International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.267-274
http://dx.doi.org/10.14257/ijdta.2016.9.5.27

Data Mining Technology Based on Bayesian Network Structure Applied in Learning

Chunhua Wang and Dong Han

College of information engineering, Huanghuai University,Henan,China
Corresponding E-mail:Flyequn@163.com

Abstract

Originated from Bayesian statistics, Bayesian network，with such characteristics as its unique expression form of uncertainty knowledge, rich probabilistic expression abilities, and the incremental learning method for comprehensive priori knowledge，indicates the probability distributions and causal relations of objects, becoming one of the most striking focus among numerous current data mining methods. Taking a brief introduction of data mining for the point-cut of the study, and combining an explanation for the data mining process and an analysis of Bayesian Network, the paper investigates the implementation of Bayesian network applied in in learning.

Keywords*: Bayesian network; Learning method; Data mining*

1. Introduction

Data mining is a process as well as a knowledge discovery in database, namely, it is from the large, incomplete, noisy, fuzzy and random data to extract implicit information that people do not know in advance, but it is potentially useful information and it is a non-trivial process for knowledge. Data mining is originated from many disciplines, including database, artificial intelligence, statistics, machine learning, *etc.*. Among them, the most important three fields are database, machine learning and statistics. These different historical influences made the different scholars hold different views on the function of data mining.

Introduction of Data Mining

With the development of global informatization, automatic data acquisition tools and mature database technologies have led to massive data stored in databases. It is very important to extract reliable, novel, and effective knowledge from massive data which can also be understood by people, hence data mining has caused great concerns to information industry. Its extensive application fields involve agriculture, learning diagnostics, business management, product control, market analysis, engineering design, scientific research, and so on.

Data mining is related to many disciplines and methods, therefore, there are data a variety of classification methods for data mining. According to the task of data mining, it can be divided into classification or warning model discovery, data summarization, clustering, association rules discovery, sequence pattern discovery and dependency relation or the dependent model discovery, exception discovery and trend discovery, *etc.*; according to the object of data mining, it can be including the relational database, object-oriented database, spatial database, temporal database, text database, multimedia database heterogeneous database, heritage database and Web, *etc*.; according to the method, it can be divided into the machine learning method, statistical method, neural network method and database method. While machine learning methods can be divided into inductive learning methods (decision trees, induction of rules, *etc*.) based on the case

ISSN: 2005-4270 IJDTA

study, active learning, genetic algorithms, *etc.* Statistical analysis methods can be divided into regression (multivariate regression and autoregressive regression, *etc*.), discriminant analysis (Bayesian discriminating, Fischer discriminant, nonparametric discriminant, *etc*.), cluster analysis (hierarchical clustering, clustering segmentation, *etc.*), exploratory analysis (principal component analysis, correlation analysis, *etc.*) and so on. The artificial neural network method can be divided into feedforward neural networks (BP algorithm), self-organizing neural network (self-organizing feature map, competitive learning, *etc*.) The database method mainly includes the multidimensional data analysis, attribute-oriented induction method, *etc.*

Data Mining Process

Data mining is a process of mining interesting knowledge from among mass data stored in databases, data warehouses or other information bases. It is a definition by JiaweiHan in his work, *Data Structures—Concepts and Technologies*. Data Mining also refers to the process of discovering knowledge from mass data, as shown in Figure 1, which represents well the process of knowledge discovery.

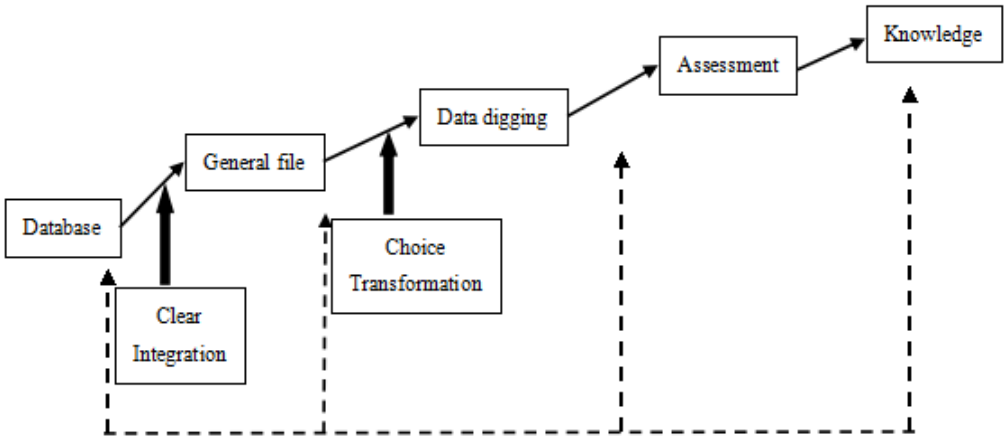

Figure 1. Flow Chart of Data Mining

The Determination of Business Logics

It is an important step for Data Mining to clearly define its business problems and determine the purposes of data mining. It has a blindness and will not be successful to mine data simply for the purpose of data mining itself.

