DEVELOPMENT OF A DATA MINING APPLICATION FOR AGRICULTURE BASED ON BAYESIAN NETWORKS

Jiejun Huang [1,*], Yanbin Yuan [1], Wei Cui [1], Yunjun Zhan [1]

[1] School of Resource and Environmental Engineering, Wuhan University of Technology, Wuhan, China, 430070

** Corresponding author, Address: School of Resource and Environmental Engineering, Wuhan University of Technology, 122 Luoshi Road, Wuhan,hubei, 430070, P. R. China, Tel:+86-27-62421206, Email: Hjjtk@21cn.com*

Abstract: Data mining is a process by which the data can be analyzed so as to generate useful knowledge. It aims to use existing data to invent new facts and to uncover new relationships previously unknown even to experts. Bayesian network is a powerful tool for dealing with uncertainties, and has a widespread use in the area of data mining. In this paper, we focus on development of a data mining application for agriculture based on Bayesian networks. Let features (or objects) as variables or the nodes in Bayesian network, let directed edges present the relationships between features, and the relevancy intensity can be regarded as confidence between the variables. Accordingly, it can find the relationships in the agricultural data by learning a Bayesian network. After defining the domain variables and data preparation, we construct a model for agricultural application based on Bayesian network learning method. The experimental results indicate that the proposed method is feasible and efficient, and it is a promising approach for data mining in agricultural data.

Keywords: data mining, Bayesian network, model, agriculture.

1. INTRODUCTION

Data mining provides an information technology to develop and utilize the data; it is very helpful for making decision by extracting regulations, patterns and models from large databases. Data mining has proved to have surprisingly broad application. Many scholars pay attention to application of data mining for agriculture (Lee et al., 1998; Bajwa et al, 2004; Abdullah et al., 2006; Andujar et al., 2006). Expert systems and geographical information systems have been used to help with an implementation of a land suitability evaluation model (Kalogirou, 2002). It shows that applying Dempster-Shafer Theory in image classification can yield thematic maps with accuracies that can support their operational use, and the potential for applying soft-classification procedures based on the Dempster-Shafer Theory of Evidence was demonstrated (Lein, 2003). Fuzzy set and interpolation techniques are applied for land suitability evaluation for maize in Northern Ghana (Braimoh et al., 2004). A linear mixture model (LMM) approach is applied to classify land covers in the eastern Nile delta of Egypt. It indicates that the LMM is a promising approach for distinguishing the different land cover types and to classify the different vegetation types using Landsat ETM+ data (Ghar et al., 2005). Furthermore, computational intelligence is used to agriculture. A novel model of land suitability evaluation is built based on computational intelligence (Liu et al., 2005).

Bayesian networks are the method for uncertainty reasoning and knowledge representation that was advanced at the end of the 20th Century. It is a probabilistic graphical model, which has been used for probabilistic reasoning in expert systems. Bayesian networks proved to have surprisingly broad applications, such as medical diagnoses, image interpretation, pattern recognition, in particular, knowledge discovery and data mining (Heckerman, 1997; Helman et al., 2004). Bayesian network as an application to agriculture is studied by some authors. An application of belief networks to assess the impact of climate change on potato production is used as an illustration, and used simulated data from a mathematical model which forecasts the impact of climate change on potato production (Gu et al., 1994). The potential of Bayesian networks to assess the yield response of winter wheat to fungicide programmes has been shown, and a Bayesian network that fits the experimental data has been produced (Tari, 1996). An example of application of Bayesian networks for modeling landuse changes is proposed (Benferhat et a.l, 2004). In addition, Bayesian networks have been used in the sustainable planning of the Eastern Mancha aquifer. The results can be obtained through Bayesian networks are the partial substitution of groundwater with surface water, the improvement of irrigation efficiency and the adequate control of water use (Martin, 2007).

In this paper, we focus on development of a data mining application for agriculture based on Bayesian networks. We believe that it will provide a potentially useful tool in the domain. In next section, we discuss Bayesian network for data mining. In section 3 we discuss an information-theoretical approach to learning Bayesian network. In section 4, we propose an example of application of Bayesian networks for agricultural land gradation. Finally, we draw conclusions and present the future works.

