to predict values of factors according to incidence of tobacco disease. SVM is a common classification algorithm, which builds the optimal hyper-plane in space based on structural risk minimization principle [10]. We use SVM to accomplish the classification and prediction of environmental factors. Here, the predictive objects are the principle components extracted in section 2.1. The following is steps of prediction:

Step 1: Selecting training set and test set. It should choose some data as training set from each label .

Step 2: Normalizing data into [0,1].

Step 3: Selecting kernel function RBF and determining optimum parameters C and g.

Step 4: Building SVM model and achieving prediction and classification by training train set and using test set to check out accuracy of the model.

3.3. Building Prevention and Control Model

1) Constructing Knowledge Graph

Relevant agricultural data come from agricultural disease database and open network. In order to build case library, the framework of knowledge graph is constructed by experts. On the basis of that, We build knowledge graph of TMV by transforming row data into RDF triple to represent heterogeneous knowledge. Figure 1 shows the example for constructing knowledge graph of TMV.

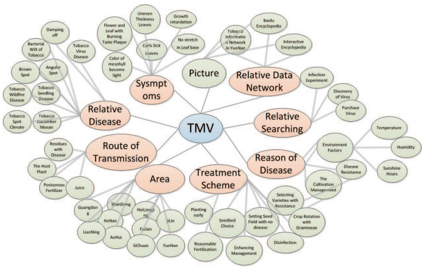

Figure 1: Knowledge graph of TMV

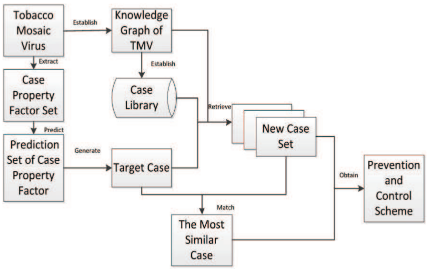

Figure 2: Framework chart of the prevention and control model

2) Building Model

Case representation: According to the framework-based case representation in section 1.1, case can be represented as:

$$Case\{Name, R :< fi, wi >, S, CaseID\}$$

Here, *Name* is the case name, *R* is the property of case, *fi* is the eigenvalue of property factors, *wi* is the weight of each property, *S* is the scheme in the case and *CaseID* is identification of case.

Setting index: The efficiency of case retrieval is improved by creating primary index and secondary index in descending order of the weight coefficients of properties. Disease occurrence time of tobacco are different because of different growth areas and planting time of that, so we set location as primary index and time as secondary index.

Building model: Figure 2 shows framework for constructing model and the following is the specific procedures.

Step 1: Extracting property factors of target case according to steps in section 3.1.

Step 2: Predicting property factors of target case in the light of steps in section 3.2.

Step 3: Building case library by using knowledge graph and normalizing it.

Step 4: Using primary index and secondary index successively to search the similar case.

Step 5: Determining weight coefficient d_i of property factors according to formula (1).

$$d_i = \frac{\sum_{k=1}^{n} a_{ki} * e_k}{\sum_{k=1}^{n} e_k} \tag{1}$$

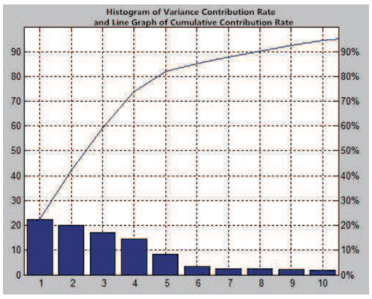

Figure 3: Variance contribution rate and cumulative contribution rate

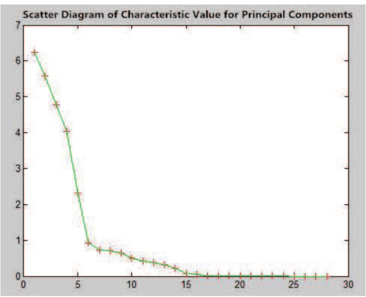

Figure 4: eigenvalue of principal component

Table 1: Major environmental indicators

Index Variables	Environment factors	Index Variables	Environment factors
X_1	Average Temperature	X_{10}	average sunshine time
X_2	Minimum Temperature	X_{22}	Maximum soil moisture in 10cm under ground
X_4	Average relative humidity	X_{23}	Average soil moisture in 20cm under ground
X_7	Average Rainfall	X_{28}	Maximum soil temperature
X_9	Maximum Rainfall		

Here, e_k indicates variance contribution rate of the k^th principal components, and a_ki stands for the i^th weight coefficient of the k^th principal components in comprehensive model.

