Mobile Netw Appl (2017) 22:418–426
DOI 10.1007/s11036-016-0797-2

Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm

Mingyang Liu[1] · **Ming Qu**[2] · **Bin Zhao**[3]

Published online: 23 December 2016

Abstract With the development of social information technology and the increasing of information data in big data era, how to query the required data accurately is becoming more and more important, the purpose of this paper is to establish a model of data mining technology. In this paper, we use the Bayesian network learning model to study the data mining technology. In this paper, a Bayesian network learning model is established, then, the parameters of the recognition and the selection of coefficients are analyzed in detail, after that, the data mining model based on Bayesian computation is deduced, and the reliability of the model is verified by the example of the students. The probability distribution pattern used by Bayes has many advantages in data mining. It further proves the applicability of Bayesian formula, and provides a reference for data mining technology.

Keywords Information technology · Big data · Data mining technology · Bayesian network · Probability distribution

1 Introduction

With the development of computer software and hardware technology, the information degree of the society is constantly deepened. On the one hand, the data processing for social production and life is more and more profound, the data stock is increasing dramatically. On the other hand, a large amount of storage space are not fully utilized properly, most people are drowning in the massive data, but they still have the feeling of lacking knowledge, as a result, the so-called "data explosion and poor information" [1] phenomenon was appeared. As for the reasons, it is due to there has a certain limitation in the information acquisition, the database, the management technology and the related data processing tools, which lead to people can only get a part of the information, so it far from meeting their needs. It is proved that these important information was hide in the kinds of data, this not only reflects the potential relationship between the data, but also has important reference value in the decision-making process. The task of data mining is to find the information that is ignored and then get benefit from it. Data mining is defined as the process of finding the pattern of data, that is, to deal with the data from a large number of incomplete, noisy, fuzzy, random data [2] in the database, and then extract the implicit and unknown knowledge, but these knowledge and information are potentially useful. The work steps are as follows: Firstly, we should determine the mining object, search the relevant internal and external data information, and then select the data that is suitable for data mining applications. Secondly, we should study the data quality and make some preparations, such as data cleaning, integration, selection, transformation, etc. Finally, we need to select the effective algorithm to analyze the huge data space, then dig the data and gather the information we are interested in.

Data mining [3–6] is a very active research field in the database and artificial intelligence field. At the same time, it has been widely used in various databases.

✉ Ming Qu
myliu0427@qq.com

[1] College of Instrumentation & Electrical Engineering, Jilin University, Changchun 130061, China

[2] College of Computer Science and Technology of Jilin University, Changchun 130012, China

[3] Jilin Animation Institute, Changchun 130012, China

Therefore, the study of intelligent and automatic extraction of valuable knowledge and information from a large amount of data, that is data mining, which has very important theoretical and practical significance, meanwhile, it has broad application prospects. At present, data mining has become a hot research topic which has an urgent need, so many researchers at home and abroad have devoted great enthusiasm to this field. In scientific research field, with the increasing of scientific simulation experiments, and the data of different experimental data are scattered in different computers, hence, scientists can hardly find the intrinsic link between the data sources by hand. This is an urgent need to study the corresponding new data mining technology and mining tools to solve these problems. At present, there are many researches on data mining technology abroad, but it is very rare in scientific data area. In China, the research in this field is at the initial stage, and the mature research results are basically blank.

2 The current situation of bayesian algorithm and data mining technology at home and abroad.

A similar term to data mining— Knowledge discovery from database (KDD) [7], this word first appeared in the Eleventh International Conference on artificial intelligence, which held in Detroit in August 1989. After 1993, the United States Computer Association (ACM) held a special meeting each year to discuss the data mining technology, the conference name is ACM SIGKDD International Conference [8] on Knowledge Discovery and Data Mining, KDD conference for short. The scale of the KDD conference is developed from the original symposium to international conference. The research emphasis is also gradually changing from the discovery method to the system application. It makes people tend to pay more attention to the integration of a variety of strategies and techniques, and the mutual penetration of various disciplines. The domestic research on DMKD (data mining and knowledge discovery) is a little late compared with foreign countries, and there is no overall force currently. In 1993, The National Natural Science Fund firstly support the Chinese Academy of Hefei Branch [9] to study the project in the field. At present, the research of data mining is mainly in the University, also, there are some in the research institute or company as well. Researches are generally focused on the learning algorithm, the practical application of data mining and the data mining theory. Most of the current research projects funded by the government, such as the National Natural Science Foundation, the 95 plan, 863 plan, etc. In China, many scientific research institutions and universities are also competing to carry out the basic theory of knowledge discovery and its application, among them, Beijing System Engineering Research Institute has an intensive study on the application of fuzzy method in knowledge discovery. Peking University also carried out research on the data cube algebra, at the same time, the central China University of science and engineering, Fudan University [10] etc. also carried out the optimization and transformation of the association rules mining algorithm. Nanjing University, Sichuan University and other units also explore and study the knowledge discovery of unstructured data and Web data mining. Compared with foreign countries, the domestic research on DMKD (data mining and knowledge discovery) is a little bit late, there is no overall force currently.