Data Preparation

- The relearning of data, eliminating inconsistent and noisy data, and combining together the data from different data sources.
- Choice of Data, searching for all internal and external data information relating to business objects, among which the data suitable for the application of data mining are chosen.
- Data Transformation, aiming at certain methods for mining algorithm, and transforming data into forms suitable for mining.

Data Mining

- Using intelligent methods, we mine the acquired and transformed data. In addition to choosing the right mining algorithm, all the rest of the work can be completed automatically.
- Knowledge Assessment. Interpreting and evaluating the results. The adopted analytical methods are generally determined by data mining operations, and

visualization techniques are often used.

- Knowledge Representation. The knowledge obtained through an analysis will be provided to the users, or integrated into the organizational structures of business information systems.

Classification and Forecasting

Based on known training sets, Classification is used to find out models or functions which describe and distinguish the data classes or concepts, and to accurately classify each group and entities according to the classified information, so as to forecast object classes with unknown signs by using models. While in Forecasting, the forecasted values are numerical data.

Cluster Analysis

Cluster analysis aims at a collection of data objects, aggregates entities with the same characteristics to become one category, enables data objects in the same category to share as many similarities as possible, and uses certain rules to describe the common properties of the category, whereas there are large differences among objects in different categories. Clustering, in essence，is an unsupervised learning method, the purpose of which is to find out similarities and differences in data sets and to aggregate the data objects sharing common characteristics into the same category, the characteristics of each cluster can usually be analyzed and explained.

Outlier Analysis

Outliers are those data which are inconsistent with general acts or models of the most of the data in data sources. Much of this data are considered noises or abnormal and are discarded. However, these data are more interesting in such fields as analyzing customer behaviors, credit fraud screening and quality control of data, network security management and fault detection than those data appearing normally.

Bayesian Network

Bayesian network, also known as probabilistic causal network, web of trust, knowledge graph and so on, is a directed acyclic graph. A Bayesian network is composed of two parts: (1) a directed acyclic graph G with k nodes (as shown in Figure 2). The nodes in the graph represent random variables, and the directed edges between nodes represent interrelated relationships among the nodes. Node variables can be the abstract of any issues, such as test values, phenomena observations, questions and comments, *etc.*.. Usually directed edges are considered to express a kind of causal relationship, so Bayesian network is sometimes called causal network. It is important that directed graphs contain the conditional independence assumption, Bayesian network stipulates that each node Vi condition in the graph is independent from any nod subset constituted by non Vi descendant nods given by parent nods of Vi, Fi *i.e.*, if A(Vi) is used to represent any node subset constituted by non Vi descendant nods, II (Vi) is used to represent direct parent nodes of Vi, then p(Vi A(Vi),∏(V i))=p(Vi ∏(Vi)).

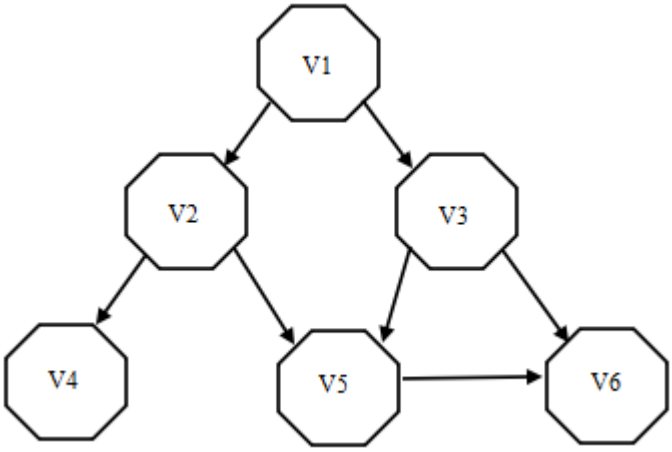

Figure 2. An Exemplified Example of Structure of Bayesian Network

P is a Conditional Probabilities Table (CPT) associated with every node. The Conditional Probability Table can be described by p(Vi ∏(V i)), which expresses correlations between nodes and their parent nodes---the conditional probability. The node probability without any parent nodes is its priori probability. A joint probability can be expressed according to a conditional probability chain, whose general form is

