Bayesian Networks Modeling for Crop Diseases

Chunguang Bi and Guifen Chen

College of Information & Technology, Jilin Agricultural University, Changchun, China
Bi_chunguan@126.com, guifchen@163.com

Abstract. Severe large-scale diseases in agricultural regions have caused significant economic damage. In order to improve crop yields, we develop a framework to predict the occurrence of crop diseases. In the presence of risk and uncertainty, this paper focuses on finding out the best pest control decision-making program which is based on the Bayesian network. The paper describes the flowchart of a Bayesian network and the principles used to calculate the conditional probabilities required in it. The practice proves that BN is an effective tool for crop disease.

Keywords: Bayesian network; Modeling; Crop Diseases; Inference.

1 Introduction

With the development of animal husbandry and processing industry, the demand for maize is growing fast. Jilin province is the main production area of spring maize and the national commodity grain base, with the corn acreage of 2 million hm2, nearly 10% of the national grain acreage. Corn borer is the most devastating disease in maize production and its occurrence ofte gets affected by lots of factors such as meteorological weather conditions and so on. Bayesian network is one of the most effective theoretical models for uncertainty knowledge expression and reasoning. It has not only a solid basis for probability theory, but also a perfect correspondence with technical knowledge structures. So we use Bayesian network to model the crop diseases.

2 Bayesian Networks Introduction

A Bayesian network, also known as belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependences via a directed acyclic graph (DAG). It could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. We may use mathematical symbols to represent a Bayesian network model as follows:

B= (V, E, P)
among :
V= {V1, V2 …Vn}   set of random variables
E= {ViVj|Vi,Vj∈V}   set of directed edge
P= {P (Vi|V1, V2… Vi-1),Vi∈V}   conditional probability table

D. Li, Y. Liu, and Y. Chen (Eds.): CCTA 2010, Part I, IFIP AICT 344, pp. 312–320, 2011.

Variables can be the abstraction of any problem, to represent interesting phenomenon, components, state or property, etc., with certain physical and practical significance. Directed edges show the dependent or causal relations among variables, the arrow of the edge representing the direction of causal influence (from parent node to child nodes), disconnected nodes representing the variables which these nodes corresponding to are conditional independent. Conditional probability table lists all possible conditional probabilities each node related to its parent. Probability shows the strength or confidence between child nodes and the parent. Probability of independent node called prior probability.

Bayesian networks can be understood in two ways: first, Bayesian networks express conditional independent relations between each node. We can directly get conditional independent relations and dependent relations from Bayesian networks; Bayesian networks also express joint probability distribution of events in another form. According to the structure of Bayesian network and condition probability table (CPT), we can get the probability of each basic event (a combination of all attributes) quickly. Bayesian network consists of two parts. One part is a directed acyclic graph, in which each node represents a random variable; each arc (connection of two nodes) represents a probability of dependence. If one arc starts from node A to node B, A is the parent node and B is the child node. Given the parent nodes, each variable is independent of the non-subclass nodes in the graph. Variables can take discrete values or continuous values. They can correspond to the actual variables or hidden variables in the data set to form a relationship.

Bayesian networks combine the dependent relations with the probability, the prior knowledge with the sample information; overcome many conceptual and calculating difficulties in the rule-based system in graphical way. Combining with statistical techniques makes Bayesian networks advanced in data analysis. Compared with decision-making tree, artificial neural networks and the density estimation methods, the advantages of Bayesian networks are as follows:

(1) Once the Bayesian network is determined, new variables can be easily added on the basis of the current structure.

(2) Bayesian network is very suitable for handling uncertainty and incomplete data sets. It uses probability theory to express the correlation between the variables, also could learn and reason under a limited, incomplete and uncertain information condition. For traditional algorithm of supervision, all possible input data must be clear. If one input data was lost, there would be some deviation for the model. In order to solve this problem, Bayesian networks get the sum or integration of all the probability of possible values.

(3) Bayesian network itself is a kind of uncertainty causal relationship model. Bayesian network is different with other decision-making models. Through graphic visualization of knowledge, it is an expression of probability knowledge, also a reasoning model, with a proper description of the causal relationships between networks variables.

(4) As combination of data and prior knowledge in a probability approach, Bayesian networks better reflect the over-fitting of the model.

