An Efficient Attribute Ordering Optimization in Bayesian Networks for Prognostic Modeling of the Metabolic Syndrome

Han-Saem Park and Sung-Bae Cho

Department of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
sammy@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Abstract. The metabolic syndrome has become a significant problem in Asian countries recently due to the change in dietary habit and life style. Bayesian networks provide a robust formalism for probabilistic modeling, so they have been used as a method for prognostic model in medical domain. Since K2 algorithm is influenced by an input order of the attributes, optimization of BN attribute ordering is studied. This paper proposes an evolutionary optimization of attribute ordering in BN to solve this problem using a genetic algorithm with medical knowledge. Experiments have been conducted with the dataset obtained in Yonchon County of Korea, and the proposed model provides better performance than the comparable models.

1 Introduction

The metabolic syndrome is composed of a cluster of metabolic disorders including abdominal obesity, insulin resistance, dyslipedemia and hypertension, and the correlation between metabolic syndrome and coronary heart disease was reported in previous studies [1]. It affects around 25% of adults over the age of 20 and up to 45% over age 50 in the United States [2]. In Asian countries, it has become a significant problem lately due to the change in dietary habit and life style. In situations like this, many groups have been studying the metabolic syndrome from all over the world [1-2].

The Bayesian network has emerged in recent years as a powerful technique for handling uncertainty in complex domains [3]. It is a model of a joint probability distribution over a set of random variables. The Bayesian network is represented as a directed acyclic graph where nodes correspond to variables and arcs correspond to probabilistic dependencies between connected nodes [3]. Bayesian networks have been used for prediction or classification problem in the medical domain and shown high performance. In particular, they have been applied successfully to the modeling of diagnosis and prognosis for diverse diseases [4-6]. There have been many black box tools that classify or predict several diseases, and neural networks are the representative example. Bayesian networks have strengths that they can use the domain knowledge easily and analyze the results compared to them [5]. Even though

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNBI 4115, pp. 381 – 391, 2006.

sometimes they are not better than neural networks in terms of accuracies, Bayesian networks are appropriate methods in the medical domain that require analyzing the results with medical knowledge.

This paper deals with a problem that predicts the metabolic syndrome with the dataset obtained in Yonchon County of Korea. This paper makes a prognostic model using Bayesian network, and has used the K2 algorithm by Cooper and Herskovits in order to learn its structure [7]. Since the result of the K2 algorithm is influenced by an input ordering of the attributes, an optimization of this ordering has been also studied [7]. This paper proposes an efficient optimization method using a genetic algorithm with medical domain knowledge in order to solve this problem. Contrary to the conventional methods, after clustering similar attributes into each group, an ordering of the groups and an ordering of the attributes in each group have been performed in turns. By applying the medical domain knowledge, an efficient and reliable evolution process can be conducted. Subsequently, the experiments using the proposed prognostic model have been done after the structure and parameter learning processes in order to show its usefulness.

2 The Problem: The Metabolic Syndrome

The definition of the metabolic syndrome was provided by the National Cholesterol and Education Program, Adult Treatment Panel III (ATP III). It requires the presence of three or more of the following components [1, 2]:

1) Abdominal obesity
(waist circumference >102 cm in men and >88 cm in women),
2) Hypertriglyceridemia ($\geq$ 150 mg/dL),
3) Low high density lipoprotein (HDL) cholesterol
<40 mg/dL in men and <50 mg/dL in women),
4) High blood pressure (systolic $\geq$ 130 mmHg or diastolic $\geq$ 80 mmHg) and
5) High fasting glucose (>110 mg/dL).

Since this original standard is not appropriate for Asian, we have used modified the definition for Asian of the abdominal obesity (waist circumference >90 cm in men and >80 cm in women) in this paper [8].

3 Prognostic Modeling of the Metabolic Syndrome

Fig. 1 illustrates the flowchart that makes the proposed prognostic model. This process can be divided into four main parts: pre-processing, attribute ordering, BN learning, and prediction processes. For the pre-processing part, medical domain knowledge has been applied, so the prediction model would be more reliable. For the attribute ordering part, a genetic algorithm has been also used in order to optimize the ordering of the model efficiently. The structure and parameter learning processes have been conducted after the attribute ordering, and the whole process is finished with prediction of input samples.

