1st World Congress of Administrative & Political Sciences (ADPOL-2012)

General Bayesian Network Approach to Health Informatics Prediction: Emphasis on Performance Comparison

Dahee Chung a, Kun Chang Lee b*, Seung Chang Seong a

a SKK Business School, Sungkyunkwan University, Seoul 110-745, Republic of Korea
b Professor at SKK Business School, WCU Professor at Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, Republic of Korea

Abstract

Health Informatics is emerging as a promising research area. As average life expectancy increases due to medical technology development, health issues remain most sensitive agenda in most of countries in the world. However, health technology requires more intelligent mechanisms by which users' requirement for more accurate prediction about their health problems can be fulfilled. Furthermore, such intelligent mechanisms must provide very flexible and robust procedures by which complicated but necessary decision support functions are allowed. In this sense, this paper proposes General Bayesian Network (GBN) to predict appropriate diets and restaurants that would benefit users' health. We compared the performance of GBN with other competing techniques such as NBN (naive Bayesian Network), TAN (Tree Augmented naive Bayesian Network), and decision tree. Experiments with real health dataset revealed that GBN results outperform other techniques with statistical validity.

2013 Published by Elsevier Ltd.
Selection and peer review under the responsibility of Prof. Dr. Andreea Iluzia Iacob.

Keywords: General Bayesian Network, Naïve Bayesian Network, Tree-Augmented NBN, Health Informatics

1. Introduction

Recently, in the era of aging society, better medical development and the improvement of overall standard of living lead to longer life expectancy. It means that healthier and longer lives become important as well as curing diseases. Together with these changes, general opinion is starting to make out that health is wealth. Hence, healthcare issues are clearly and highly important concerns. Particularly, a lot of disease such as obesity, diabetes and high blood pressure are closely related to daily eating habits, so the diet is considered much than before. Also, individuals' psychological state such as level of stress and anger are considered as important factor of health and well-being.

Since it is especially known that a lot of physical problem is affected by psychological state, it is getting more important to control psychological state.

Together with these changes, even when people go out for a meal, they usually want to find something to eat which is proper to their health state and preference. Though internet has become a useful source for health information and we can find a lot of useful health information about disease, treatment and prevention, it also dissemi-
nates wrong and bad information. In other words, as people are exposed to too much information, they need to re-
ommend systems which help their decision making. Nowadays, studies of recommender systems on healthcare 
service have been actively conducted (Kim et al, 2009).

A lot of researchers studied various approaches such as collaborative filtering, content-based filtering and heuris-
tic method to enhance their prediction accuracy (Bobadilla et al, 2010). These issues are essential to make effective 
recommender systems. In response, this present paper we adopted General Bayesian Network to develop accurate 
infrastructure engine, and examined the performance comparing to other methods such as Naïve Bayesian Net-
work(NBN), and Decision Tree.

The paper is organized as follows: In section 2, we review related works such as health and healthcare and 
Bayesian Network. The process of experiment and details are presented in section 3. The result and discussion are 
presented in section 4. Finally, theoretical and practical implication of the present paper is presented in section 5.

2. Related Works

2.1. Healthcare

In the aging society, generally people consider a lot their health and well-being issues than before. As the medical 
improvement, people’s life expectancy is much longer than before, so people’s main concerns are staying healthy in 
their whole life. In response, the health related industry is growing at a fast pace. Compare to other industry, health 
related industry has peculiar characters. Just imagine that you want to buy a pair of jeans or something else, the best 
choice is buying a pair of jeans which is good quality with lower price. That is, the salient issues are in terms of cost 
and quality. In contrast, they are reluctant to apply same reasoning to their decisions of their health. Healthcare is 
different in that quality and reliability are viewed as necessary primary considerations. In addition, these are espe-
cially important in cases which decisions are seriously related to personal health problem. In this context, healthcare 
services have been framed as credence goods, which means people rely on trusted experts recommendation to make 
decisions (Kay, 2007). Also if they assess the recommendation is reliable one, they is likely to accept the recom-

mendation easily and follow that as they are recommended without reactance (Grandpre et al, 2003).