3 Bayesian Networks Modeling

3.1 Modeling Process

The construction of Bayesian network is a very complex process, a good Bayesian network will directly affect the correctness. In building a Bayesian network, we first need to identify the main variables and their relationships which lead to the corn borer attack, and on this basis can we build a Bayesian network model; then, we specify the conditional probability distribution for each node in the graph. After we get the preliminary conditional probability table (CPT), the system can further amend the conditional probability according to experts' experience and the actual experimental data. The process of constructing a Bayesian network is as follows:

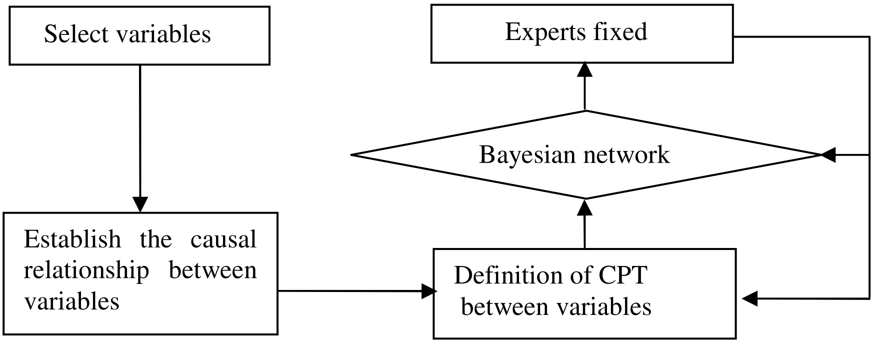
Select variables
Experts fixed
Bayesian network
Establish the causal relationship between variables
Definition of CPT between variables

Fig. 1. Flowchart of Construct Bayesian network

3.2 Problem Description

Corn borer, a major corn pest, leads to above 40% rate of corn victimization every normal year with a 10-15% reduction in output, and above 70% rate of corn victimization in serious year with a 20% reduction. Reinforcing the study on meteorological conditions of corn borer, establishing Bayesian networks to predict this attack, will be of great significance for prevention work.

The outbreak of corn borer is affected by following variables:

(1) Mar-Jul accumulated temperature (ACCT). Corn borer, a kind of cold-blooded animal, with a primitive nervous system, is less able to regulate the body temperature itself. Therefore, its body temperature basically depends on the temperature of external environment.

(2) Mar-Jul average maximum temperature (MAXT). In Mar-Apr, the average temperature is still low; the overwintering larvae mainly use the highest temperature to get more calories to accelerate development process.

(3) Mar-Jul average minimum temperature (MINT). The minimum temperature is higher and this is conducive to the growth of corn borer.

(4) Mar-Jul 5 CM ground temperatures (GT).

(5)Feb-May cumulative sunshine duration (CSD). Corn borer is very sensitive to photoperiod. It does not hibernate or diapause in the long-day. More hours of sunshine, higher ground and air temperature will do well to the development of corn borer.

(6) Jan-Jun precipitation (MP). High precipitation, high soil humidity, low temperature is not conducive to the development of overwintering larvae and worse to their pupation and eclosion.

If we can grasp the inner relations of these factors, we can predict the occurrence of core borer more accurately. Now, we build a Bayesian network prediction model.

3.3 Establish Causal Relationships Table

First, establish the causal relationship table between variables and the occurrence of core borer. The contents of the table are identified by experts. The arrows to right show that rows attributes are father nodes and the columns attributes are child nodes. On the contrary, the arrows to left show that columns attributes are the father nodes and the rows attributes are child nodes. Two-way arrows show the relationship can not be determined and two-way arrows with a slash show that there is no relationship between the two.

Table 1. Demand forecasting of causality

	ACCT	MAXT	MINT	GT	CSD	MP	Occur
ACCT	——	←/→	←/→	←/→	←/→	←/→	→
MAXT	←/→	——	←/→	←/→	←/→	←/→	→
MINT	←/→	←/→	——	←/→	←/→	←/→	→
GT	←/→	←/→	←/→	——	←	←	→
CSD	→	←→	←/→	→	——	←/→	→
MP	←/→	←/→	←/→	→	←/→	——	→
Occur	←	←	←	←	←	←	——

As we can see from the table, experts couldn't identify the relations between Feb-May cumulative sunshine duration (CSD) and Mar- Apr average maximum temperature (MAXT). So we got opinions from another two experts who suggest adopting evidence synthesis method.

