Applied Computer Letters (ACL)

http://doi.org/10.26480/acl.02.2017.07.09

RESEARCH ON APPLICATION OF DATA MINING BASED ON BAYESIAN NETWORK

Guoliang Zhao[1*], Guolin Zhao[2]

ARTICLE DETAILS

Article History:

Received 12 July 2017
Accepted 12 August 2017
Available online 8 October 2017

ABSTRACT

This paper is committed to applied research in Bayesian networks in data mining, for there is not a complete Bayesian networks in data mining algorithms, propose a heuristic in data mining to construct Bayesian network using sample data algorithm thought, this algorithm solves the problem by using sample data design the Bayesian network in data mining. Finally, the experimental analysis, established the model of college students study section, the experimental results show that the algorithm proposed in this paper, the design of simple practical method, the application of effective, compared with other algorithms and the characteristics of high precision, but also show the algorithm mining advantages in the aspects of data, to practice management, analysis, prediction and decision etc.

KEYWORDS

Bias network, data mining, conditional independence, mutual information.

1. INTRODUCTION

Based on a study, data mining (DM) is a new cross subject, based on a large amount of data on a comprehensive and profound understanding of the case, through calculation, induction and reasoning part, extracted from the common, general and the essence of the phenomenon or characteristics [1]. According to research, the common methods include genetic algorithm, decision tree, Bayesian network, rough set, neural network [2]. The Bayesian network is used to represent the variables set the probability graph model, it provides a method to represent causal information [3]. Bias network to considering the prior information and the sample data, make full use of expert knowledge and experience, qualitative analysis and quantitative analysis. The subjective and objective organically, can avoid over fitting of the data, but also avoid the subjective factors that may result in bias [4]. A research shows that good prediction effect, in the face of massive data shows its unique advantage [5]. This paper presents a Bayesian network in data mining algorithm. For there is not a complete Bayesian networks in data mining algorithms, to put forward a heuristic Bayesian network using sample data algorithm in data mining, this algorithm solves the problem with sample data, the design of Bayesian network in data mining.

2. ALGORITHM OF DATA MINING BASED ON BAYESIAN NETWORKS

Because Bayesian network to considering the prior information and the sample data, make full use of expert knowledge and experience, the characteristics of subjective and objective are organically and many other than other method, for there is not a complete construct Bayesian network in data mining algorithms, to put forward a heuristic Bayesian network using sample data algorithm in data mining:

(1) According to the goals and tasks of data mining, data analysis and variable selection, determine the need which variables describing the field, and the exact meaning of each variable [6].

(2) There are dependencies between assume that any two variables, representing the association between variables with connecting edges form a connected graph, there are n (n1) > 2 edges.

(3) The use of mutual information measure and conditional independence test edge deleting algorithm based on prior information, combined with expert knowledge and learning a minimum undirected graph.

Mutual information is the information measurement of a random variable contains another random variable, it shows that a random variable and reduce the uncertainty due to get another variable information [7]. The definition of mutual information is as follows:

The discrete random variables X and Y have a joint probability function p (x} Y) and marginal probability function p (x), and P(x), the mutual information I (X; Y) is defined as:

$$I(X;Y) := I(X;Y) = \sum_{x}\sum_{y} p(x,y) \frac{p(x,y)}{p(x)\,p(y)} \qquad s$$

The Z conditions, random variables X and Y conditional mutual information is defined as the calculated between any two nodes Vi and Vj mutual information I (Vi, Vj) are arranged in descending order, constitute a node pair (Vi, Vj) set K0 application edge deleting algorithm, set a smaller threshold values: K1 > 0, if the mutual information I (Vi, Vj) < k1, then delete the connection between Vj and Vi, the retention of the connections between the nodes, or. Then the conditional independence of the node (Ci) test, further delete redundant connection, until a minimum undirected graph.

(4) Determine the connection side direction. Use of prior information and expert knowledge, can eliminate a large number of different. Reasonable structure, and at the same time by using the maximum a posteriori probability (MAP) and minimum description length (MDL) criterion, adjustment [8]. Connecting edges between nodes, at the same time determined between graph nodes connected edge direction. A variable X is initialized to a value set {X1, X2, ..., Xn}, if a value corresponding to the concept of X Yu Xinyuan.

$$S_0=\arg\max\left[P(D/S)P(S)\right]=AEGMIN\left[L(D/S)+L(S)\right] \quad (2)$$

The rate of $P(x)=P(X=x)$, the value of the amount of information of a x is defined as: $I(X=x)=I(x)=-\log 2p(x)$. The best of the information describing the source language should be the minimum description length variable X, hence the description length L(x) and I(x) is equal to the amount of information. Based on this relationship, the selection of the optimal structure of Sn maximum a posteriori probability (MAP) and minimum description length (MDL) of the quasi

$$S_0=\arg\max\left[P(D/S)P(S)\right]=AEGMIN\left[L(D/S)+L(S)\right] \quad (3)$$

(5) The relevant expert knowledge and rule, using sample data, the DAG is revised and advantages, the adjustment of the structure of the Bayesian network, network, through the calculation and analysis, and the sample data to study the optimum matching network structure.

(6) Determine the function table of the probability distribution of nodes by using conditions of parameter learning algorithm (CPT). When the data set is complete, the probability parameter learning in general there are two ways: one is the classical sample statistics, two is Bias statistics; when the data set is incomplete, the general use of the method of approximate calculation, approximate maximum likelihood function, and the values of the estimated parameters as the point. Through each variable network structure is determined by experts in the conditional probability distribution function, dependent on the relationship between quantitative variables, using prior information and expert knowledge, obtain the optimal Bayesian network model.