Step 6: Measuring case similarity. Assume m is the number of the similar cases, n is the number of property factors and $w = w_1, w_2, \cdots w_n$ is weight feature of the property factors. The Sum of weight is:

$$Sub * w_\tau = \begin{pmatrix} |sub_{11}| & |sub_{12}| & \cdots & |sub_{1n}| \\ |sub_{21}| & |sub_{22}| & \cdots & |sub_{2n}| \\ \vdots & \vdots & \ddots & \vdots \\ |sub_{m1}| & |sub_{m2}| & \cdots & |sub_{mn}| \end{pmatrix} \times \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_m \end{pmatrix} \tag{2}$$

Here, Sub is the absolute value of difference between the target case and the similar cases. Δ is the distance matrix of target case with m cases. The smaller the distance, the more similar the cases are.

Step 7: Determining prevention and control scheme. We determine the status of disease via comparing the target case with the most similar case and then make prevention and control scheme.

4. Experimental Results

In this paper, data is historical meteorological and environmental data about tobacco planting base which is located at Bozhou, Anhui. TMV is an example for our research. Due to the occurrence time of TMV, we collect environmental data from mid-April to mid-July in 20 years.

4.1. Extracting Influencing factors

In this paper, the number of samples n sums to 200 and the number of observation indicators of each sample equals to 28. We extract the principal components of environmental data using software Matlab. As shown in figure 3 and figure 4, the total contribution rate is more than 80% and it changes smoothly after the fifth principal component. Therefore, we choose the first five principal components to replace the original indicators. Table 1 shows environmental factors reflected by the first five comprehensive indicators.

4.2. Predicting influencing factors

We choose temperature as first object to be predicted. On the basis of analyzing effect of temperature on TMV, we divided temperature into five grades 0-4. As shown in table2, grade 0 stands for virus suppressed while grade 4 indicates the greatest impact for disease occurrence. According to the above method, we

Table 2: Temperature label set

Temperature(T_d) Range	10°C	10°C ≤ T_d15°C	16°C ≤ T_d20°C	21°C ≤ T_d25°C	26°C ≤ T_d30°C
Range Divided	0	1	2	3	4

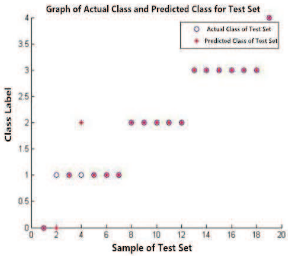

Figure 5: Results of rough selection

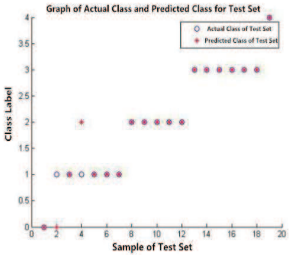

Figure 6: Results of accurate selection

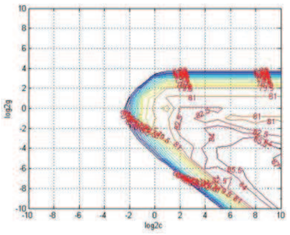

Figure 7: Prediction result

mark whole data set and sort data by labels and divide the 200 records are into 181 for a training set and 19 for a test set.

Before classification and prediction, we need to obtain the optimal value of C and g. We repeat to perform roughly filtering and accurate filtering. The results are shown in figure 5 and figure 6. It always gets the same values of $C = 1024$ and $g = 0.03125$. After parameter selection, we train network of SVM by using the optimal parameters. Figure 7 shows classification and prediction results of test set, the accuracy of which equals to 89.4737% (17/19).

We exploit the same method to predict other factors and mark the properties values according to labels classification. Thereby, the environmental factors of the next year, shown in table 3, are predicted.

4.3. Determining scheme for disease prevention and control

On the basis of PCA, We determine the weight coefficient of each factor via variance contribution rate and factors analysis. According to formula (1), weight coefficients are calculated and shown in table 4.

According to formula (2), the similarity measurement are calculated and the results are shown in table 5. Comparing the environmental status of the target case with that of the similar case to judge disease status. Afterwards, scheme for prevention and control is determined accordingly. The specific solution is described in the following. Firstly, tobacco field need to be worked with deep tillage, weed and the other plant residues need to be basked and remove to reduce casual factors of TMV. Then, choosing tobacco varieties with disease resistance. Finally, after tobacco seedling, special medical solution can be sprayed on tobacco. Moreover, top dressing, hilling and watering are necessary for boosting tobacco root, strong growth and improving resistance ability.