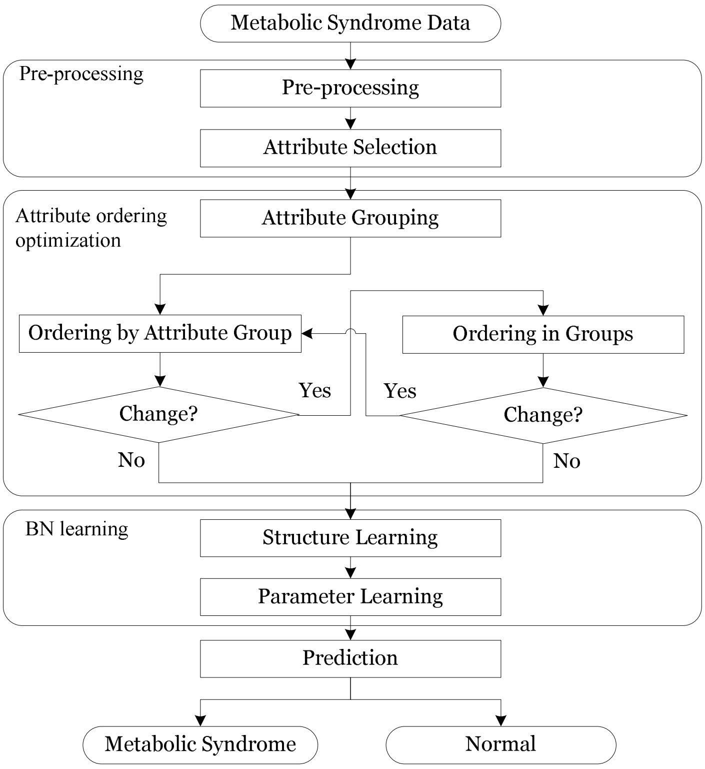

Fig. 1. The overall flowchart of the proposed method

3.1 Data Pre-processing and Attribute Selection

The pre-processing part subdivides into the pre-processing and an attribute selection. Here, we have decided the discrete states of the attributes since most attributes have continuous values and BN requires discrete states. The medical literatures [1, 2, 8] and experts have given a help for this process.

Table 1 demonstrates the attributes and their possible states. Explaining a few important attributes here, fasting glucose is divided into normal (<110), impaired glucose metabolism (110~125) and diabetes (>125). Triglyceride is increased if its value is larger than 150, or normal. In case of HDL cholesterol, it is decreased (abnormal) when it is smaller than standard value, 40 in men and 50 in women. The standards for state discrimination used here are significant since they have been used by medical experts [8].

We have chosen the informative attributes for the prognostic model after the pre-processing. As described in section 2, basic attributes required for defining metabolic syndrome are eight. However, there are more attributes that influence the metabolic syndrome such as age, 2-hour postprandial glucose, hypertension and body mass index [9]. Among these four attributes, the information of hypertension is redundant with the information of basic attributes. Therefore, we have proposed the Bayesian network model using 11 attributes, which include eight basic attributes and additional three informative attributes of age, 2-hour postprandial glucose and body mass index for the prognostic model of the metabolic syndrome. It is well-known that age is closely related to the metabolic syndrome and 2-hour postprandial glucose influence fasting glucose of an important attribute [9, 10]. Body mass index is selected because it is related to obesity, which is related to the metabolic syndrome.

Table 1. Attributes and the possible states

Attribute	Possible states
Age	young, middle aged, old
Sex	male, female
Fasting glucose	normal, impaired glucose metabolism, diabetes
2-hour postprandial glucose	normal, impaired glucose metabolism, diabetes
Waist circumference	normal, abdominal obesity
Triglyceride	normal, increased
HDL cholesterol	decreased, normal
Body mass index	low weight, normal, over weight, obesity
Ratio of waist-hip circumference	normal, abdominal obesity
Systolic blood pressure	normal, hypertension
Diastolic blood pressure	normal, hypertension

3.2 Evolutionary Optimization of Attribute Ordering

For the structure learning process of the Bayesian network, we have used the K2 algorithm, which is subject to an ordering of attributes when they are input. Therefore, an attribute ordering optimization has been studied to make Bayesian network more accurate. Larranaga *et al.* regarded this ordering problem as an ordering of cities in the TSP (Traveling Salesman Problem), and applied several genetic algorithms that had been used for the TSP [7]. We have also used the genetic algorithm to optimize attribute ordering and applied medical domain knowledge to make a model more efficiently.

