A SIMPLIFIED BAYESIAN NETWORK MODEL APPLIED IN CROP OR ANIMAL DISEASE DIAGNOSIS

Helong Yu [1,2] , Guifen Chen [1,2,*] , Dayou Liu [1]

[1] *College of Computer Science and Technology, Jilin University, Changchun, Jilin Province, P. R. China, 130012*

[2] *College of information Technology， Jilin Agricultural University,Changchun, Jilin Province, P. R. China 130118*

[*] *Corresponding author, Address: College of Computer Science and Technology，Jilin University， Changchun，Jilin Province, P. R. China, 130012， Tel: +86-431-84532775, Fax: +86-431-84532775, Email:guifchen@163.com*

Abstract: Bayesian network is a powerful tool to represent and deal with uncertain knowledge. There exists much uncertainty in crop or animal disease. The construction of Bayesian network need much data and knowledge. But when data is scarce, some methods should be adopted to construct an effective Bayesian network. This paper introduces a disease diagnosis model based on Bayesian network, which is two-layered and obeys noisy-or assumption. Based on the two-layered structure, the relationship between nodes is obtained by domain knowledge. Based on the noisy-model, the conditional probability table is elicited by three methods, which are parameter learning, domain expert and the existing certainty factor model. In order to implement this model, a Bayesian network tool is developed. Finally, an example about cow disease diagnosis was implemented, which proved that the model discussed in this paper is an effective tool for some simple disease diagnosis in crop or animal field.

Keywords: bayesian network, crop or animal, disease diagnosis, noisy-or, certainty factor

1. INTRODUCTION

There exists a lot of uncertainty phenomenon and problem in agriculture, especially in the field of crop or animal disease. Uncertainty in disease diagnosis is more extensive and complex. So, in order to create an effective disease diagnosis system, uncertain knowledge must be dealt with.

Bayesian network is a kind of probabilistic graphical model(P.Larranaga,S. Moral.2008), namely a combination of probability theory and graph theory. By graph, Bayesian network can represent knowledge naturally and intuitively. By probability, Bayesian network can solve uncertain problem.

Bayesian network can reason in dual direction, which can be used in both prediction and diagnosis. Also, Bayesian network can use prior knowledge effectively and make the best of knowledge form domain expert.

However, the construction of the Bayesian network needs a great amount of probability. Generally, this is very complicated and difficult.

As far as this paper is concerned, a simplified Bayesian network is proposed and used in disease diagnosis system.

The simplified model is a two-layered structure and obeys noisy-or assumption.

In order to decrease the difficulty of constructing Bayesian network, domain knowledge and existing certainty factor knowledge base are used.

Then a Bayesian network tool was developed and an example about cow disease diagnosis was implemented.

2. METHODS AND TECHNOLOGY

2.1 The defination of Bayesian network

Bayesian network is a binary group, namely S=<G，P>，in which:

（1）G is a directed acyclic graph. The nodes correspond to random variable and the directed arcs represent probabilistic dependence between variables. The meaning of the arc from x to y is that x have direct influence to y.

（2）P is the set of local probability distribution , $P=P\{P(x|\pi_x)\}$ is conditional probability, which is used to measure the strength of causal dependencies and π_x is the set of parent nodes of x.

2.2 A two-layer model of disease diagnosis

There are two types of nodes in the disease Diagnosis System，which are disease nodes and symptom nodes(Fig.1).

Both disease nodes and symptom nodes are Boolean variables. Disease nodes contain states: ‘happen’ and ’not happen’. Symptom nodes contain states: ‘find’ and ‘not find’(P.J.F Lucas.2005, Radim Jirousck.1997).

This BN is a two-layer network, in which the upper layer is composed of disease nodes and the lower layer is composed of symptom nodes. Obviously， the arc direction is from disease nodes to symptom nodes.