Comparing the past, people consider psychological stability a lot as well as physical health (Danna & Griffin, 
1999). It means we live in a world of stress and anger which threaten our psychological stability. A lot of research-
ers took into accounts psychological states. Some studied showed the negative effects of stress in workplace. This 
presented that stress made many serious problem including reduced productivity, poor decision making and job 
dissatisfaction in the perspective of organization, and the consequence of stress to the individual include anxiety, 
fatigue and cardiovascular problems (Teasdale, 2006). That is, healthcare service providers consider both physical 
and psychological factors to offer more reliable services and gain profits from providing healthcare services. Take 
these into account the most important thing is making reliable personalized services reflecting various factors. To 
make personalized recommendation inference engine, we consider various factors such as chronic illness, level of 
fatigue, stress and anger in the phase of data collection.

2.2. Bayesian Network and Decision Tree

Bayesian network is a probabilistic graphical model that represents as directed acyclic graph (DAG) based on the 
conditional probability tables (CPTs) identifies the causal relationships in the simplest way. In the Bayesian network, 
node represents variables and arc represents the relationship between the dependent variables. Node A is node B’s 
parent node if arc from node A to node B. The node’s given value is called the evidence node. The one node can be 
any kind of variable such as measured values, arguments, latent variables and hypotheses. Bayesian network could 
represent the probabilistic relationships between natural disaster and sign. Given a certain sign, the network can 
predict the presence of various natural disasters throughout computing the probabilities (Darwiche, 2010). The ap-
lication fields include cognitive science, statistics, computer engineering and medical.
The most widely well-known Bayesian networks include naive Bayesian network (NBN), Tree augmented NBN (TAN), and general Bayesian 0network (GBN). First, NBN consists of a simple probabilistic classification. Meaning according to the probability model as a more accurate portrayal is an independent feature model. Probability model can be derived by using Bayes theorem and strong independence assumption that cannot occur because called 'Naive'. NBN shows a simple structure that has the classification node as the parent node of all other nodes. There are two advantages about NBN. One of the advantages is easy to construct, as the structure is a given priority. And the inference process using NBN is very fast and efficient. Besides, NBN’s inference performance is significantly good with some unrealistic limitations. Hence, TAN was proposed by (Friedman et al, 1997). The class node in TAN also directly points to all attribute nodes same as NBN, but there is no limitation on the arc among attribute nodes (Madden, 2009).

GBN is an unrestricted Bayesian network. In general Bayesian network, all nodes treat as normal node and can have a present node and also be a child node of some attribute node. K2 and Hill Climb (HC) algorithm adopt to generate the GBN. K2 algorithm is simple and fast algorithm, kind of greedy algorithm. It starts with a given ordering of the nodes. Hill Climb algorithm starts from an empty or random network. If there is no information on the conditional probability distribution of the data in the network structure learning is required. An important element in the network structure learning is actual sample data for each event. GBN can have reasoning ability throughout the network structure learning based on the actual sample data. A typical method for structural learning is the score-based learning. This method is to maximize the score according to the degree of matching of generated network and actual data. Of this approach is a numerical score reflects the degree of data and network that matches (Heckerman, 2008). In addition, GBN provides effective analysis methods which are what-if and goal-seeking analysis. In summary, GBN is well reflected in the actual data (Cooper, 1992). (i.e., situation of domain)

ID3 is a typical decision tree approach proposed by (Quinlan, 1986), which is one of the data mining techniques to find a set of decision-making rules from the target dataset made of categorical variables. Decision tree considers chance event outcomes, resource costs, and utility. Decision tree is strong analysis techniques for prediction, tabulate decision-making rules and classify into several subgroups. Decision tree algorithms have different formation process. Therefore, J48 is another version of ID3, which generates a decision tree from dataset represented in continuous variables. J48 is identical with C4.5 (Quinlan, 1993). A decision tree has advantages that can be built easily and is simple to understand and interpret. But, only predefined inference is possible. However, complex reasoning based on decision tree is not possible, which makes it difficult to handle exception case. Therefore, we expected general Bayesian network shows better performance.

3. Experiments

Bayesian Network (BN) is a graphical model that encodes causal relationship among variables. It is also a probabilistic modeling method, so a BN is more reliable technique under conditions of uncertainty. Before constructing inference engine, we collected data from 392 students who took the lecture of management information system. Then, we adopted BN and Decision Tree to construct inference engine using this data, then compared the performance to find the most accurate method.