In the past, the research of Bayesian networks is mainly divided into two categories, one is based on probability statistics theory; the other is based on information theory. Bayesian methods based on probability statistics include Bayesian averaging and maximum posteriori criterion. Cooper& Herskovits [10] first proposed the Bayes maximum a posteriori method for multi-link structure. This method uses the Bayesian score to find the maximum possible network, that is, using the product of the likelihood function and the prior probability of the structure of the given network structure data, at the same time, the prior probability of the structure is the score criterion. Just like other Bayesian methods, this method has to assume a prior distribution of the structural space. However, since the prior distribution is consistent, which makes the method more similar to ML estimation. By selecting a same prior, a more accurate network without the influence of structure complexity is obtained. In order to avoid the limitation of structure, the feasible method is to use the minimum description length (MDL) criterion. The minimal description length standard (MDL) was firstly put forward by Rissanen [11], and it was as a new criterion for the statistical model then. By using of MDL, it is assumed that the prior value of the network structure is replaced by the description of the structure, and the most important reason is that the length can be calculated. The application of MDL in Bayesian network learning is studied by Bouckaert [12], Later, and the use of MDL to evaluate the structure of learning methods has a further promote. In recent years, many scholars have tried to study the structure of incomplete data and Latent Variable. Whereby, Friedman proposed a method to extend the EM algorithm of the structure learning, which called EM Structural algorithm. In general, Structure EM search it in the structure and parameters of the joint space. However, because of the discontinuity of the structure space, the joint space is not continuous, which makes the EM algorithm may not reach the expected effect. In addition, Chickering [13] will realize the network structure evaluation through transforming an equivalent class space; Gamez & Puerta [14] apply the search optimization of the network structure to the ant colony behavior; Campos & Huete proposed a method based on the

independence criterion, in addition, many scholars tried to use different methods to study the structure as well.

3 Research on data mining technology based on bayesian algorithm

3.1 Conditions for the establishment of bayes algorithm

The condition of Bayes's algorithm is the independence of all time. The so-called conditional independence means that under certain conditions, the occurrence of an event is not influenced by another event. In solving practical problems, if certain prior knowledge is known. That is, the basic conditions was given, so that we can eliminate some factors and the correlation between the results, also, we can effectively save time and energy, greatly improve the efficiency of decision-making. And in the Bayesian network, conditional independence relationship has an important position and function. Bayes network structure is a qualitative description of the problem areas, the network structure of each node (variable) in conditions which known to their parent node is independent of all other non-descendant nodes [15]. According to the conditional independence, the Bayesian network can decompose the joint probability into the product of several conditional probabilities, which can effectively save the storage space of the parameters. At the same time, it makes the probability reasoning tend to be more intuitive and convenient, and it also simplified the process of knowledge acquisition and domain modeling. In the process of quantitative reasoning, the use of conditional independence can reduce the number of prior probabilities, it also can reduce the computational complexity of the reasoning process and improve the efficiency of learning.

3.2 Bayes network

Bayes network generally includes two parts, one is the Bayes network structure, which is a directed acyclic graph (DAG) [16]. Each node in the graph represents the corresponding variable, and the connection between the nodes represents the conditional independence of the Bayesian network. The other part is the conditional probability table (CPT), which is a series of probability values. If a Bayesian network provides sufficient conditional probability values, which can calculate any given joint probability, and we call it a computable. Figure 1 is a Bayesian network with 6 nodes, which expresses a series of conditional independence properties: After the state of the father node is given, each variable is independent of its non-inheritance node in the graph. The graph captures the qualitative structure of probability distribution and is developed to make efficient inference and decision making. Bayes networks can represent arbitrary probability distributions, and

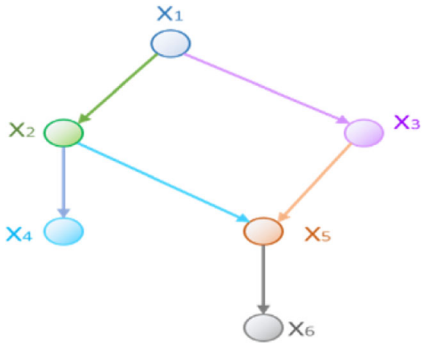


Fig. 1 Bayesian network with six nodes

they can be used to express the distribution of simple structures. Hypothesis for vertexX_i, its parent node set isP_{ai}, the conditional probability of each variable $P(X_i|P_{ai})$ isX_i, and the joint probability distribution of the$X = \{x_1, x_2, L, x_n\}$of the vertex set is calculated as follows:

$$P(X) = \prod_{i=1}^{n} P(X_i|P_{ai}) \tag{1}$$

The simplified joint probability formula in Fig. 1 of the Bayesian network is as follows:

$$\begin{aligned} &P(x_1, x_2, x_3, x_4, x_5, x_6) = \\ &P(x_6|x_5)gP(x_5|x_2, x_3)gP(x_6|x_1, x_2) \\ &\cdot gP(x_3|x_1)gP(x_2|x_1)gP(x_1) \end{aligned} \tag{2}$$