$$p(V1,V\ 2,...,V\ k)= \prod_{i=1}^{n} P(Vi\ Vi\text{-}1...V1)$$

The Network constituted by a graph G and a probability table p is called Bayesian Network, which represents causal relationships among random variables through the form of directed digraphs, and quantifies the relationships through conditional probabilities, and which can contain joint probability distributions of random variable sets and is an information representation framework combining causal knowledge and probabilistic knowledge.

Model Construction

Constructing the model is the core of data mining. Once the data cleaning and the transformation of variables are completed, the model construction is began. Before the construction of model, it must understand the target if the data mining and the types of data mining. In this stage, it needs for the cooperation with the relevant analysts who had the relevant knowledge. After understanding the task of data mining, selecting the appropriate algorithm becomes relatively easy. Each data mining task should be correspond to the appropriated algorithm. In most cases, we don't know what kind of algorithm is the most suitable before constructing the model. The accuracy of data depends on the properties of the algorithm. The correct way is to use different algorithms to construct multiple models, and then use the tools to evaluate the accuracy of these models. Even with the same algorithm, it should also set different parameters to build multiple models, so that it is conducive to adjust the accuracy of the model.

Clustering Analysis Model

Clustering analysis model must meet the requirements: (1) each model must contain a unique key column, it can be a numeric or text column, which is used to identify each record uniquely. (2) each model must contain at least one input column, the input column should contain the value which is used to generate the classification. The input column can be set more at random, while adding additional input column will increase the time for the model setting. (3) as far as this model is concerned, the predictable column is not required necessarily, but the predictable column can be added. The value of the

predictable column can be considered as the input of a clustering model, which also can be specified only for the purpose of prediction, Shown in Figure 3.

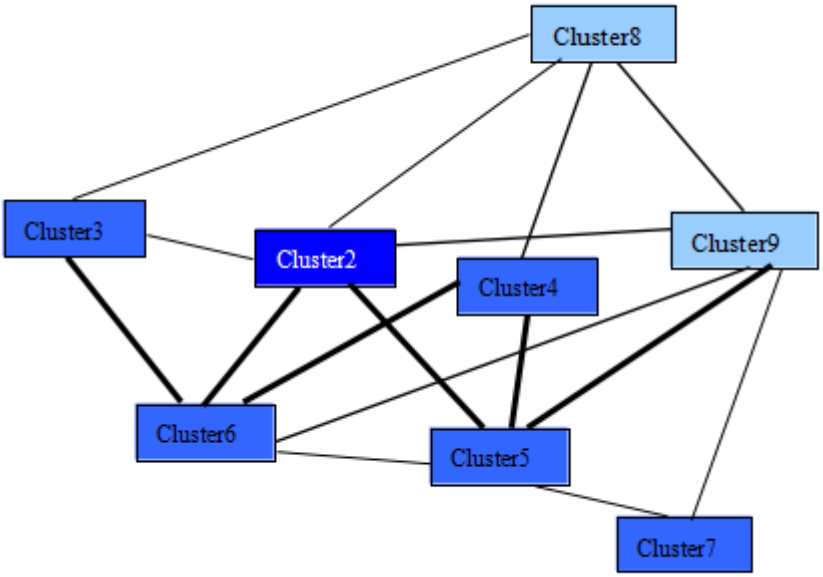

Figure 3. Clustering Process

Verifying Mining Model

Verification is the process of assessing the implementation of mining model over the real data. Before setting the mining model in the production environment, it must verify it through the understanding of its quality and characteristics. It can use a variety of methods to assess the quality and feature of data mining model. First of all, it can use various statistical validity information to determine whether there are problems in data or models. Secondly, the data can be divided into the set of training and set of test so as to test the accuracy of prediction. Finally, it can ask the business experts to check the result of the data mining model to determine whether the pattern of findings in the target business scheme is meaningful. All of these methods in data mining methods are very useful in creating, testing and optimizing models to solve specific problems which can be used repeatedly.