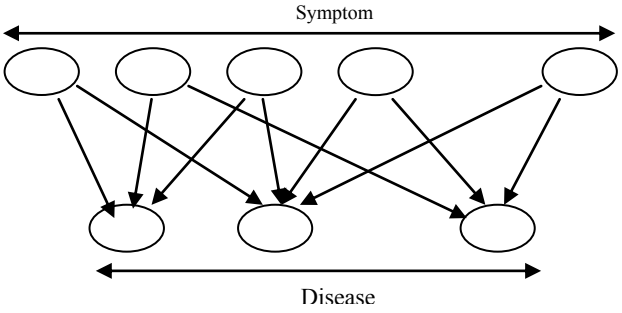

Fig. 1 two-layered structure

2.3 Noisy-or technology

One problem faced in knowledge engineering for Bayesian networks is the exponential growth of the number of parameters in their conditional probability tables (CPTs). The most common practical solution is application of the noisy-OR (or their generalization, the noisy-MAX) gates, which take advantage of independence of causal interactions and provide a logarithmic reduction of the number of parameters required to specify a CPT.

This model has three assumptions: Parents and child are Boolean variables.Inhibition of one parent is independent of the inhibitions of any other parents.All possible causes are listed. In practice this constraint is not an issue because a leak node can be added (a leak node is an additional parent of a Noisy-or node).

Now, we can have a definition of noisy-or:

（1）A child node is false only if its true parents are inhibited.

（2）The probability of such inhibition is the product of the inhibition probabilities for each parent.

（3）So the probability that the child node is true is 1 minus the product of the inhibition probabilities for the true parents.

According to Fig. 2, Given the reason nodes, suppose that $p_i = P(F = true \mid H_i = true)$, we can get the following conclusions:

If all reason nodes are false $P(F = true) = 0$;

if only one reason node is true, $P(F = true) = p_i$;

else $P(B = false) = \prod_{i=1}^{m} (1 - p_i)$,in which m is the counts of true reason nodes.

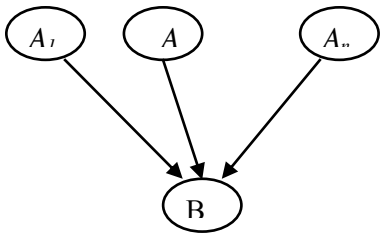

Fig.2 : Noisy-or model

Generally, for node having k parent nodes, if use 'Noisy-or', it needs $\Theta(k)$ parameters, if not, it needs $\Theta(2^k)$ parameters. Obviously, BN is simplified by the use of noisy-or technology.

3. THE CONSTRUCTION OF BN

Before being deduced, Bayesian network must be constructed. As we know, Bayesian network has two parts: structure and CPT, so the process of constructing Bayesian network is to construct structure and CPT (David J.Spiegelhalter.1993).

Construcion of a Bayesian network for a domain problem needs communication and cooperation of Bayesian network expert, domain expert and BN software tool(P.J.F Lucas.2000).

There are three methods to construct Bayesian network: manual construction, machine learning and combination of them. This article mainly introduces manual method, which constructs Bayesian network by domain expert elicitation(E.Charles,J.Kahn, etc.1997).

3.1 Elicitation of BN structure

In this process, variables and relationships between them should be determined.

First, select variable set. It is important to limit the number of variables. So, it is necessary to choose important variables which are

(1) Target variables: or Query variables, they are outputs of net and what we want to know.

(2) Observation variables: or Evidence variables, they are inputs of net and used to reason states of query variables.

In the relationship between cause and effect, there are two cases:

(1)Multi-causes, one effect.

(2)One cause, multi-effects.

According to the Bayesian formula

$$P(X_1 \ldots X_n) = P(X_n \mid X_{n-1}, \ldots, X_i) P(X_{n-1} \mid X_{n-2}, \ldots, X_1) \ldots P(X_2 \mid X_1) P(X_1) = \prod_{i=1}^{n} (X_i \mid Parents(X_i))$$

we can find that the right sequence of adding nodes is:

(1)Add symptom nodes.

(2)Add disease nodes that be influenced directly by symptom nodes.

(3)Repeat the above two steps until all nodes are added.

3.2 Elicitation of Condition Probability Table

In most of the cases, each relationship between an influenced node and its parents were estimated separately from different sources of data or—in a few cases—given by experts. The conditional probabilities were calculated from the sources in different ways (Kristian Kristensen etc.2002).

In our model, conditional probability table is obtained through expert knowledge, existing certainty factor model and parameter learning.