Once the correlation between the propositions is represented by a directed arc, the conditional probability is represented by the weight of the arc. The knowledge about the relation between the static structures of the proposition is expressed. When acquiring a new evidence, the possible values of each proposition should be comprehensively examined, and then define a trust degree of each node as$B(x)$, it can be specified:

$$B(x) = P(X = x_i|E) \tag{3}$$

It indicates that all facts and evidence was provided in the current E conditions, the proposition Xvalue the trust degree ofx_i, and then based on the evidence and the fact to calculate the$B(x)$ trust degree. By using the formula (2), the type joint probability formula, to greatly simplify the calculation of$B(x)$.

3.3 Bayes network learning

Bayes network learning, that is, to find a way to reflect the existing data in the existing database of the dependent relationship between the Bayes network models, the Bayesian network is constructed according to the prior knowledge of the user, which is called a priori Bayesian network. The Bayesian network which is obtained by combining the prior Bayesian network with the data is called a posteriori Bayesian

network, the process of Bayesian network is obtained by a priori Bayesian network is a Bayesian network learning, Bayes network can continue to learn, the last time to learn the Bayes network can become the Bayes network which can be leaned in the next time. Every time before learning, the user can adjust the Bayesian network, so that the new Bayesian network [17] can reflect the knowledge of the data, which was shown in Fig. 2.

Learning based on Bayesian network includes two parts: parameter learning and structure learning. At the same time, according to the different properties of the sample data, each part includes two aspects [18]: Data completeness and data incompleteness. Parameters learning methods are mainly based on classical statistics and Bayesian statistics. The structure learning method is based on the Bayesian statistical measures and the encoding theory. The following introduction is based on the structure of the study.

For the learning process of Bayesian networks, we propose the following 3 assumptions:

(1) Random sample D is complete, that is, no missing data in D;
(2) Parameter variables are independent of each other, that is:

$$p(\theta|S^h) = \prod_{i=1}^{n} p(\theta_i|S^h) \tag{4}$$

$$\sum_{k=1}^{r_1} \theta_{ijk} = 1 \tag{5}$$

$p(X|\theta, S^h)$Is the probability density for the variable i.

$$p(\theta_i|S^h) = \prod_{j=1}^{q_i} p(\theta_{ij}|S^h) \tag{6}$$

In Bayesian networks, we should firstly define a random variableS^h, and the database$D$ is a random sample assumption from the network structure S, then gives a priori probability distribution $p(S^h)$to express the network structure's uncertainty and calculate the posterior probability distribution$p(S^h|D)$. On the basis of Bayesian theory we can get:

$$\begin{aligned} p(S^h|D) &= p(S^h, D)/p(D) \\ &= p(S^h)p(D|S^h)/p(D) \end{aligned} \tag{7}$$

In which, $p(D)$ is a normal constant which has nothing to do with structure learning, $p(D|S^h)$ is a structural likelihood constant, thus, the posterior distribution of the network structure is only required for each possible structure of the data structure. In the premise of no constraint multinomial distribution, parameter independence, by using Dirichlet prior and

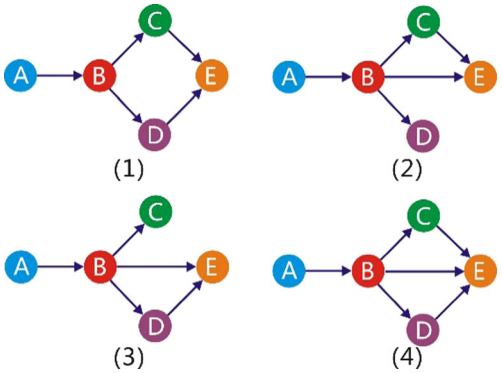


Fig. 2 Bayes network continuous learning plan

data integrity, the data structure is just the product of the likelihood of each pair of structure(i, j), i.e.

$$p(D|S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_1} \frac{\Gamma(a_{ij})}{\Gamma(a_{ij} + N_{ij})} \prod_{k=1}^{r_1} \frac{\Gamma(a_{ijk} + N_{ijk})}{\Gamma(a_{ijk})} \tag{8}$$

When a_{ij}>0, it is the distribution coefficient of Dirichlet, its value relate to $S^h$and$D$. This shows that the joint density $p(D|S^h)$only decided by Dirichlet distribution coefficienta_{ij}, what's more, it indicates that Bayes network learning process is to find the appropriate index coefficienta_{ij}, which can make joint probability $p(D|S^h)$maximum.

