Prediction of Multiple Diseases of Soybean in Complex Environment Based on Improved Apriori Algorithm

Xiaonan Hu¹, Fangyi Deng¹, Yu Zou¹, Yan Guo¹²*, and Yuchao Fang³

¹College of Information Engineering, Sichuan Agricultural University, Ya’an, Sichuan, China
²Key Laboratory of Agricultural information engineering of Sichuan Province, Sichuan Agricultural University, Ya’an, Sichuan, China
³College of electronics and information engineering, Liaoning University of Technology, Jinzhou, Liaoning, China

*Corresponding author e-mail: 14403@sicau.edu.cn

Abstract: The existence and persistence of soybean diseases are not conducive to the effective operation of the global soybean market. Many detection and prediction methods have been used to prevent and detect soybean diseases, but the practicability of these methods has always been a big challenge for researchers due to there are too few variables in the prediction model, which show bad prediction effect of soybean disease in complex environment. In this paper, the popular Apriori algorithm in data mining is used to analyze the common disease data of soybean in complex environment, so as to achieve the goal of early prediction and control of soybean disease. The variables used in this paper are the characteristic factors of 35 kinds of Soybean under 18 common diseases. The experimental results show that the improved Apriori algorithm can complete the better prediction of soybean diseases in complex environment, so as to reduce the impact of diseases on soybean yield, which is of great significance for economic development, agricultural production and other fields.

1. Introduction

Soybean is one of the important food crops in China. It has a long history, rich in nutritional value and rich in plant protein. It is often used to make various bean products, extract soybean oil and extract protein. In the future, soybean will occupy a higher and higher proportion of agricultural market. It is reported that China needs to import more than 90 million tons of soybeans every year, while Brazil, the United States and Argentina are the main exporters. In 2016, China's soybean imports cost 34 billion US dollars in foreign exchange, more than 200 billion yuan. This shows that it is difficult to meet the demand of soybean in China, partly because of the low production. In the growth process of soybean, the yield will be affected by many factors, among which diseases and insect pests are the main factors.

The traditional identification method of soybean diseases and insect pests relies on people's manual inspection, which not only consumes a lot of time, but also wastes a lot of manpower. Along with time, there are more and more agricultural data, which appears some problems in data acquisition [1]. Even if the data is obtained, how to use the data to identify the diseases and pests remains to be studied. If we can effectively predict the occurrence of diseases and insect pests, to some extent, we can prevent them, so as to improve the yield of soybean. With the development of data processing technology,
some data mining methods have developed rapidly. By means of modern data mining technology, this paper studies and analyzes the data of soybean diseases and insect pests in the past, trying to find out the relationship between plant characteristics and the occurrence of diseases and insect pests, and making effective disease prediction for the future soybean planting industry, which plays significant role in advance prevention and improved soybean yield. This goal has become the research direction of many agricultural scholars.

Due to the progress of data mining technology, it is necessary to mine and search frequent patterns in data set to extract effective information from a large number of data and find the relevance between data. Apriori algorithm was proposed by Agrawal and r. Srikant in 1994. It is a layer by layer search iterative algorithm [2]. It is often used to find correlation and association rules in a large number of data sets, and extract potential links hidden in a large number of data sets [3].

In this paper, 18 common diseases of soybean plants were investigated, and 35 characteristics of soybean plants were studied by association rule analysis in order to find out the relationship between the internal factors and the disease regularity of soybean plants as well as help people to effectively prevent soybean diseases.

2. Literature review
Soybean diseases and insect pests have been affecting the yield. Liu et al. studied soybean yield in Northeast China in recent 15 years to find the interaction between root diseases, pests and yield [4].

Eastburn et al. observed the occurrence probability of soybean downy mildew by changing the concentration of carbon dioxide and ozone, sought the relationship between them, so as to and draw a conclusion that the occurrence probability of soybean downy mildew was influenced by environmental factors [5]. Figueredo et al. established a probability model through Bayesian network to verify the effectiveness of predicting the spraying time and quantity of fungicides to prevent soybean rust and helped farmers to reasonably use fungicides to prevent soybean diseases [6]. Fall et al. found out the occurrence rule of the disease with the analysis of the correlation between the temperature, precipitation and other factors and the stem rot of Sclerotinia sojae, and carried out cost-effective management of soybean planting [7]. Wagner et al. established two prediction models of soybean rust and compared the prediction results of the two models on soybean yield [8]. Herrmann et al. measured the spectrum of leaves and canopy in soybean growing season, and discussed the effect of fusarium-wilt on soybean yield [9]. By analyzing the evaluation function based on the difference between high yield and low yield, Umejima et al. found the influencing factors of soybean growth [10].