Table 3: The values of factors in different periods

Factors Time(Ten Days)	Average Temperature	Minimum Temperature	Average Relative Humidity	Average Rainfall	Maximum Rainfall	Average Sunshine Hours	Maximum Soil Moisture	Average Soil Moisture	Maximum Soil Temperature
First	0	0	1	0	0	3	1	0	3
Second	1	0	2	0	1	2	2	0	2
Third	2	1	2	2	4	2	4	2	2
Forth	2	1	3	1	2	2	3	3	3
Fifth	3	2	2	0	1	4	3	3	1
Sixth	3	2	2	0	0	2	4	1	3
Seventh	3	2	4	2	3	2	2	4	1
Eighth	4	1	1	1	3	1	2	2	3
Ninth	4	2	2	1	2	4	2	3	2
Tenth	4	2	1	1	4	1	3	2	1

Table 4: Normalized weighted coefficient

Factors Extracted	Comprehensive Index Coefficient	Index Weight
(Average Temperature)	0.0792	0.0818
(Minimum Temperature)	0.0796	0.0821
(Average Relative Humidity)	0.0917	0.0946
(Average Rainfall)	0.1339	0.1382
(Maximum Rainfall)	0.1130	0.1167
(Average Sunshine Hours)	0.1196	0.1235
(Maximum Soil Moisture 10cm)	0.1326	0.1369
(Average Soil Moisture 20cm)	0.1263	0.1303
(Maximum Soil Temperature)	0.0928	0.0958
Sum	0.9688	1

Table 5: Analysis of similar attribute factors

Ten Days	Similar Year	Location	Minimum Value	Disease Status	Ten Days	Similar Year	Location	Minimum Value	Disease Status
firsts	2003	Qiao Cheng	0.7927	I	sixth	2006	Xia Yi	0.8745	III
second	2004	Li Xin	0.6732	II	seventh	2004	Li Xin	0.6612	II
third	2002	Wo Yang	0.7792	II	eighth	2002	Qiao Cheng	0.5489	I
forth	2007	Lu Yi	0.5682	II	ninth	2015	Qiao Cheng	0.7842	I
fifth	2009	Qiao Cheng	0.6843	III	tenth	2007	Qiao Cheng	0.9872	II

5. Conclusions

This paper proposes the model for tobacco disease prevention and control using knowledge graph and CBR and achieves accurate prediction for occurrence of TMV at Bozhou, Anhui. However, the location of tobacco growth is limited and has lowered universality of the model in the paper. In order to validate universality of the model, we plan to add the new regions in our next research.

References

[1] F. Jingyuan, Distribution and Identification of the tobacco fungal disease in Chongqing, Southwest University, 2014.

[2] X.Y Niu, et al., Inhibition of fungi polysaccharide on cucumber mosaic virus, Journal of Northwest A&F University (Nat. Sci. Ed.) 41 (2013) 103–108.

[3] Y.L Zhang, Research on automatic identification system of tobacco disease, Shandong Agricultural University, 2015.

[4] B. Cho, K.J. Kim, J.W. Chung, CBR-based network performance management with multi-agent approach, Cluster Computing 20 (2017) 1–11.

[5] M. Cindy, S. Jay, S. Frank, Toward Case-Based Reasoning for Diabetes Management: A Preliminary Clinical Study and Decision Support System Prototype,Computational Intelligence 25?(2010) 165–179.

[6] H. Paulheim, Knowledge graph refinement:a survey of approaches and evaluation methods,Semantic Web 8 (2017) 489–+.

[7] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A review of relational machine learning for knowledge graphs, Proceedings of the IEEE 104 (2015) 11–33.

[8] F.M. Aliyu, A. Uyar, Evaluating search features of Google Knowledge Graph and Bing Satori Entity types, list searches and query interfaces. Online Information Review 39(2015) 197–213.

[9] H. Abdi, L.J. Williams,Principal component analysis, Wiley Interdisciplinary Reviews Computational Statistics 2 (2010) 433–459.

[10] F. Mordelet, J.P. Vert, A bagging SVM to learn from positive and unlabeled examples,Pattern Recognition Letters 37 (2014) 201–209.