Evidence synthesis is an effective method in dealing with uncertain reasoning problem. Synthesis of evidence, from the theory, first proposed by Dempster, is promoted and developed by Shafer in dealing with uncertainty reasoning theory. The earliest synthesis evidence formula of the data theory is the Dempster formula[1]:

$$m(\varphi) = 0 \tag{1}$$

$$m(A) = \frac{1}{1-K} \sum_{A_i \cap A_j \cap A_k \cap \ldots = A} m_1(A_i) m_2(A_j) m_3(A_k) \cdots\cdots \tag{2}$$

$$K = \sum_{A_i \cap A_j \cap A_k \cap \ldots = \phi} m_1(A_i) m_2(A_j) m_3(A_k) \cdots \tag{3}$$

In Dempster synthetic formula, 1-K, a normalized factor, completely abandon the conflict between evidence and distribute all probability related with these conflicts to an empty set. This is a very strict with operations. Therefore, the results often perverse in the synthesis of high degree conflict evidence.

In order to solve the problem, Yager proposed a new synthetic formula:

$$m(\varphi) = 0 \tag{4}$$

$$m(A) = \sum_{A_i \cap A_j \cap A_k \cap \ldots = A} m_1(A_i) m_2(A_j) m_3(A_k) \ldots \quad A \neq \phi, X \tag{5}$$

$$m(X) = \sum_{A_i \cap A_j \cap A_k \cap \ldots = \phi} m_1(A_i) m_2(A_j) m_3(A_k) \ldots + K \tag{6}$$

Through the evidence synthesis method above, we got the results:

Table 2. Evidence synthesis results

	CSD→MAXT	CSD←MAXT	CSD←/→MAXT	uncertainty
E1	0.5	0	0.3	0.2
E2	0.4	0	0.4	0.2
E3	0.4	0	0.3	0.3
compound	0.61	0	0.36	0.03

Synthesis results show that the parent node is SSD and child node is MAXT, then we build the Bayesian network model as shown in figure2:

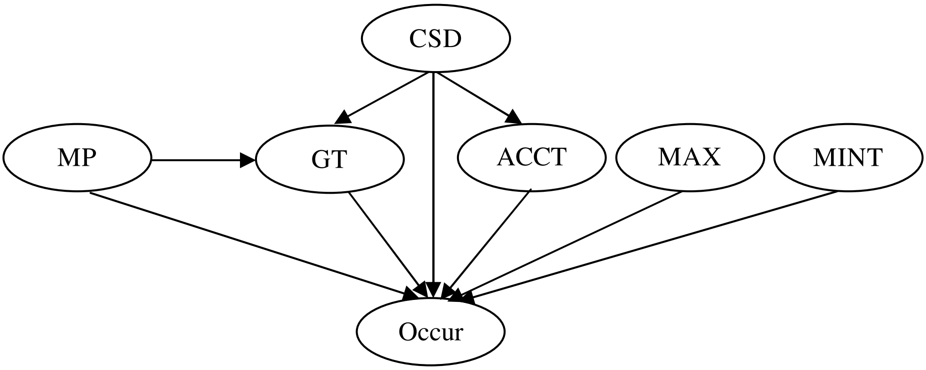

Fig. 2. Prediction problem of Bayesian network initial model

3.4 Determine Conditional Probability Tables

This stage is to determine each variable's state and qualitative probability information. This information could be gotten from experts and relevant literature [2].

After variables selection, we need to confirm each variable’s state. In order to reduce the size of Bayesian networks, we should limit the number that each variable contains and only select those interesting state that users prefer to. When select the state, make sure that the selected state is mutex to each other.

Determine the scope of variables:

(1) Mar-Jul accumulated temperature (ACCT). {2540-2550℃.d, 2551-2560℃.d, 2561-2570℃.d}

(2) Mar-Jul average maximum temperature (MAXT). {24.1-25.0℃, 25.1-26.0℃}

(3) Mar-Jul average minimum temperature (MINT). {20.1-23.0℃, 23.1-26.0℃}

(4) Mar-Jul 5 CM ground temperature (GT). {19.1-20.0℃, 20.1-21.0℃, 21.1-22.0 ℃}

(5) Feb-May cumulative sunshine duration (CSD). {1001-1100h, 1101-1200h, 1201-1300h}