3. EXPERIMENTAL RESULTS AND ANALYSIS

Nowadays, the employment situation is very grim, college students after graduation from the university are faced with is to graduate or move towards. The two choices of social work, according to a survey of the study section below a certain area of university graduates, the use of several factors decisive, statistics of 10000 college students, to construct a Bayesian network, verify the feasibility of the algorithm and effectiveness, and the advantages of Bayesian network in data mining. To construct a Bayesian network to identify the causal relationships among these variables for data mining, the specific process of constructing network are as follows: to define the variables for gender (A), academic achievement (B) (C), family economy, the employment situation (D), is going to graduate students (E); according to the existing expert knowledge we select only the most likely a and B two kinds of network structure. The difference between them is the causal relationship between students' employment situation and family economy are different, the causal relationship between academic achievement and family economic differences. As shown in Figure 1:

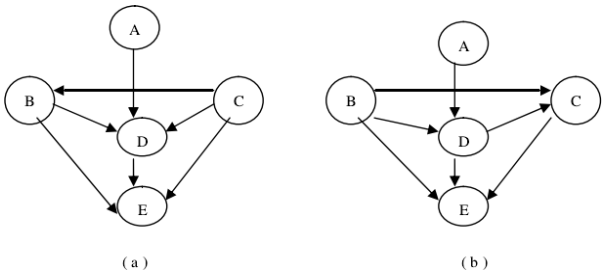

Figure 1: The two most likely network structures of A and B.

For this example, probability parameter, calculating the variables using sample statistics such as the classic Figure 1 (a), the calculation of P (B, C), to find P (B, C) and P (C), by P (B, C) = p (B, C) / P (C) using table. The data, obtained the B conditional probability distribution function, other parameters can be obtained by the same method. After calculation and Analysis, the conditional probability distribution function adjustment of network structure of Bayesian network and the variables, finally draw the map 1(b) for the optimal Bayesian network structure.

For the case, use the decision tree method to analyze the data, compared the results obtained with Bayesian network method, the learning curve for the Bayesian network method and decisions tree method is applied to the case as shown in Figure 2.

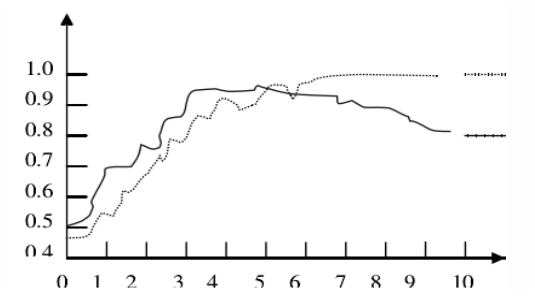

Figure 2: Comparison of the case based on the learning curve of the two different methods

The learning curve for about before the number of 5000 people from the decision tree method, numerical value on the test set ratio. The total is increased, but more than 5000 people in proper proportion with the increase in the number of reduced, show that the decision tree. In a large amount of data and the data of complex cases, the algorithm is insufficient, the performance is getting worse. While Bias network approach to learning curve with the increase in the number of correct, proportion of the increase, when the arrival of 7000. The proportion of people around, numerical tends to 1, the performance is getting better and better. Get in the larger amount of data, the Bayesian. Si network method is better than the decision tree method. Bias network is causal and probabilistic semantics, which can be.

organically the knot Combining prior knowledge and sample data, subjective and objective are organically, clear expression, it can reflect and the essence of data objects inside, to facilitate the realization of data mining.

4. CONCLUSIONS

Bias network is a powerful tool for uncertainty reasoning and knowledge representation, when used in combination with statistical method, showing many of the advantages of data processing. This paper presents a heuristic to construct the Bias network using sample data algorithm in data mining, and applies it to the actual case, and will be the result of the decision tree method and the traditional credit scoring method were compared, obtained the Bias network should be used for data mining, which can fully exploit the data implied information and the inherent nature, has a good prediction ability etc.

organically the knot Combining prior knowledge and sample data, subjective and objective are organically, clear expression, it can reflect and the essence of data objects inside, to facilitate the realization of data mining.

REFERENCES

[1] Yang, L., Sun, H., Zhu, H.C. 2007. Multi objective optimization algorithm based on Bayesian network, Journal of North China Electric Power University, 34, 1, 128-131.

[2] Wang, S.C., Zhang, M., Chen, N.J. 2007. Learning Bayesian network structure based on causal semantic orientation, Computer Engineering and Application, 43, (8): 29-31.

[3] Tian, F.Z., Huang, L. 2005. Bayesian network containing hidden variables incremental learning method, Acta Electronica Sinica, 33, (11): 1925-1928

[4] Ji, J.Z., Yan, J., Liu, C.N. 2006. Improved learning algorithm of Bayesian network structure based on I-B&B-MDL, Journal of Beijing University of Technology, 32, (5) 437-441.

[5] Ji, J.Z., Liu, C.N., Jiang, C. 2002. Application of Bayesian networks and probabilistic reasoning in intelligent teaching, Journal of Beijing University of Technology, 2002, 353-357.

[6] Zhou, Z.B., Zhong, B., Dong, D.D., Zhou, J.L. 2006. Application of Bayesian networks in reliability analysis, System Engineering Theory and Practice, 2006, 6, 95-100.

[7] Ji, C.C., Sha, Z.Q., Liu, C.N. 2005. Application research of Bayesian network model in the recommendation system, Computer Engineering, 31, (13): 32-34.

[8] 2006. The field Zide, Study on the application of Bayesian networks in adaptive hypermedia system, Information Science, 24, 7, 1049-1