In general, it is difficult to rule out the possible network structures of n variables that are larger than the n as a function of the exponential function. Two methods can be used to deal with this problem: "model selection" and "selective model averaging". The former is to choose a "good" model from all possible models, and to take it as the right model; the latter select all possible models for a reasonable number of "good" models, and these models represent all cases [19].

3.4 Bayes parameter selection

Each event of Bayes is independent, but the weight of each event is also different from each other. We simply consider that the parameters are used to specify the weights of the input variables. Therefore, for the absolute value of the parameter in the larger model, if we get the derivative of the differential input value, the derivative tend to be larger. It shows that the prediction value of the model is more sensitive to the change of the input, and the prediction curve is also relatively steep, it does not conform to our intuitive sense of the simple model as well. As the following picture shows, 10 blue dots represent the observed data [20]. We hope to get a smoother curve, although it is not strictly cross through each dot (the green curve in the Fig. 3).

The use of Bayesian methods in the selection of model parameters is described below. Assuming that the observed data are given by$D = \{<x_1, t_1>, \cdots, <x_n, t_n>\}$, the learning objectives are to obtain a set of parameters$W$, which can make

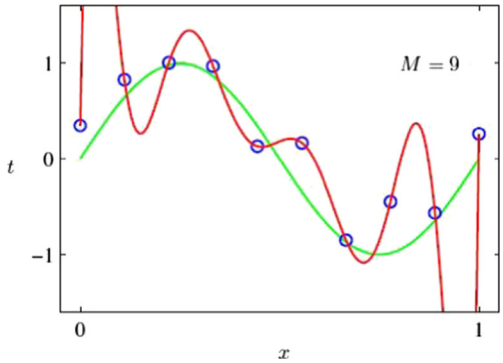


Fig. 3 Coefficient prediction curve

the conditional probability$p(W|D)$ maximum. According to the Bayes formula:

$$p(W|D) = p(W|D)p(W)/p(D) \tag{9}$$

Whereby, $p(W|D)$ is the maximum likelihood estimator, it indicates the coincidence level between the observed data and the model. We assume that the predicted values of the model accord with the$y(x, W)$ as the mean value, setβ as the normal distribution of variance (normal distribution is represented by the symbol$N$), and assume that there are N observation data is independent, then

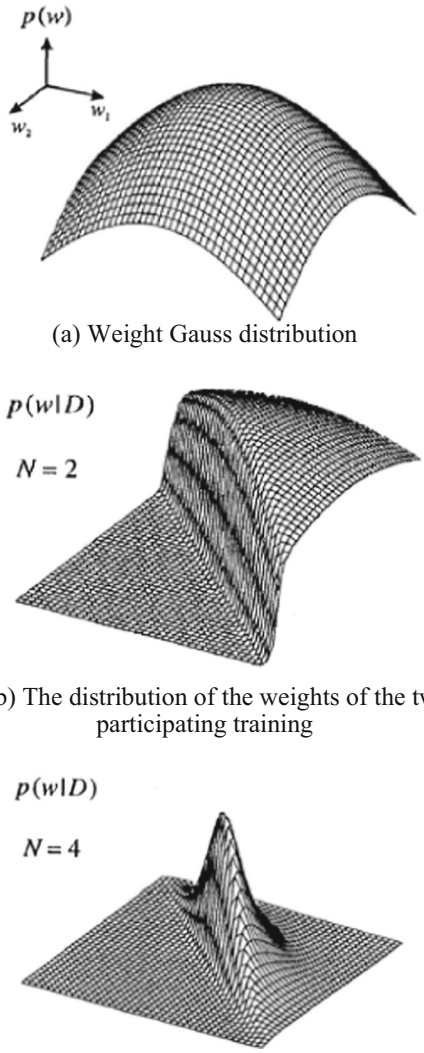
$$p(D|W) = \prod N(t_n|y(x_n, W), \beta)\Big) \tag{10}$$

As for the prior distribution$p(W)$ of the parameters$W$, it is not a uniform distribution. First of all, our human being prefer a simple model, and in the nature, the model which is simpler and its appearance probability is larger. This is because, the more simple things the more stable, the more complex the more sensitive to subtle changes, so, it's hard to be stable. Therefore, the reason why we can observe things, a large probability is dominated by a simple rule. But the smaller the model $\|W\|$ of$W$, the model will tend to be simpler (when$\|W\| = 0$, the model is reduced to a constant). So we assume that$p(W)$ is in normal distribution with 0 as the mean andα as the variance. That is:

$$p(W) = N(W|0, \alpha) \tag{11}$$

Through several graphs we can see the weight of a priori [21], as well as the posttest with the new data is constantly observed after the change (see Fig. 4).