2. BAYESIAN NETWORKS FOR DATA MINING

A Bayesian network is a directed acyclic graph representing the causal relationships between variables that associate conditional probability distributions to variables given their parent variables. It is represented at two levels, qualitative and quantitative. At the qualitative level, we have a directed acyclic graph in which nodes represent variables and directed arcs describe the conditional independence relations embedded in the model. At the quantitative level, the dependence relations are expressed in terms of conditional probability distributions for each variable in the network. Suppose a data set D is given, which is defined by n variables V={ V_1 , V_2 , … , V_n}, each variable respond to a node, let G represents a DAG, L is a set of directed links, P is a set of conditional probability distributions associated with every node. Using Bayes chain rule, and let Pa_i is the set of parents of the variable V_i, so we can get the joint probability distribution:

$$P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i \mid Pa_i) \qquad (1)$$

A Bayesian network represents a joint probability distribution of a set of random variables of interest. Information we can obtain comes in the form of evidence about a subset of variables. The basic task of inference is to update the joint probability distribution of the variable set conditioned on the given set of evidence. Bayesian networks as data mining have several aspects:

(1) Causality discovery. Bayesian networks give a graphical representation of the domain problems and results. As for a given problem, we assume that it is fully described by a finite set of random variables. Each variable is fully defined in a finite frame, i.e. set of all possible states. The set of relations among variables is called the structure of the Bayesian network, which represents the qualitative knowledge about the problem domain. If all the variables are identified, and each variable is defined with a frame, we say that we have a Bayesian network structure which is a representation of the causality of the variables. Here the structure *S* refers to a set of directed edges.

$$S = \{U \rightarrow V \mid U, V \in \mathbf{V}\} \qquad (2)$$

Where for each directed link $U \rightarrow V$, *U* is a parent of *V*. On the ground that, we can discover the causality between variables by finding the directed links which we called learning Bayesian network structure.

(2) Uncertainty reasoning. Inference in data mining is rigorous based on Bayesian probability theory. This rigor will not decrease along with inference passages of any length. Inference always remains rigorous irrespective of the size of the network and how far the information variables are from the target variables in the network. There is no need to distinguish between forward reasoning and backward reasoning. Bayesian network are

capable of learning, the structure can be constructed by training the data sets. And the resulting model has a clear interpretation. The conditional probabilities associated with relations correspond to the quantitative aspects of the expert knowledge. Furthermore, information in the form of evidence may come into the network from any location (variables or nodes) in the network, and such incoming information will then be propagated throughout the rest of the network.

(3) Adaptive learning. Adaptivity is closely related to learning. When adaptivity is concerned, learning becomes a decision problem. Bayesian networks can incorporate expert knowledge and historical data for decision-making. In addition, data mining based on Bayesian network can be controlled by artificial conditioning. For example, in predictive mining, the data model usually consists of a large sample set of cases, with each case containing a certain number of features. Formulating a predictive problem trains the system to "learn" which patterns match predefined criteria within existing cases and which don't, and to accept or reject new cases based on these criteria.

3. LEARNING BAYESIAN NETWORK STRUCTURE FROM DATA

The main obstacle for using Bayesian networks is to construction the domain model. Creation of Bayesian models is a complex task involving participation of a knowledge engineer and domain experts. A Bayesian method was developed for the induction of Bayesian networks from data (Cooper et al., 1992). An information-theory based approach to learning Bayesian networks from data was provided (Cheng et al., 2002). And many authors have studied on learning Bayesian networks and proposed some relative algorithms (Huang et al., 2005; Tsamardinos et al., 2006). Learning a Bayesian network from data involves two tasks: Estimating the probabilities for the conditional probability tables (learning parameters) and deriving the structure of the network. Although ideally the structure and parameters should be learned simultaneously, finding the optimal structure of the network is the most difficult part of the whole problem. It comprises a heuristic search through the space of possible structures. Candidate structures are evaluated by calculating how well the network fits the data.