3.1. Constructing Inference Engines

To develop inference engine, we use WEKA (Waikato environment for Knowledge Analysis) (Hall et al, 2009) which is a well known data mining tool. In this work, the structure of the General Bayesian Network (GBN) was learned K2, and Hill Climber which are classifiers provided in WEKA with setting the maximum number of parent node as five. We also try to construct inference engine using other classifiers such as Naïve Bayesian Network (NBN), Tree augmented Naïve Bayesian Network (TAN) which is an extended version of the NBN, and Decision Tree classifiers.
3.2. Classification Performance comparison

We compared the classification performance of seven different classifiers. Table 1 shows the performance of each classifier and t-test result. The performance in the table is the average after 10 runs of 10 folds-validation on each classifier to obtain its prediction accuracy, and then conducted t-test at the 5% significance level to compare each classifier with the baseline GBN-HC. Since GBN-HC showed better performance in comparison of other Classifiers, GBN-HC was selected as the baseline. As can see in Table 1, the GBN-HC and GBN-K2 outperformed all other classifiers and the performance differences were statistically significant in almost all cases. Though between GBN-HC and GBN-K2, the actual value is different, it is not significant statistically. That is, it makes sense GBN-HC and GBN-K2 is more accurate classifiers than others.

Table 1. Prediction Accuracy of Classifiers

|          | GBN-HC | GBN-K2 | NBN   | TAN   | ID3   | J48   |
|----------|--------|--------|-------|-------|-------|-------|
| Accuracy (%) | 82.53  | 80.83  | 70.73*| 57.28*| 62.82*| 69.33*|

(* means bad performance level comparing a baseline classifier (GBN-HC) at 5% significant level

4. Discussion and concluding remarks

In the current paper, firstly, we collect data for learning to use inference engine. The data include individuals’ demographic characteristics such as age, gender and job, chronic illness and psychological state. We try to reflect various factors to make reliable recommendation engine. We also examined the performance of classifiers comparing GBN and others to find the most accurate classifiers. The result showed that GBN classifiers outperformed comparing to other classifiers. Adopting GBN, the inference engine not only more accurate prediction but also can provide additional information using what-if and goal-seeking analysis techniques.

Acknowledgements

This research was supported by the World Class University (WCU) program through the National Research Foundation of Korea, funded by the ministry of Education, Science and Technology, Republic of Korea (Grant No. R31-2008-000-10062-0).

References

Bobadilla, J., Serradilla, F., Bernal, J. (2010). A new collaborative filtering metric that improves the behavior of recommender systems. Knowledge-Based Systems, 23(6), 520-528.
Cooper, G., & Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. Machine Learning, 9(4), 309-347.
Danna, K., & Griffin, R. W. (1999). Health and well-being in the workplace: A review and synthesis of the literature. Journal of Management, 25(3), 357-384.
Darwiche, A. (2010). Bayesian networks. Communications of the ACM, 53(12), 80-90.
Friedman, N., Geiger, D., Goldszmidt, M.(1997). Bayesian network classifiers, Machine Learning, 29(1), 131-163.
Grandpre, J., Alvaro, E. M., Burgoon, M., Miller, C. H., Hall, J. R. (2003). Adolescent reactance and anti-smoking campaigns: A theoretical approach. Health Communication, 15(3), 349-366.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA data mining software: An update. SIGKDD Exploration Newsletter, 11(1), 10-18.
Heckerman, D. (2008). A tutorial on learning with bayesian networks. D. Holmes, & L. Jain (Eds.), Springer Berlin Heidelberg,156, 33-82
Kay, M. J., (2007). Healthcare marketing: what is salient?. International Journal of Pharmaceutical and Healthcare Marketing, 1(3), 247 - 263
Kim, J.,Lee, J.,Park, J., Lee, Y., Rim.K. (2009). Design of diet recommendation system for healthcare service based on user information. Fourth International Conference on computer sciences and convergence Information Technology(ICCIT), 516-518.
Madden, M.G.(2009). On the classification performance of TAN and general Bayesian networks, Knowledge-Based Systems, 22(7), 489-495
Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
Quinlan, J. R. (1993). C 4.5: Programs for machine learning, Morgan Kaufman Publishers.
Teasdale, E. L. (2006). Workplace stress. Psychiatry, 5(7), 251-254.