1) Optimizing Attribute Group Ordering

As illustrated in attribute ordering part of Fig. 1, grouping of similar attributes is performed first. It is so complicated that we cannot simplify the causal relationship among attributes, but it is known that the attributes related to obesity are expressed first and ones related to metabolic disorder follow generally [10]. We have grouped twelve attributes, 11 attributes and a label, based on this information, so three attributes (waist circumference, body mass index, and ratio of waist-hip circumference) related to obesity and four attributes (fasting glucose, 2-hour postprandial glucose, triglyceride, and HDL cholesterol) related to metabolic disorder are grouped, respectively. The remaining five attributes are treated as five groups with a single attribute, respectively.

Subsequently, an ordering of these seven attribute groups is optimized. Ordering of attribute groups is optimized first, and then ordering of attributes in each group is optimized. These two processes are repeated until there is no change of ordering after each optimization step.

2) Application of the Genetic Algorithm

Details of chromosome representation, selection, and genetic operations are as follows.

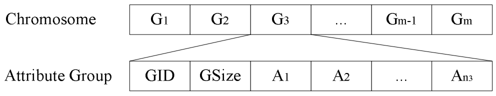

Fig. 2. Chromosome representation

As shown in Fig. 2, supposing each chromosome consists of *m* attribute groups, each group has group ID, the number of attributes in each group, and attributes themselves in each group, general genetic operations can be conducted. Attribute ordering of initial chromosome is decided randomly, and the fitness of each individual is evaluated in terms of the predictive accuracy because this paper focuses on the prognostic modeling.

After evaluating the fitness of individuals, individuals for the next generation are selected using rank-based selection method. In (1), $I_{(g,j)}$ means the *j*th individual in the *g*th generation, and Rank($f(I_{(g,j)})$) means the rank of each individual based on the fitness. n represents the number of individuals, and $p_{(g,j)}$, the probability that each individual $I_{(g,\mathrm{j})}$ is selected, can be provided as shown in (1).

$$p_{(g,j)} = \frac{n - Rank(f(I_{(g,j)})) + 1}{n(n+1)/2} \tag{1}$$

Crossover and mutation operations follow the selection process. Larranaga *et al.* compared several crossover operators on the TSP, and cycle crossover (CX) operator was the best [9]. This operator provides high performance even though the number of individuals is small. It attempts to create an offspring from the parents where every position is occupied by a corresponding element from one of the parents [7]. Fig. 3 illustrates an example of CX operation. There are two cycles in each individual, and they are crossed over each other.

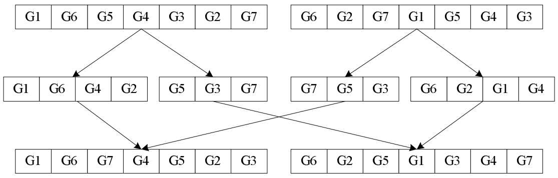

Fig. 3. An example of CX operation

Displacement mutation (DM) operator has been used for mutation operation. This operator provided the best performance when it was used with CX operator in Larranaga's work [7]. First, it selects a random substring and removes it, and it inserts this substring into random position. Fig. 4 shows an example of DM operation.