Fundamental to various approaches to learning Bayesian networks are statistical learning theory, Bayesian learning theory, and computational learning theory. In this section, we introduce an information-theoretical approach: Minimum Description Length (MDL) criterion. Learning Bayesian network structure from a data set can be regarded as a problem of explaining the given set of data using a learned Bayesian network as a model. Given a data set **D** out of the data space $\mathcal{D}$, the MDL principle selects the best model M$_{best}$ out of the model space $\mathcal{M}$ with

$$M_{best} = \underset{M \in \mathcal{M}}{\arg\min}\ L(\mathbf{D}, \mathrm{M}) \tag{3}$$

The model space $\mathcal{M}$ is a Cartesian product of the structure space $\mathcal{S}$ and the parameter space Θ_S given structure S

$$M = (\mathcal{S}, \Theta_S) = \mathcal{S} \times \Theta_S, \qquad S \in \boldsymbol{S} \tag{4}$$

The joint description length of a given data set and a model – a Bayesian network- can be defined as

$$\begin{aligned} L(\mathbf{D}, M) &= L(\mathbf{D}, \Theta, S) \\ &= L(\mathbf{D} \mid \Theta, S) + L(\Theta \mid S) + L(S) \\ &= L(\mathbf{D} \mid \Theta) + L(\Theta \mid S) + L(S) \end{aligned} \tag{5}$$

For structure learning purpose, we may only need to evaluate

$$L(\mathbf{D}, S) = L(\mathbf{D} \mid S) + L(S) \tag{6}$$

According to the MDL criterion, a data set can be represented by the Bayesian network structure, whose description length as

$$L(g, x^N) = H(g, x^N) + \frac{k(g)}{2} \log N \tag{7}$$

Where $H(g, x^N)$ is the empirical entropy for a network structure g

$$H(g, x^N) = \sum_{j \in J} H(j, g, x^N) \tag{8}$$

$k(g)$ is the number of independent conditional probabilities embedded in the network structure g

$$k(g) = \sum_{j \in J} k(j, g) \tag{9}$$

In this sense, learning Bayesian network from data is the problem same as optimal problem. The best model of all alternative models will be the one with the shortest total description length. This information-theoretical approach can avoid explicitly defining the structure prior and easy to be comprehended. Nevertheless, it is difficult for tackling the incomplete or soft data problem and the computation would be too complex if the network has many nodes.

4. DATA MINING BASED ON BAYESIAN NETWORKS FOR AGRICULTURE

In this section, we propose an example of application of Bayesian networks for agricultural land gradation. The data set contains 2 000 cases, part of the data and the variables with theirs status showed in Table1. The domain problem has 6 variables; each of them has several attributes. And one variable responds to one node in the model respectively. The variables and their implications describe as follows. 1) Soil texture: the relative proportions of sand, silt, and clay particles in a mass of soil. 2) Organic matter: consists of plant and animal material that is in the process of decomposing. 3) Gradient: A measure of slope (soil-surface), i.e. the rate of inclination for land topography changes. 4) Drainage: The capability of draining off the water when the field doesn't need the water. The means of draining collectively, as a system of conduits, trenches, etc. 5) Soil pH: indicates the acidity of the soil, it can be determined by having a soil analysis carried out, and has a range approximately from 0 to 14; 6) Land grades: the quality of the agricultural land measured by the natural and economic characteristics.

*Table 1.*Part of data and the variables with theirs status

Land Code	Soil texture	Organic matte	Gradient	Drainage	Soil pH	Land grades
0616	sand	2.06	3	2	6.45	III
0896	silt	1.38	3	2	6.55	IV
1025	sand	2.06	2	2	6.42	III
1380	sand	1.38	2	3	6.42	IV
1620	silt	1.38	2	2	6.42	IV
1698	clay	1.95	2	3	6.42	I
1806	clay	1.38	3	3	6.50	II
1912	clay	1.30	3	3	6.85	II

After defining the domain variables and data preparation, we can get a Bayesian network model for agricultural land gradation by using the approach described in previously section. Based on the model, and learning the parameters of each node with the dataset by using Bayes criterion, the complete network including conditional probability distributions was got.

The experimental study is done on 200 cases to test its validity. The gradation accuracy is 87.5%. Then we compare results given by Bayesian networks with the ones of naive bayes which are simple Bayesian networks, and decision tree method, the results by Bayesian networks is the best one (Table2). The experimental results validate the practical viability of the proposed approach for data mining in agricultural data.