3.2.1 Elicitation of CPT from parameter learning

Suppose there is much disease diagnosis data, in which there exists the value of symptom node and diseases node(Table 1). S_i=f represents this symptom is found and S_i=nf represents it is not found. D_i=h represents this disease have happened and D_i =nh represents it did not happen. The amount of data record in the table is m.

Table 1. Some data record about symptom and diseases

label	S_1	S_2	…	S_p	D_1	D_2	…	D_q
1	f	nf	…	f	h	nh	…	nh
2	nf	f	…	f	nh	h	…	nh
….	…	…	…	…	…	…	…	…
m	a	a		a	b	b		b

According to the table, we can achieve the following probability.

（1）prior probability

Assume the amount of $S_i = true$ in the table is r, $P(S_i = true) = \frac{r}{m}$

（2）conditional probability

Assume the amount of ($S_i = true$ and $S_j = true$) is s, the amount of $S_j = true$ is t, $P(S_i = true \mid S_j = true) = \frac{s}{t}$

3.2.2 Elicitation of CPT from domain expert

In this process, the state and qualitative probability of each variable should be determined. This can be obtained by domain expert and literature.

Humans tend to think in categories ("likely", "unlikely", etc.) rather than in terms of exact probability (Chard T.1991). So the transformation from qualitative probability to quantitative probability is necessary, which was achieved by consulting domain expert repeatedly. The corresponding relationship can be represented by binary group <Qualitative Probability, Quantitative Probability >, for example, <always, 0.99>, <often, 0.78>,etc.

3.2.3 Elicitation of CPT from certainty factor model

There exists some knowledge in certainty factor model. In order to use this part of knowledge in Bayesian network, some actions should be taken(F.trai.1996, Kevin B.Korb, Ann E.Nicholson.2006).

In the certainty factor model, the knowledge given by domain expert is rule-based, and measurement for the belief is certainty factor, that is:

IF A Then $B : CF(B \mid A)$

Definition of certainty factor (CF):

$$CF(B \mid A) = \begin{cases} \dfrac{P(B \mid A) - P(B)}{1 - P(B)}, & if \quad P(B \mid A) > P(B) \\ 0, & if \quad P(B \mid A) = P(B) \\ \dfrac{P(B \mid A) - P(B)}{P(B)}, & if \quad P(B \mid A) < p(B) \end{cases}$$

However, in the Bayesian network, uncertainty is measured by probability. So, in order to construct Bayesian network, it needs to transform CF to probability(Nevin Lianwen Zhang.1996, Wang ronggui etc.2004). From above formula, we can obtain:

$$P(B|A) = \begin{cases} CF(B|A)(1 - P(B)) + P(B), if \ CF(B|A) \geq 0 \\ (CF(B|A) + 1)P(B), if \ CF(B|A) < 0 \end{cases}$$

So, in order to get probability, it needs to know $P(B)$, which is prior probability of node B.

$P(B)$ can be obtained from domain expert, literature, or existing data. If not, assume $P(B)$=0.5,which represents ignorant.

If $P(B)$=0.5,the formula is as follows:

$$P(B|A) = \begin{cases} 0.5 \times CF(B|A) + 0.5, if \ CF(B|A) \geq 0 \\ (CF(B|A) + 1) \times 0.5, if \ CF(B|A) < 0 \end{cases}$$

The transformation from CF model to Bayesian network brings two advantages. One is that the relationship between variables can be demonstrated visually and intuitively, the other is that more information, namely the probability, can be achieved from the Bayesian net.

3.3 The developing of Bayesian network tool

In order to implement this model, a Bayesian network tool was developed. This tool includes two components. One is the Bayesian building component, the other is Bayesian reasoning component. Through the building component, the structure and CPT parameters, namely the knowledge base, can be constructed visually, which are stored in a XML file. Through the reasoning component, the posterior probability of a node can be computed.

4. AN EXAMPLE

In order to verify the disease diagnosis model, an example about cow disease diagnosis is introduced, which is from an existing certainty factor model. Table 2 shows the detail. The column of S represents symptom nodes, D representing disease nodes, CF representing certainty factor, P representing corresponding probability, IP representing inhibition probability (noisy probability).

