Advance Journal of Food Science and Technology 12(2): 70-73, 2016
DOI:10.19026/ajfst.12.2841
ISSN: 2042-4868; e-ISSN: 2042-4876
© 2016 Maxwell Scientific Publication Corp.
Submitted: August 24, 2015 Accepted: October 11, 2015 Published: September 15, 2016

Research Article
The Application of Data Mining Technology Based on Bayesian Network Structure in Food Science Learning

Zhen-Feng Jiang
School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China

Abstract: The paper investigates the implementation of Bayesian network in food science learning. 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. 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 prior knowledge, indicates the probability distributions and causal relations of objects, becoming one of the most striking focus among numerous current data mining methods.

Keywords: Bayesian network, data mining, food science learning

INTRODUCTION

Food science teaching is an interdisciplinary subject of preventive medicine and food science, closely relating to human health. In the teaching contents, it mainly introduces functional factors, development of a variety of functional food sciences and evaluation and detection of processing technologies for functional food sciences, among which the development of functional food sciences includes such contents as lowering blood pressure, lowering blood fat, improving diabetes, improving one’s look, antitumor, immune enhancement and improving osteoporosis. In order for the students to have a better grasp of this course and achieve the training goal of food science and engineering specialty and the training goal of food science nutrition and safety, we must renew our education concepts, continuously organize reasonable teaching contents, improve teaching methods and trains of thought, increase amount of information as much as possible and improve the teaching quality of relevant food science specialized courses. Therefore, we conduct a study on teaching methods for relevant food science specialized courses and introduces data mining techniques, thus enhancing students' learning validity (Pearl, 1997). 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, medical diagnostics, business management, product control, market analysis, engineering design, scientific research and so on.

MATERIALS AND METHODS

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 Charniak (1991) in his study, Data Structures-Concepts and Technologies. Data Mining also refers to the process of discovering knowledge from mass data, as shown in Fig. 1, which represents well the process of knowledge discovery.

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 pretreatment 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.

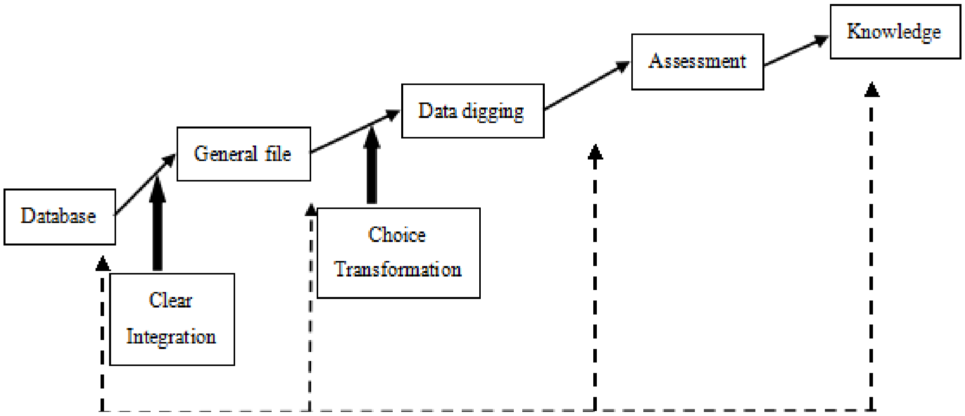

Fig. 1: Flow chart of data mining

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.

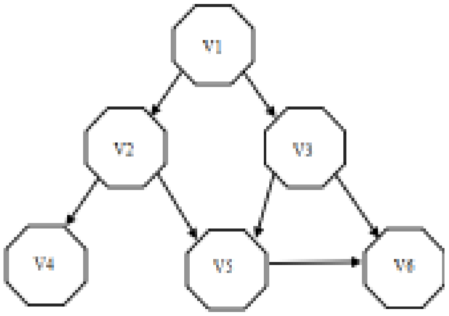

Fig. 2: An exemplified example of structure of Bayesian network

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 (Cooper and Herskovits, 1992). 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:

- A directed acyclic graph G with k nodes (Fig. 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

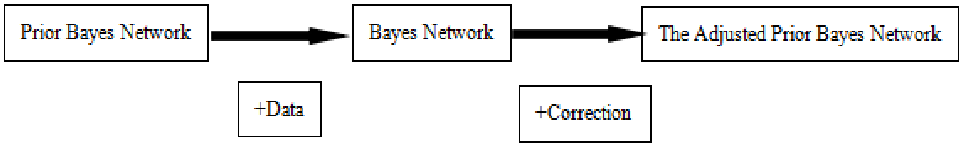

Fig. 3: Continuous learning graph of a Bayesian network

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, 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 $p(Vi\ A(Vi),\prod(V\ i)) = p(Vi\ \prod(Vi))$.

RESULTS AND DISCUSSION

P is a Conditional Probabilities Table (CPT) associated with every node. The Conditional Probability Table can be described by $p(Vi\ \prod(V\ i))$, 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.

The implementation of food science learning based on Bayesian network: A Bayesian network constructed according to users' prior 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 (Sewell and Shah, 1998). 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 (Fig. 3).

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 (Spirtes *et al*., 1993).

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.

CONCLUSION

The food science 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.

REFERENCES

Charniak, E., 1991. Bayesian networks without tears: Making Bayesian networks more accessible to the probabilistically unsophisticated. AI Mag., 12(4): 50-63.

Cooper, G.F. and E. Herskovits, 1992. A Bayesian method for the induction of probabilistic networks from data. Mach. Learn., 9(4): 309-347.

Pearl, J., 1997. Graphical Models for Probabilistic and Causal Reasoning. In: The Computer Science and Engineering Hand-Book. Kluwer Academic Publishers, NY, pp: 697-714.

Sewell, W.H. and V.P. Shah, 1998. Social class, parental encouragement, and educational aspirations. Am. J. Sociol., 73(5): 559-572.

Spirtes, P., C. Glymour and R. Scheines, 1993. Causation, Prediction, and Search. Springer-Verlag, New York, pp: 25-29.