Generally speaking, in the past, when people analyzed soybean diseases, they did not use data mining technology to find the law between the various factors and the occurrence of soybean diseases during the growth period. In previous studies, most researchers only considered the influence of 2-3 factors on a soybean disease, but did not collect data to analyze. In this paper, the characteristics of soybean diseases are fully considered, and some characteristic factors associated with soybean diseases are found from more than 30 characteristic factors. Finally, good results are obtained.

3. Model and experimental research

3.1. data preprocessing
The variables used in this study include 18 diseases and 35 characteristic factors which are obtained from UCI. Firstly, the data are screened, the unknown quantity in some factors is removed, and the remaining data are named. These 18 diseases are: diaporthe-stem-canker, charcoal-decompose, root rot caused by rhizoctonia, rot caused by phytophthora, stem rot of brown disease, powdery mildew, downy mildew, brown spot, bacterial blight, bacterial abscess, purple spot of seed, anthracnose, phylosticta leaf spot, leaf black spot, leaf frog ophthalmopathy, stem wilt caused by permeable pods, cysticercosis, plant injury caused by herbicides. The 35 characteristic factors are temperature, disease area, leaf shape, seed size and shape, root condition, etc. The 35 characteristic factors of soybean are shown in Table 1.
Table 1. The 35 characteristic factors of soybean.

| Variable | Description | Value |
|----------|-------------|-------|
| C1       | Date        | April, May, June, July, August, September, October |
| C2       | Plant-stand | Normal, Less than normal |
| C3       | Precip      | Less than normal, Normal, Greater than normal |
| C4       | Ambient Temperature | Less than normal, Normal, Greater than normal |
| C5       | Hailstorm   | Yes, No |
| C6       | Crop-history | Different-lst-year, Same-lst-year, Same-lst-two-years, Same-lst-several-years |
| C7       | Area-damaged | Scattered, Low-areas, Upper-areas, Whole-field |
| C8       | Severity    | Minor, Pot-severe, Severe |
| C9       | Seed-thousand metric tons | None, Fungicide, Other, |
| C10      | Germination | 90-100%, 80-89%, Less than 80% |
| C11      | Plant-growth | Normal, Abnormal |
| C12      | Soybean leaves | Normal, Abnormal |
| C13      | Soybean leafspots-halo | Absent, Yellow-halos, No-yellow-halos |
| C14      | Soybean Leafspots-marg | Wrong-side-marg, No-wrong-side-marg, Less than 1/8, Greater than 1/8, |
| C15      | Soybean Leafspot-size | Less than-1/8, Greater than-1/8, Don't apply |
| C16      | Soybean Leaf-shred | Absent, Present |
| C17      | Soybean Leaf-malf | Absent, Present |
| C18      | Soybean Leaf-mild | Absent, Upper-surface, Lower-surface |
| C19      | Soybean Stem | Normal, Abnormal |
| C20      | Lodging     | Yes, No |
| C21      | Stem-cankers | Absent, Below-soil, Above-soil, Above-second-none destructive examination, |
| C22      | Canker-lesion | Don't adjust, Brown, Dark-brown, Brighter-than-dark, Tan |
| C23      | Fruiting-bodies | Absent, Present, |
| C24      | External decay | Absent, Firm-and-dry, Watery, |
| C25      | Mycelium    | Absent, Present, |
| C26      | Int-discolor | None, Brown, Black |
| C27      | Sclerotia   | Absent, Present, |
| C28      | Fruit-pods  | Normal, Diseased, Few-present, Not apply |
| C29      | Fruit spots | Absent, Colored, Brown-withered/Brighter-than-dark-specks, Distort, Not apply |
| C30      | Soybean Seed | Normal, Abnormal |
| C31      | Mold-growth | Absent, Present, |
| C32      | Seed-discolor | Absent, Present, |
| C33      | Seed-size   | Normal, Abnormal |
| C34      | Shrveling Leaves | Absent, Present, |
| C35      | Soybean Roots | Normal, Rotted, Galls-cysts |
Because there are many characteristic factors, not all characteristic factors have influence on the occurrence of the disease. In this paper, the corresponding multiple linear regression models are established for each 18 diseases.

It is found that the contribution rate of some characteristic factors to the disease is almost 0, which can be ignored. The multiple linear regression equation is as follows:

\[ y_a = \beta_0 + \beta_1 x_{ia} + \beta_2 x_{2a} + \cdots + \beta_k x_{ka} + \epsilon_a \]  

where \( \beta_0, \beta_1, \ldots, \beta_k \) is the undetermined parameter and \( \epsilon_a \) is the random variable.

Next, according to the above-mentioned multiple regression model, the disease the feature factors are removed, the remaining feature factors are retained, and then the association rule data mining are conducted. Apriori algorithm is used to analyze the association.