(6) Jan-Jun precipitation (MP). {51-100mm, 101-150mm}

In most cases, experts only orally describe the causality. The probability we obtained is usually frequency and some qualitative but not quantitative terms such as "may", "often", "very little". Therefore, engineers need to make the transition from qualitative to quantitative. Through test, Renooij found that these oral quantifiers have some relations with actual probability[3]. She created a probability benchmarking, as shown in figure3:

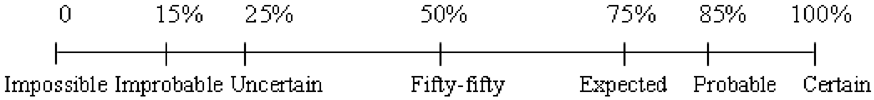

Fig. 3. Probability benchmarking

The probability benchmarking, making discrete psychological variables to continuous, help us determine probability distribution more accurately. Using this benchmarking, experts could clearly get the relationship between psychology and quantitative probability. Experts make a mark which corresponds to a certain number, and this number equals to the probability distribution. The average result from that of different experts is final probability distribution of the node [4]. With probability benchmarking, identify the conditional probability of each node in the network. For example as shown in table 3:

Table 3. Conditional Probability Table

CSD ACCT	1001-1100	1101-1200	1201-1300
2540-2550	0.24	0.25	0.23
2551-2560	0.39	0.37	0.37
2561-2570	0.37	0.38	0.40

3.5 Bayesian Network

A Bayesian network is shown in figure 4 (with conditional probabilities):

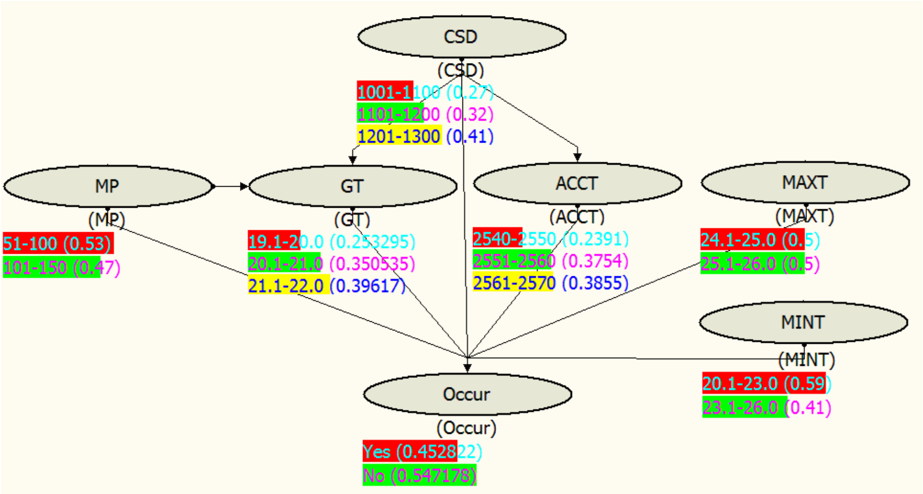

Fig. 4. Core borer Bayesian network

However, the most important application of Bayesian network is to reasoning backward according to actual events. For example, if we know the ground temperature is 19.1-20.0, we can enter the evidence that 'GT' =19.1-20.0. The conditional probability tables already tell us the probabilities of corn borer's occurrence (0.452822).As figure 5 shows:

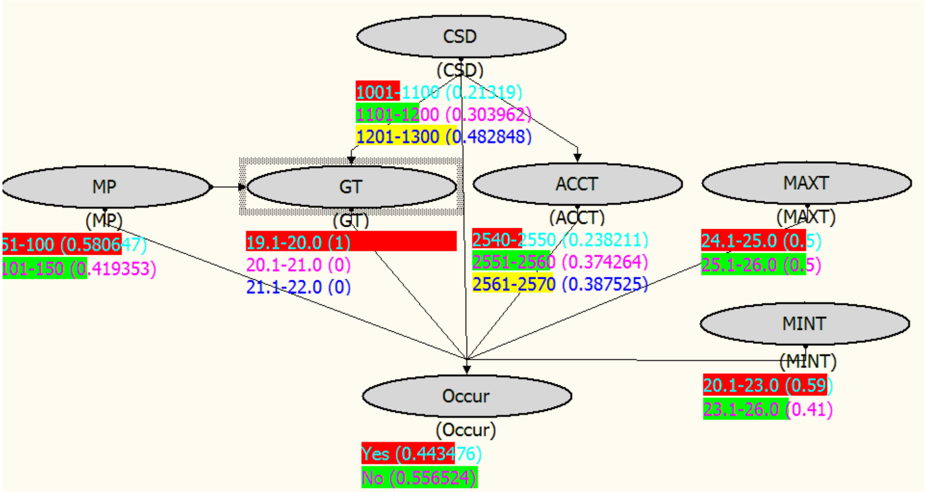

Fig. 5. Reverse reasoning

If we know corn borer breaks out, then we can enter the evidence that 'occur' =yes and we can observe the result to get other nodes' revised probability.