While Apriori algorithm is used for the Boolean data, so the historical evaluation data need to be converted. The reflection of food category is S, wherein x represents the number of the category, y represents the number of sub-class. The reflection of time and quarter is J_1-J_4. The reflection of month is M_1-M_{12}, the enterprise risk level is set Q_L, Q_M, Q_S respectively as high risk, medium risk, low risk, for example, the official risk is set as G_L, G_M, G_S, the trade risk is set as T_M, T_L, T_S, the result of the risk assessment is set as P_L, P_M, P_S, which is encoded as follows in Table 1:

Table 1. Data after Being Encoded

$S_{01},S_{015},Q_S,G_S,T_S,J_3,M_9,P_S$
$S_{02},S_{021},Q_M,G_M,T_L,J_2,M_4,P_M$
$S_{02},S_{022},Q_M,G_S,T_L,J_2,M_4,P_M$
$S_{01},S_{15},Q_L,G_M,T_S,J_3,M_8,P_M$
$S_{02},S_{012},Q_M,G_M,T_M,J_2,M_5,P_M$
$S_{02},S_{021},Q_M,G_L,T_M,J_2,M_6,P_M$
$S_{02},S_{022},Q_S,G_M,T_M,J_1,M_1,P_S$
$S_{01},S_{15},Q_L,G_L,T_M,J_2,M_5,P_L$

The Implementation of Learning Based on Bayesian Network

A Bayesian network constructed according to users' priori knowledge is called a priori Bayesian network, and a Bayesian network obtained by the combination of priori Bayesian networks and data is called a posteriori Bayesian networks, the process of obtaining posteriori Bayesian networks from priori Bayesian networks is known as Bayesian network learning. Bayesian network can keep learning, the posteriori Bayesian network obtained by the last learning can become the prior Bayesian network for next learning. Before each learning, users can make adjustments to the prior Bayes networks, enabling new Bayesian networks to be able to better reflect the knowledge contained in the data, as shown in Figure 4.

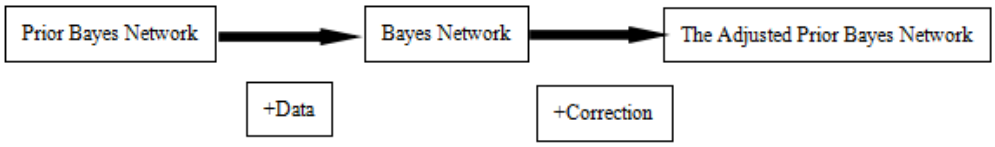

Figure 4.Continuous Learning Graph of a Bayesian Network

The learning based on a Bayesian network includes two contents: parameter learning and structure learning, meanwhile, according to different natures of the sample data, each part includes two aspects: complete instance data and incomplete instance data. Parameter learning methods are mainly the learning based on classical statistical learning and the learning based on Bayesian statistics - conditional probability table (CPT). Structure learning methods are mainly based on the Bayesian statistical measurement methods and based on coding theory measurement methods. The learning based on structures is presented below.

In a Bayesian network, firstly a random variable Sh is defined, representing that the database D is a random sample assumptions from the network structure S and is given a priori probability distribution $p(S^h)$ which indicates the uncertainty of the network structure, and then the posterior probability distribution $P(S^h\ D)$ is calculated. According to the Bayesian theorem, we have:

$P(S^h\ D)=P(S^h,D)/P(D)=P(S^h)P(D\ S^h)/P(D)$, where P(D) is a normalization constant which is irrelevant to structure learning, and P(D Sh) is a structure likelihood. Then the posterior distribution of the network structure is determined, while the only need is to calculate the structure likelihood of the data for each possible structure.

On the premises of multinomial distribution without constraints, independent parameters, and the adoption of Dirichlet priori and complete data, the structure likelihood of the data is exactly equal to the product of the structure likelihood of every (i,j) pair.

Research on Key Technology

Information Standardization

The related technology, standards, protocols and interfaces used by the regional information platform should follow relevant regulations of international, national and industry. The data structure and the design of the application software should implement the relevant standards of the industry [3]. First the regional learning information platform should refer to regulations of the basic framework and data standard of the learning records (for trial implementation) and regional learning information platform construction guide based the learning records (for trial implementation) issued by the Ministry of learning.