It can be seen that the weight can affect and control the shape of the classification function of the whole network. The data make the posterior probability of the weight space become smaller and smaller, because these weights will make the decision surface toward the wrong direction. At the same time, the posterior probability of the other parts of the weight space is not changed, and the prior distribution is preserved. When the 4 data are involved in training, because there is no



(a) Weight Gauss distribution

(b) The distribution of the weights of the two participating training

(c) 4 data are involved intraining distribution

Fig. 4 Weight distribution of bayes parameter

decision can be very perfect corresponding to their classification, so the most likely solution is a decision making with a special shape.

Therefore, only a very narrow the area of the posterior probability tend to large, and most of the region is very small.

3.5 Bayesian network method applied to data mining

Data mining technology contains many algorithms, and the commonly used methods are: decision tree, genetic algorithm, Bayesian network method, rough set, neural network, etc. Each algorithm has its own features and advantages. The advantage of the decision tree is obvious, it's understandable and intuitive. It is mainly used for classification and induction mining, but in the case of large data and data complexity, the law appears to be inadequate. Genetic algorithm is good at data clustering, and it has a unique advantage in combination optimization problem. Rough set has a very important role in

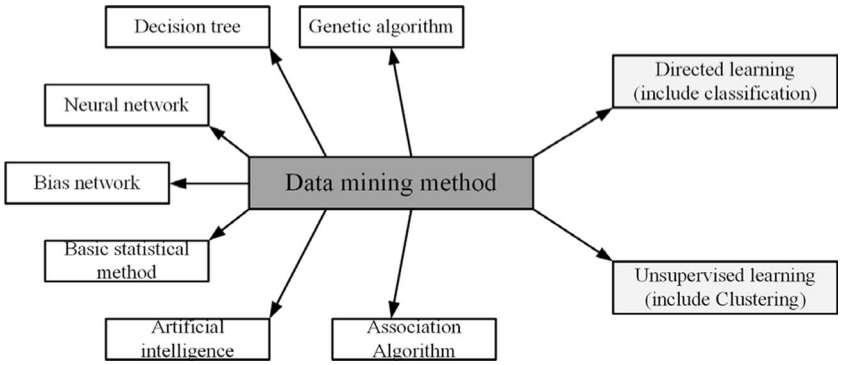


Fig. 5 The main frame of data mining

data mining, it is often used to deal with the problem of ambiguity and uncertainty, the problem of feature induction and correlation analysis, at the same time, using rough set for data preprocessing can improve the efficiency of knowledge discovery. Neural networks can be used to predict the complex problems, which is widely used in the business world [22]. For credit customer identification, stock forecast and stock market analysis, its effect is good. The Bayesian network has the function of classification, clustering, prediction and causal analysis. It is easy to understand, and the forecast effect is good. In the face of large-scale data, it has a unique advantage, as shown in Fig. 5:

Bayesian networks is a kind of decision analysis tool which is developed with the influence diagram, it provides knowledge representation, reasoning and learning methods under uncertainty environment. And it can accomplish the tasks of decision making, diagnosis, prediction and classification. Bayesian network has the functions of classification, clustering, prediction and causal analysis. It is easy to understand, and the forecast effect is good. It has been widely used in speech recognition, industrial control, economic forecasting, and medical diagnosis etc. In recent years, the application of data mining has opened up a new research space. What's more, Bayesian network is the carrier of probability information, and it is the form of joint probability distribution.

A Bayesian network is usually made up of two parts: The first part is a directed acyclic graph, each node represents a random variable, and each arc represents a probability dependent; the second part is a conditional probability table for each attribute (CPT) [23]. Figure 6 gives a simple Bayesian network of 6 Boolean variables. For example, smoking is concerned with lung cancer, and it also influenced by the family history of lung cancer. In addition, the arc is also indicated, when the parent family history of lung cancer (FH) and smoking (S) data was given, the variable lung cancer (LC) conditions are independent of emphysema (E), this means that once the family history of lung cancer (FH) and smoking (S) data is known, the variable emphysema (E) does not provide additional information about lung cancer (LC). Table 1 gives the CPT of lung cancer (LC), for each possible combination of its family history of lung cancer (FH) and smoking (S), the conditional probability of each value of LC is given in the table. It was shown in Fig. 6. For example: P (LC = "no" FH = "no ", S = "yes") = 0.5.

Corresponding to the attribute or variable $X_1, X_2, \cdots, X_n$, the joint probability of the arbitrary tuple$(X_1, X_2, \cdots, X_n)$ was calculated by the following formula:

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} p(X_i|\text{parents}(X_i)) \quad (12)$$

Among them, the value of $p(X_i|\, parents(X_i))$ corresponds to the value of$X_i$in CPT.

3.6 Data mining algorithm based on bayesian network

Since the Bayesian network can take into account the prior information and sample data, and it can make full use of expert knowledge and experience, meanwhile, it can combine the subjective and objective, and has many features that are better than other methods. At present, there isn't a complete algorithm for constructing Bayesian networks in data mining, and a heuristic method is proposed, which is based on sample data in data mining, to construct the algorithm of Bayesian network.