Table 2. The accuracy of result by using different methods

Methods	Test data set	Correct	Accuracy
Naïve bayes	200	156	78.0%
Decision tree	200	166	83.0%
Bayesian network	200	175	87.5%

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we use Bayesian networks as a data mining method for agriculture. Firstly, we presented an overview of Bayesian network for data mining. Secondly, we discussed an information-theoretical approach to learning Bayesian network structure. Then we propose an example of application of Bayesian networks for agricultural land gradation. Furthermore, we compare results given by Bayesian networks with Naive bayes and decision tree. From the practice of applying Bayesian network, it can deal with all kinds of data timely, and has other functions such as agricultural land evaluation and agricultural machine diagnosis. Bayesian networks as data mining for agriculture have several characteristics. Its representation and reasoning can be carried out simultaneously, and combined with prior knowledge and observed data. Moreover, it can overcome the noise of data set, and provide the scientific evidences in decision making for exploiting agriculture resource. It is undoubted that Bayesian networks will be a promising approach for data mining and get the surprisingly success in the application domains. At the same time, Bayesian network is not almighty. It can not obtain satisfying results from small data or sparse data. Therefore, the future work should be focused on how to deal with sparse data and missing values. Furthermore, we will apply Bayesian network to other agricultural domains with Geological information system and remote sensing.

ACKNOWLEDGEMENTS

The work presented in the paper was supported by the National Natural Science Foundation of P. R. China (No. 40601076, 40572166 and 40571128), the Natural Science Foundation of Hubei Province (2007ABA322) and Open Research Fund of State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (WKL(06)0303).

REFERENCES

Abdullah A, Hussain A. Data mining a new pilot agriculture extension data warehouse. Journal of Research and Practice in Information Technology, 2006, 38(3): 229~248

Andujar J M, Aroba J, et al. Contrast of evolution models for agricultural contaminants in ground waters by means of fuzzy logic and data mining. Environmental Geology, 2006, 49(3): 458~466

Bajwa S G, Bajcsy P, Groves P, Tian L F. Hyperspectral image data mining for band selection in agricultural applications. Transactions of the American Society of Agricultural Engineers, 2004, 47(3): 895~907

Benferhat S, Cavarroc M, Jeansoulin R. Modeling landuse changes using bayesian networks. Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, 2004, 615~620

Braimoh A k, Vlek Paul L, Stein Alfred. Land Evaluation for Maize Based on Fuzzy Set and Interpolation, Environmental Management, 2004, 33(2): 226~238

Cheng J, Greiner R, Kelly J, et al. Learning Bayesian networks from data: An information-theory based approach. Artificial Intelligence, 2002, 137(1-2): 43~90

Cooper G F, Herskovits E A. Bayesian method for the induction of Bayesian networks from data. Machine Learning, 1992, 9: 309~347

Ghar M A, Renchin T, Tateishi R, Javzandulam T. Agricultural land monitoring using a linear mixture model. International Journal of Environmental Studies, 2005, 62(2): 227~234

Gu Y, Peiris DR, Crawford J, et al. An application of belief networks to future crop production. Proceedings of the 10th Conference on Artificial Intelligence for applications, San Antonio, Texas, 1994, 305~309

Heckerman D. Bayesian Network for data mining, Data mining and knowledge discovery. 1997, 1: 79~119

Helman P, Veroff R, Atlas S, et al. A Bayesian network classification methodology for gene expression data. Journal of Computational Biology. 2004, 11(4): 581~615

Huang J J, Pan H P, Wan Y C. An algorithm for cooperative learning of Bayesian network structure from data. Lecture Notes in Computer Science, 2005, 3168: 86~94

Kalogirou S. Expert systems and GIS: An application of land suitability evaluation. Computers, Environment and Urban Systems, 2002, 26(2-3): 89~112

Lee S W, Kerschberg L. Methodology and life cycle model for data mining and knowledge discovery in precision agriculture. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 1998, 3: 2882~2887

Lein J K. Applying evidential reasoning methods to agricultural land cover classification International Journal of Remote Sensing, 2003, 24 (21): 4161~4180

Liu Y, Jiao L. Model of land suitability evaluation based on computational intelligence. Wuhan Daxue Xuebao (Xinxi Kexue Ban), 2005, 30(4): 283~287(in Chinese)

Martin de Santa Olalla F, Dominguez A, Ortega F, et al. Bayesian networks in planning a large aquifer in Eastern Mancha, Spain. Environmental Modelling and Software, 2007, 22(8): 1089~1100

Tari F. A Bayesian Network for predicting yield response of winter wheat to fungicide programmes. Computer and electronics in agriculture, 1996, 15: 111~121

Tsamardinos I, Brown L, Aliferis C. The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning, 2006, 65(1): 31~78