Bayesian network make the dependant relations of different variables more explicit. In general there may be relatively fewer direct dependencies (modeled by arcs between nodes of the network) and this means that many variables are conditionally independent.

The existence of unlinked (conditionally independent) nodes in a network drastically reduces the possibility of calculating all the probabilities. Usually, all the probabilities can be calculated by joint probability distribution. However, we need to do some simple calculation when there are some independent nodes[5].

After completing basic construction of Bayesian network, we still need to adjust conditional probabilities and revise the model, then improve the accuracy.

4 Implementation of Crop Disease Forecast System Based on Bayesian Network

Construction of a Bayesian network needs the communication and cooperation of domain experts and Bayesian network experts[6].

There are two types of nodes in the Crop Disease forecast system--- disease nodes and meteorological nodes. The disease nodes are Boolean variables, which contain states of 'occur' and 'dis-occur'' while meteorological nodes may indicate multiple states.

This BN is a multi-layer network.

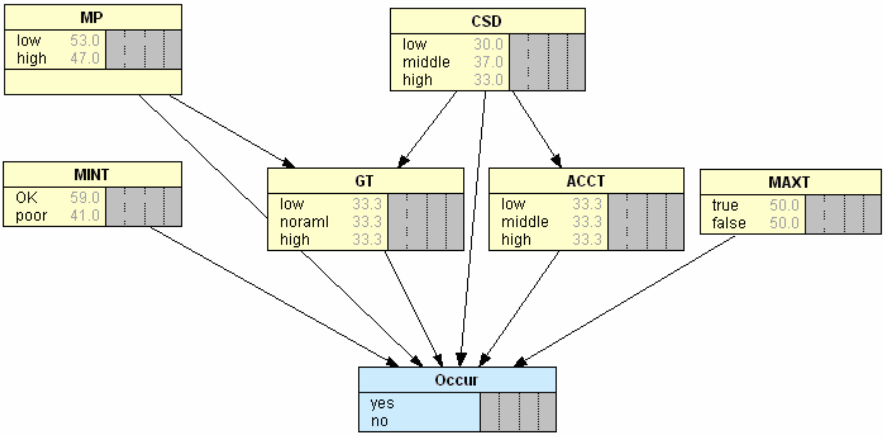

Fig. 6. Inference results of the BN

5 Conclusions

As Bayesian network applies probability knowledge with complex systems' reasoning, and experts always supply the probability value we need. So, this paper starts building BN model with these experts knowledge. On this basis, we establish the Crop Disease Forecast System. There is a large number of uncertain knowledge in the diagnosis of crop pests, while Bayesian network has unique advantage in dealing with uncertain factors.

In following study, we found that adopting application of Ontology in building Bayesian networks will make the whole network perfect. In addition, Bayesian network could revise the conditional probability table and make more accurate prediction of crop disease for its self-learning function.

References

1. Mingzhu, X., Guangju, C.: A Modified Combination Rule of Evidence Theory. Electronic Journals 3(9), 1715–1716 (2005)
2. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs, pp. 18–32. Springer, Heidelberg (2007)
3. Du, T., Zhang, S., Wang, Z.: Learning Bayesian Networks from Data by Particle Swarm Optimization. Journal of Shanghai Jiao Tong University E-11(4) (2006)
4. Friedman, N., Linial, M., Nachman, I., Peter, D.: Using Bayesian networks to analyze expression data. Computational Biology 7, 601–620 (2000)
5. Friedman, N., Koller, D.: Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. Machine Learning 50 (2003)
6. Guifen, C., Helong, Y.: Bayesian Network and Its Application in Maize Diseases Diagnosis. In: Proceedings of First IFIP TC 12 International Conferences on Computer and Computing Technologies in Agriculture (2007)