1) At first, according to the data mining's objectives and tasks, we should make a data analysis and variable selection,

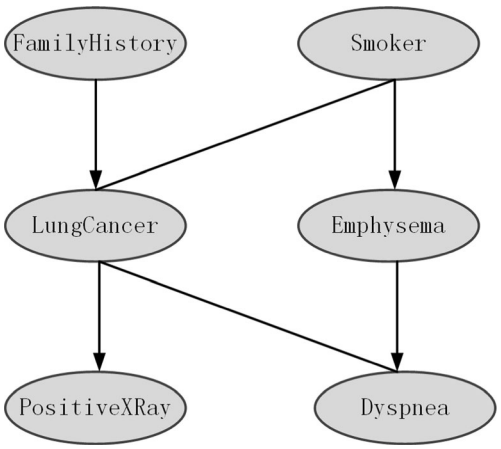


Fig. 6 The structure of a Bayesian network

Table 1 conditional probability attribute

	FH,S	FH,~S	~FH,S	~FH,~S
LC	0.7	0.4	0.5	0.2
~LC	0.3	0.6	0.5	0.8

and determine which variables are needed to describe the field, to understand the exact meaning of each variable.

2) Assume that there is a dependency between any two variables, by using the connection edge to represent the relationship between the variables, and then form a fully connected graph.

3) Based on mutual information measure and conditional independence test (CI), with a priori information and expert knowledge, a minimal undirected graph is obtained.

The mutual information is a random variable that contains a random variable information, it indicates that a random variable is reduced by the information of another variable. The mutual information is defined as follows:

Two discrete random variables$X$ and$Y$, it has joint probability function$P(x, y)$, marginal probability function$P(x)$and$P(y)$, its mutual information $I(X, Y)$was defined as:

$$I(X,Y)=\sum_{x\in Q_x}\sum_{x\in Q_y}P(x,y)\frac{P(x,y)}{P(x)P(y)} \tag{13}$$

4 Case analysis

4.1 Experimental results of cases

We use a social survey study to illustrate the computational methods of Bayesian Networks. Through the investigation of the students in a certain area, the following factors are found which has an effect on the students' schooling:

Gender ( ): male, female.

Intelligence Quotient ( ): low, lower middle, upper middle, high.

Family economy ( ): low, lower middle, upper middle, high.

Family encouragement ( ): low, high.

Whether intend to go to college ( ): yes, no.

Table 2 is a statistical result of 10,318 students. The first row of the Table 1 indicates that = male, = low, = low, = low, = yes and its student number is 4. The second part of the data table indicates that = male, = low, = low, = low, = no and its students number is 349, and so forth, in the lower part of the Table 2 (i.e., 1 to 8 line). The value of is female.

On the basis of Bayesian network, data mining can find out the cause and effect relationship between these variables, and the specific calculation process is as follows:

1) Select the appropriate network structure according to the prior knowledge.

For data samples with n variables, the network structure may be composed of $n!$ species, it is impossible to calculate the structure of each network. By using the existing expert knowledge, we can eliminate a large number of unreasonable combination. For instance, in this case, there is no relationship between the gender of the students and the economic situation of the family, so in the Bayes network, and there has no connection between X_1 andX_3. In the following calculations, we only choose two kinds of network structure: S_1 andS_2, which was shown in Fig. 7 and Fig. 8. The only difference is that the IQ of the students is different from the family economy.

2)The calculation of Dirichlet distribution coefficient a_{ij};$p(D|S^h)$ Was only determined by distribution coefficienta_{ij}, now it is necessary to estimate it firstly. Under the assumption that the network structure is known, the prediction formula for the case$l$ $C_l(l=2,3,\cdots,m)$is:

$$p\left(C_l|D,S^h\right)=\prod_{k=1}^{n}\frac{a_{ijk}+N_{ijk}}{a_{ij}+N_{ij}} \tag{14}$$

Whereby,$a_{ij}=\sum_{k=1}^{r_1}a_{ijk}, N_{ij}=\sum_{k=1}^{r_1}N_{ijk}$ and the statistical dataN_{ijk} is known, so the prediction probability $p(C_l|D,S^h)$ can be get easily. For instance, one case $C_l(X_1$= male, X_2= low, X_3= low, X_4= low, X_5= no), its probability is 349/10318 = 0.03382. Therefore, we can list m-1 equation, by using the minimum variance, we can estimate the value ofa_{ijk}.

(3) NetworkS^hstructure selection.