In this model, we assume the prior probability of symptom nodes and disease nodes is 0.5

Table 2. Part of data from a cow disease knowledge base of certainty factor model

S	D	CF	P	IP
tbsz21	w13	0.25	0.625	0.375
bcs11	w13	0.15	0.575	0.425
Ydyc02	w13	0.2	0.6	0.4
sz05	w13	0.2	0.6	0.4
tbcz12	w13	0.2	0.6	0.4
sz02	w14	0.15	0.575	0.425
ydyc19	w14	0.3	0.65	0.35
bcs21	w14	0.2	0.6	0.4
ydyc02	w14	0.2	0.6	0.4
tbcz12	w14	0.15	0.575	0.425
tsjc18	w15	0.3	0.65	0.35
tbcz20	w15	0.2	0.6	0.4
tbsz15	w15	0.3	0.65	0.35
tbcz13	w15	0.1	0.55	0.45
bcs11	w15	0.1	0.55	0.45

According to the noisy-or assumption and data from Table 2, we can get conditional probability table of the Bayesian network about cow disease diagnosis.

From domain knowledge and table 3, the structure of cow diseases diagnosis can be obtained, which is constructed by the building component. See Fig.3.

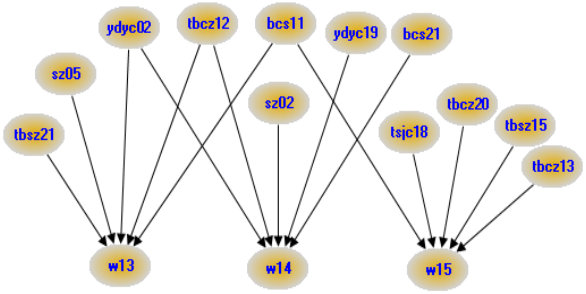

Fig.3 The Bayesian structure of cow disease diagnosis

The posterior probability of a node can be computed by reasoning component. In practice, the reasoning result is conformed with the domain knowledge and existing certainty factor model.

5. CONCLUSIONS

BN is a strong tool for representing and dealing with uncertain knowledge. There exists a lot of uncertainty knowledge in crop or animal disease diagnosis. So it is natural to use BN to build disease diagnosis system. However, in general the data for disease is scarce, so a simple and effective Bayesian network is needed.

While building the two-layered BN, Noisy-or model and transformation from CF to probability are used to decrease network scale and simplify the network structure.

In running the cow disease diagnosis system, we find that the reasoning result is conformed with the solution given by domain expert and the existing certainty factor model, which proves that it is effective to use Bayesian network to represent and deal with uncertain knowledge in disease diagnosis.

ACKNOWLEDGEMENTS

Funding for this research was provided by National “863”project “Research and application of agricultural knowledge grid” (No.2006AA10Z245-2)and China National “863”project “Research and application of precision working system for maize” (No.2006AA10A309) .

REFERENCES

Chard T.Qualitative probability versus quantitative probability in clinical diagnosis: a study using a computer simulation. Med Decis Making. 1991 Jan-Mar;11(1):38-41.

David J.Spiegelhalter. Bayesian Analysis in Expert Systems, Statistical Science, 1993.Volume 8, Issue 3: 219-247.

E.Charles,J.Kahn, etc. Construction of a Bayesian network for mammographic diagnosis of breast cancer, Comut.Biol.Med,1997:19-29.

F.trai..A Bayesian network for predicting yield response of winter wheat to fungicide programs, Computers and electronics in agriculture.1996: 111-121.

Kevin B.Korb, Ann E.Nicholson. Bayesian Artificial Intelligence, CRC Press.2006:.225-260

Kristian Kristensen etc, The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides, Computers and Electronics in Agriculture,2002(33):197-217

Nevin Lianwen Zhang. Exploiting causal independence in Bayesian network inference, Journal of artificial intelligence, 1996: 301-328.

P.J.F Lucas. Bayesian network modeling through qualitative patterns. Artificial Intelligence, 2005: 233-263.

P.J.F Lucas. Certainty-Factor-Like structures in Bayesian belief networks, Knowledge-based systems,2001: 327-335.

P.Larranaga，S.Moral. Probabilistic graphical models in artificial intelligence. Applied soft computing.2008:1-18.

Radim Jirousck. Constructing probabilistic models, International journal of medical informatics 1997(45): 9-18.

Wang ronggui etc, From Certainty Factor Model to Bayesian Network. computer science ,2004,31(10).