Data Acquisition and Access Technology

Due to the large student amount and the large volume of business the amount of data to interact in the student is also large. In order to ensure the normal operation of the student business system and the information sharing of the regional learning information platform the pre-machine can be used in the student to complete the interaction with the regional learning information platform. When the student registers in the student the identity card number is uploaded to the pre-machine and the pre-machine downloads the electronic learning information of the student to the teacher for the use of learning. In addition the pre-machine automatic extracts the learning information from the student systems and uploads to the regional learning information platform.

The community learning service stations are all small in scale, so the data acquisition and accessing can use directly way of uploading and accessing. If the system is not easy to directly upload, it can also use the mode of manual importing.

Network Transmission Technology

Regional learning information platform involves all the students and the community learning service station in the region. The locations of the various learning institutions are scattered and some may be relatively remote. Therefore the network transmission can use a variety of combination. For the large students that the locations are near and the conditions are ripe these can use dedicated line access. But the cost of the dedicated line access is higher and the flexibility is not good. For the learning institutions that the locations are remote and there are no dedicated lines those can use the access mode of VPN (Virtual Private Network). VPN has to access methods that are SSL (Secure Sockets Layer) VPN and IPsec (Internet Protocol Security) VPN according to the use of different protocols. The way of SSL VPN is more flexible and convenient and the cost is low. The way of IPsec VPN need to install equipment in the client and the realization is complicated.

Data Storage Technology

The regional learning information platform stores the electronic learning record information and the electronic learning record information. The data storage of the regional learning information platform can use those ways including centralized, distributed and mixed ways. The centralized storage is to build a unified data center to store all the data in the students and communities such as the electronic learning record information. Virtualization technology can be used in order to ensure the efficient management of the centralized data store and reduce the cost. The distributed storage is to store the data in each institution and save the indexed in the regional learning information platform such as the learning image information of the residents. The mixed storage is the combined way of the centralized way and the distributed way.

Information Security Technology

The electronic learning record information of the regional learning information platform refers to the privacy of the residents. Once the electronic learning record information is leaked that will seriously affects the social order. So the construction of the regional learning information platform should be strictly in accordance with requirements of the national security level protection system to ensure the authenticity, integrity, confidentiality, availability, reliability and controllability of the information. The regional learning information platform should have the ability of safety protection, the ability of discovering the hidden trouble and the ability of emergency response. Therefore it needs to deploy equipment's of safety isolation, intrusion prevention, malicious code defense and application layer firewall in front of the intranet hosts. The security of the services provided to external by the regional learning information platform need further strengthen. By employing the new method of the application layer defenses the attacks of the WEB such as the SQL injection and cross site scripting to ensure the anti-leakage of the information and the tamper proof of the portal site.

2. Conclusion

Therefore, the exported learning in the actual application should be combined with professional knowledge to analyze the rules produced by data mining, at the same time, these rules also provide some clues and inspiration for the professional research.

References

[1] E. Charniak, "Bayesian Networks without Tears", AI Magazine, vol. 12, **(1991)**, pp. 50-63.
[2] J. Pearl, "Graphical Models for Probabilistic and Causal Reasoning", The Computer Science and Engineering Hand-book. Kluwer Academic Publishers, **(1997)**, pp. 697-714.
[3] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data", Machine Learning, vol. 9, **(1992)**, pp. 309-347.
[4] W. Sewell and V. Shah, "Social class, parental encouragement and educational aspirations", American Journal of Sociology, vol. 73, **(1968)**, pp. 559-572.
[5] P. Spirtes, C. Glymour and R. Scheines, "Causation, Predication and Search", New York: Springer-Verlag, **(1993)**, pp. 25-29.
[6] H. Wang, W. Liu and S. Zhang, "Journal of learning Informatics", vol. 7, **(2010)**, pp. 14-17.
[7] J. Zhong, B. Liu and J. Chen, "Journal China Digital Medicine", vol. 9, **(2013)**, pp. 12-14.
[8] B. Song, H. Chen, C. Zheng, J. Saiyu, "Hospital Administration in Journal China PLA", vol. 9, **(2010)**, pp. 819-821.
[9] J. Li, H. Zhang and X. Li, "Journal China Digital Medicine", vol. 5, **(2014)**, pp. 102-104.
[10] Y. Tang, J. Liu, H. Gan, W. Chen, C. Feng and C. Bu, "Chinese Journal of learning Informatics and Management", vol. 2, **(2013)**, pp. 96-104,129.

Authors

Chunhua Wang, assisted professor, Research direction: Computer analysis. Data mining.

Dong Han, professor, Research direction: Computer analysis. Data mining.