3.2. Apriori algorithm

Apriori algorithm is an algorithm of mining frequent itemsets by Boolean association rules, which uses a layer by layer search iterative method. In the soybean disease detection data in this paper, firstly, by scanning all databases, accumulating the count of each item, and recording the items that meet the minimum support degree set by rules, results may find the set of frequent 1-items in the soybean disease data, which is recorded as \( I_1 \). In the set \( I_1 \), the set \( I_2 \) of frequent 2-items is found. Then using \( I_2 \) to find \( I_3 \) until no more frequent k-itemsets can be found. The minimum support degree is represented by \( sup_{min} \). Association rules are measured by confidence and support. The support of item set is expressed by \( sup \), and the calculation formula is as follows:

\[ sup(x) = \frac{\text{count}(x \subset T)}{|D|} \]  

where the \( |D| \) represents the number of transactions in the transaction database.

In order to measure the distance between variable data, Euclidean distance formula used in this study, is as follows:

\[ d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]  

3.3. Experimental research

In the study of this paper, the 35 variables mentioned above are set as the first item, and the name of disease is set as the second item. First, suppose the minimum support degree as 15% and the minimum confidence degree as 70%, and then the correlation between the soybean disease results and each attribute are obtained, as shown in Table 2.

According to the experimental data in Table 2, the conclusions are as follows. The occurrence of disease phytophthora-rot is related to abnorm-plant-growth, abnorm-stem, none-int-discolor and abnorm-leaves. The occurrence of disease frog-eye-leaf-spot is related to abnorm-stem, leafspot-size, no-yellow-leafspots-halos, leafspots-marg and absent-leaf-shread. The occurrence of disease alternarialeaf-spot is related to leafspot-size, hail, absent-leaf-shread, absent-fruiting-bodies, absent-leaf-mild and absent-mold-growth. The occurrence of disease brown-spot is related to yellow-leafspots-halos, absent-leaf-mild, absent-mold-growth, absent-seed-discolor and absent-leaf-malf. The occurrence of disease brown-stem-rot is related to present-leaf-shread, hail, norm-fruit-pods, absent-fruiting-bodies, norm-seed-size, absent-shrivel and norm-roots. The occurrence of disease anthracnose is related to norm-fruit-pods, none-int-discolor, norm-roots and absent-leaf-malf.

4. Results and discussion

In the process of soybean growth, disease is an important factor affecting its harvest. If the disease can be predicted in advance by a certain method, then the corresponding disease prevention work can be carried out, and finally achieve a desired effect achieve the expected goal on the yield.
Table 2. Association rules and results of soybean diseases and characteristic factors.

| Front item | Back item | Sup. | Con. |
|------------|-----------|------|------|
| erer_2 and yiyi_1 and yijiu_1 and erliu_0 and yier_1 and ershi_0 | phytophthora-rot | 15.5 | 88.8 |
| yijiu_1 and yiwu_1 and yisan_2 and yisi_0 and yiliu_0 and ershi_0 | frog-eye-leaf-spot | 15.2 | 71.0 |
| yijiu_1 and yiwu_1 and yisi_0 and yiliu_0 and sanyi_0 and ershi_0 | leaf frog | 15.2 | 71.0 |
| yiwu_1 and hail and yiliu_0 and ersan_0 and yiba_0 and sanyi_0 | ophthalmopathy | 19.0 | 77.5 |
| yisan_2 and erjiu_0 and hail and yiliu_0 and ersan_0 | alternaria-leaf-spot | 19.5 | 75.6 |
| yisi_0 and hail and yiliu_0 and sanshi_0 and ersan_0 | alternaria-leaf-spot | 19.0 | 75.0 |
| yisan_2 and yiba_0 and sanyi_0 and saner_0 and ershi_0 and yiqi_0 and sanwu_0 | brown-spot | 30.5 | 76.9 |
| gt-precip-normal and erba_0 and yiba_0 and saner_0 and ershi_0 and sanwu_0 | brown-spot | 30.0 | 70.5 |
| yisi_0 and erba_0 and yiba_0 and ershi_0 and yiqi_0 and sanwu_0 | brown-spot | 33.5 | 70.1 |
| yijiu_1 and hail and erba_0 and ersan_0 and sansan_0 and sansi_0 and yiqi_0 and sanwu_0 | brown-stem-rot | 18.4 | 70.8 |
| erba_1 and yijiu_1 and gt-precip-normal and yiliu_0 and yiba_0 and erliu_0 and yiqi_0 and sanwu_0 | anthracnose | 19.1 | 80.9 |
| yiwu_0 and erjiu_0 and yiba_0 and yier_1 and saner_0 and sansan_0 | bacterial-blight | 15.5 | 71.4 |
| yiliu_0 and yiba_0 and erba_0 and saner_0 and yiqi_0 | charcoal-rot | 15.5 | 71.4 |
| yiyi_1 and yijiu_1 and yisan_0 and temp and sanshi_0 and yiliu_0 | diaporthe-stem-canker | 15.5 | 71.4 |
| gt-precip-normal and yiliu_0 and erba_0 and ersan_0 and ershi_0 | downy-mildew | 15.5 | 71.4 |
| yiwu_1 and yisan_2 and yier_1 and ersan_0 and sanwu_0 and sansi_0 | phyllosticta-leaf-spot | 15.5 | 71.4 |
| yisi_2 and gt-precip-normal and sanwu_0 and sansan_0 | rhizoctonia-root-rot | 15.5 | 71.4 |