In the practical computation of Bayesian networks, the main difficulty is the estimation ofa_{ijk}. Because of the estimation of the model is nonlinear, it is very difficult to calculate. In literature [24], G. Coper and E. Herskovits provide that by using $a_{ijk}=1$to estimate, and it has little influence on the

Table 2 student survey results

4	349	13	64	9	207	33	72	12	126	38	54	10	67	49	43
2	232	27	84	7	201	64	95	12	115	93	92	17	79	119	59
8	166	47	91	6	120	74	110	17	92	148	100	6	42	198	73
4	48	49	57	5	47	123	90	9	41	224	65	8	17	414	54
5	454	39	44	5	312	14	47	8	216	20	35	13	96	28	24
11	285	29	61	19	236	47	88	12	164	62	85	15	113	72	50
7	163	36	72	13	193	75	90	12	174	91	100	20	81	142	77
6	50	36	58	5	70	110	76	12	48	123	81	13	49	360	98

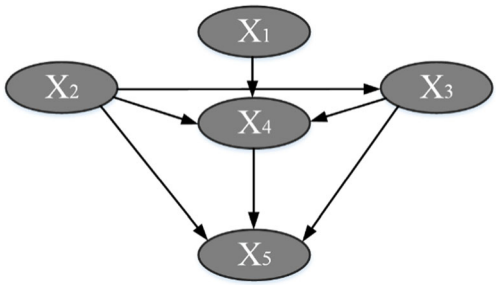


Fig. 7 S_1network structure

calculation results of the network. In our cases, we use a_{ijk} = 1to calculate the above two networks respectively.

$$p\left(S_1^h|D\right) = 1.0 \tag{15}$$

$$p\left(S_2^h|D\right) = 1.2 \times 10^{-12} \tag{16}$$

Thus, the network structure S1 can reflect the causality between the variables. At the same time, we have also noticed that the structural difference between S1 and S2 is not big, but the results of the calculations are quite different, it shows that the Bayesian network has a better sensitivity.

4.2 Experimental result analysis

As for the previous cases, we need to use the decision tree method to analyze the data, and the results we got are compared with the Bayesian network method. Figure 9 shows the Bayesian network method and decision tree method applied to the case of learning curve.

Judging from the learning curve of the decision tree method, the ratio of the test set is increasing at a rate of about 5000 people. But when the number is more than 5000 people, the correct proportion decreases with the increase of the number of the sample. It shows that the algorithm is unable to appear in the case of large amount of data or something complex. But with the increase of the number of the correct proportion, the Bayesian network method learning curve is increasing, when the number arrive to 7000 people, the proportion of the value began to tend to 1, as the same time, the performance is getting better and better. When the data is large, the Bayesian network

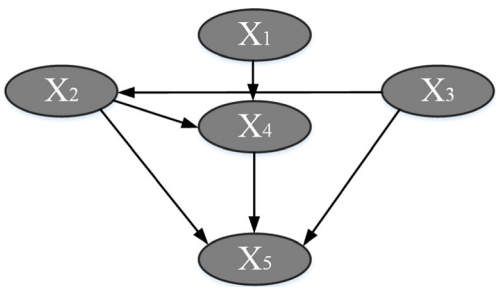


Fig. 8 S_2network structure

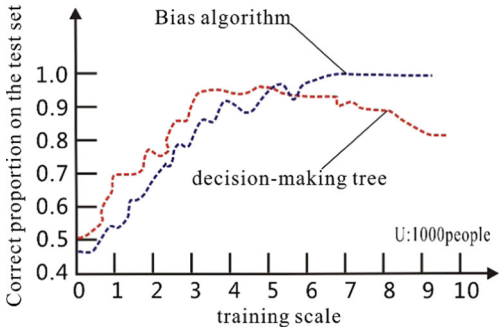


Fig. 9 Comparison of learning curves based on different methods

method is better than the decision tree method. Bayesian network has a causal and probabilistic semantics, which can be organically combined with prior knowledge and sample data, what's more, it can combine subjective and objective organic ground. Therefore, it is more comprehensive and objective to reflect the inherent connection and the essence of the data object, so it is convenient to realize the purpose of data mining. After repeated verification, the conclusion has universal applicability.

5 Conclusion

Bayes network is a graphical model used to represent the continuous probability distribution of a set of variables, it provides a natural way of representing causal information, which used to discover the potential relationship between the data. Bayes network learning, that is, to find out the one that can most truly reflect the existing database of the data in the dependent relationship between the variables of the Bayes network model. Through the research of this paper, we also see that the use of Bayesian networks is not only capable for handling large amounts of complex data in real applications, but also it can use its reasoning and self-learning ability, excavating causal relationship link from the database, the multilayer, and multi – point relationship. It reflects the universal connection between the objective world object, which is not available in the traditional data mining methods. Compared with the traditional method of data mining, it has many advantages. Some problems need to be further studied if Bayes network was applied to data mining, for example, the determination of the prior density is still a difficult problem. Bayes network requires a variety of assumptions for the premise, it has no existing rules, which brings difficulties to the practical application. Nonetheless, the Bayesian network will become a powerful tool in the near future.