The identification and prediction of soybean diseases and insect pests is of great economic significance to the actual production practice, and plays an important role in reducing the occurrence of soybean diseases and improving soybean yield. With the development of modern information technology, intelligent agriculture is also developing rapidly. There are still some defects in the prediction of soybean diseases and insect pests. In this paper, the Apriori algorithm is used to analyze the characteristic factors of soybean diseases, to extracting and discretizing the data of 35 characteristic factors of soybean, to changing the minimum support degree and the minimum confidence degree, and finally to find the relationship between the disease and each factor, which can not only verify the known soybean disease law, but also find out the hidden information behind the law. According to the above results, in the following soybean planting links, we can reasonably allocate various resources, so as to prevent diseases and insect pests in advance, and ultimately bring huge economic and social benefits. In general, this method provides a simple and effective way for the analysis and prevention of soybean diseases.
For the study of soybean diseases in this paper, we can increase more variables it is advised that more variables are increased, such as precipitation, gas concentration and other factors in the current month. In addition, the research methods used in this paper can not only analyze the relationship between soybean diseases employ correlation analysis of soybean diseases, but also expand to other fields to achieve different purposes. For example, Apriori algorithm is used to analyze the correlation of other plant diseases and insect pests, so as to find out the occurrence rules, prevent in advance and improve crop benefits.

5. Acknowledgments
The authors are extremely grateful to the journal editorial team and reviewers who provided valuable comments for improving the quality of this article. This research was funded by Social Science Foundation of Sichuan Province in 2019, grant number 19GL030 and the construction of practice teaching system of information technology specialty in agricultural and forestry colleges with the characteristics of agricultural informatization, grant number JG2018-348.

References
[1] Cheng-Kun G, Yan-Zhong L, Ying-Yi C. Major Issues and Missions in Agricultural Big Data[J]. Journal of Anhui Agricultural Sciences, 2014.
[2] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases[C]/Acm sigmod record. ACM, 1993, 22(2): 207-216.
[3] Awadalla M H A, El-Far S G. Aggregate function based enhanced apriori algorithm for mining association rules[J]. International Journal of Computer Science Issues (IJCSI), 2012, 9(3): 277.
[4] Liu X, Herbert S J. Fifteen years of research examining cultivation of continuous soybean in northeast China: a review[J]. Field Crops Research, 2002, 79(1): 1-7.
[5] Eastburn D M, Degennaro M M, Delucia E H, et al. Elevated atmospheric carbon dioxide and ozone alter soybean diseases at SoyFACE[J]. Global Change Biology, 2010, 16(1): 320-330.
[6] Figueiredo G V C, Fantin L H, Canteri M G, et al. A Bayesian Probability Model Can Simulate the Knowledge of Soybean Rust Researchers to Optimize the Application of Fungicides[J]. International Journal of Agricultural and Environmental Information Systems (IJAEIS), 2019, 10(4): 37-51.
[7] Fall M L, Boyse J F, Wang D, et al. Case study of an epidemiological approach dissecting historical soybean Sclerotinia stem rot observations and identifying environmental predictors of epidemics and yield loss[J]. Phytopathology, 2018, 108(4): 469-478.
[8] Igarashi W T, De França J A, De Aguiar E S M A, et al. Application of prediction models of asian soybean rust in two crop seasons, in Londrina, Pr[J]. Semina: Ciências Agrárias, 2016.
[9] Herrmann I, Vosberg S, Ravidran P, et al. Leaf and canopy level detection of Fusarium virguliforme (sudden death syndrome) in soybean[J]. Remote Sensing, 2018, 10(3): 426.
[10] Umejima K, Arimitsu F, Ozawa S, et al. Optimal Pattern Mining from Time-Series Cultivation Data of Soybeans for Knowledge Discovery[C]/Proceedings of the Workshop on Time Series Analytics and Applications. ACM, 2016: 19-24.