Acknowledgements This work is supported by new technology development projects of Jilin Provincial Science & Technology Department, No: 20130305020GX.

References

1. Thangaraju Mr P., and Mehala R. (2015) Novel Classification based approaches over Cancer Diseases. system 4.3
2. Bijalwan V, Kumar V, Kumari P et al (2014) KNN based machine learning approach for text and document mining. Int J Database Theory Appl 7(1):61–70
3. Yukselturk E, Ozekes S, Türel YK (2014) Predicting dropout student: an application of data mining methods in an online education program. Eur J Open, Distance e-Learning 17(1):118–133
4. Bala S, Kumar K. (2014) A literature review on kidney disease prediction using data mining classification technique[J]
5. He W, Yan G, Da Xu L (2014) Developing vehicular data cloud services in the IoT environment. IEEE Trans Ind Inf 10(2):1587–1595
6. Peña-Ayala A (2014) Educational data mining: a survey and a data mining-based analysis of recent works. Expert Syst Appl 41(4): 1432–1462
7. Abdelhamid N, Ayesh A, Thabtah F (2014) Phishing detection based associative classification data mining. Expert Syst Appl 41(13):5948–5959
8. Chen F, Deng P, Wan J et al (2015) Data mining for the internet of things: literature review and challenges. Int J Distrib Sens Netw 501:431047
9. Dhakar M, Tiwari A (2014) A novel data mining based hybrid intrusion detection framework. J Inf Comput Sci 9(1):037–048
10. Zhang ML, Zhou ZH (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
11. Chaurasia V, Pal S (2014) Data mining approach to detect heart diseases. Int J Adv Comput Sci Info Technol (IJACSIT) 2:56–66
12. Jelinek HF, Yatsko A, Stranieri A et al (2014) Novel data mining techniques for incomplete clinical data in diabetes management. British JAppl Sci Technol 4(33):4591–4460
13. Bijalwan V, Kumari P, Pascual J, et al. (2014) Machine learning approach for text and document mining[J]. arXiv preprint arXiv: 1406.1580
14. Xing W, Guo R, Petakovic E et al (2015) Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory. Comput Hum Behav 47:168–181
15. Okazaki S, Díaz-Martín AM, Rozano M et al (2015) Using twitter to engage with customers: a data mining approach. Internet Res 25(3):1066–2243
16. Bounhas M, Hamed MG, Prade H et al (2014) Naive possibilistic classifiers for imprecise or uncertain numerical data. Fuzzy Sets Syst 239:137–156
17. Charninda T, Dayaratne TT, Amarasinghe HKN, et al. (2014) Content based hybrid sms spam filtering system[J]
18. Shukla DP, Patel SB, Sen AK (2014) A literature review in health informatics using data mining techniques. Int J Softw Hardw Res Eng 2(2):123–129
19. Olsson A, Nordlöf D. (2015) Early screening diagnostic aid for heart disease using data mining: An evaluation using patient data that can be obtained without medical equipment[J]
20. Zhou X, Lim JS, Kwon IK, et al. (2014) EM algorithm with GMM and Naive Bayesian to Implement Missing Values[J]. Proceedings of April 17th, 15–19
21. Dey M, Rautaray SS. (2014) Disease Predication of Cardio-Vascular Diseases, Diabetes and Malignancy in Lungs Based on Data Mining Classification Techniques[J]
22. Jayakameswaraiah M, Ramakrishna S. (2014) A study on prediction performance of some data mining algorithms[J]. International Journal, 2 (10)
23. Xiao-feng Z, Shu W (2014) Data mining method of road transportation management information based on rough set and association rule. J South China Univ Technol (Natural Science Edition) 2:021
24. Zhao Y, Niu Z, Peng X (2014) Research on data Mining Technologies for Complicated Attributes Relationship in digital library collections. Appl Math 8(3):1173–1178

Mingyang Liu is a native person in Jilin Province of China. He received his Master degree in Jilin University in 2010. He graduated from the College of Software Engineering Department. His subject of Master was related to data mining. He got his Ph.D degree in Jilin University in 2013. He have been focusing on big data processing till now. At present, he is working as a teacher in College of Instrumentation and Electrical Engineering of Jilin University. He is still focusing on the data mining domain these years.

Ming Qu is working as a teacher in the College of Computer Science and Technology of Jilin University. Also he is currently working as a post-doctoral in the College of Electronic Science and Engineering of Jilin University. His research interest is on Computer-Supported Cooperative Work. Being focus on the problems of information security for Integrated Circuit, he is also working on modeling and simulating IC cryptogram system by using big data analyze and mining.

Bin Zhao , is currently working as the executive vice President of academy of arts, director of the foundation education of Jilin Animation Institute. Also he was the vice general manager of Jilin Yushuo film Co. Ltd. He graduated from Luxun academy of fine arts department of sculpture. Since 1980s, he has worked in animation design and production industry, also focusing on the data mining in cartoon database.