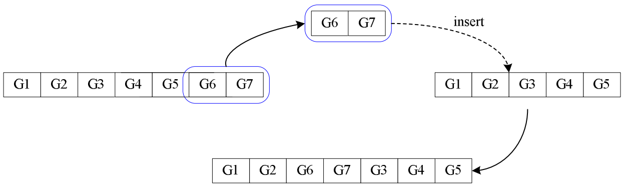

Fig. 4. An example of DM operation

```
The K2 Algorithm
Input:
    A set of $n$ nodes,
    An ordering on the nodes,
    An upper bound $u$ on the number of parents a node can have,
    A dataset $D$ including $m$ cases.
Output:
    A printout of each node and its parents.

for ( i=1; i<n; i++ ) {
$\Pi_i$ = 0;
$P_{old}$ = g($i$, $\Pi_i$);
OK2Proceed = TRUE;
    while ( OK2Proceed &&$|\Pi_i|<u$ ) {
        Let Z be the node in Pred($X_i$)-$\Pi_i$ that maximizes g($i$, $\Pi_i \cup \{Z\}$);
        $P_{new}$= g($i$, $\Pi_i \cup \{Z\}$);
        if ($P_{new} > P_{old}$) {
            $P_{old} = P_{new}$;
            $\Pi_i = \Pi_i \cup \{Z\}$;
        }
        OK2Proceed = FALSE;
    }
    print("node: ", $X_i$, "parents: ", $\Pi_i$);
}
```

Fig. 5. The K2 algorithm

3) Structure and Parameter Learning

With the evolved ordering decided before, we train the Bayesian network using the K2 algorithm [13]. The K2 algorithm narrows down the search space by fixing an order of the attributes. The probabilities of nodes have been calculated from the frequency counts of the learning data. Fig. 5 provides the pseudo-code of the K2 algorithm.

Given dataset *D*, the K2 algorithm searches the set of parent nodes that maximizes P(B_S , *D*) for every node. K2 starts by assuming that nodes do not have parents. For every step, it incrementally adds specific parent whose addition increases the probability of the result structure most. These processes continue until the addition of a single parent cannot increase the probability. K2 is a greedy heuristic, and it does not guarantee the structure with the highest probability. The details are provided in [13].

Once the structure of Bayesian network is decided, parameter values of each node are calculated from the frequency of learning data, and then the prognostic model of the metabolic syndrome is completed.

4 Experimental Results

4.1 Experimental Data

The dataset used in this paper was examined for epidemiological research of the local community. The surveys were conducted twice in 1993 and 1995 in Yonchon County of Korea. 2,293 subjects were participated in the first survey, and 1,193 of them were participated in the second survey [12]. We have used the data of 1,135 subjects who did not include missing values and the 18 attributes that could influence the prediction of the metabolic syndrome. After that we have decided whether each sample would belong to the metabolic syndrome or not. 18 attributes and the distribution by states are shown in Table 2.

4.2 Parameters and Settings

In order to compare the result of Bayesian network model with the other models, neural network (NN) and *k*-nearest neighbors (kNN) prediction models have been used. 11 input nodes, 20 hidden nodes, and 2 output nodes are used for NN, and *k* of kNN is fixed as 3 having shown the best performance in a preliminary experiment.

For the genetic algorithm, population size is set as 20, and the process has evolved by the generation of 100. The population size is small because the total search space is not so large. The number of possible combinations is 7!, but the time cost is expensive using an exhaustive method since the learning and inference processes are required for every individual of every generation. The selection rate of 0.8 and mutation rate of 0.02 are used. Crossover rate is set as 1.0, but it is less than that because the result of the CX operation sometimes is not changed.

For experiments before evolution, 10-fold cross validation is repeated 30 times, and the average is used for the result. For evolution process, we have divided the data into three parts with the ratio of 3:1:1. The first part is used for the learning, the second one is used for the validation, and the last one is used for the test.

Table 2. The attributes of tfhe dataset and the distribution by states

Attributes	n	Percentage
Age		
Young	111	9.78
Middle aged	512	45.11
Old	512	45.11
Sex		
Female	646	56.92
Male	489	43.08
Fasting glucose		
Normal	942	82.99
Impaired glucose metabolism	166	14.63
Diabetes	27	2.38
2-hour postprandial glucose		
Normal	985	86.78
Impaired glucose metabolism	150	13.22
Diabetes	0	0.00
Waist circumference		
Normal	456	40.18
Abdominal obesity	679	59.82
Triglyceride		
Normal	731	64.41
Increased	404	35.59
HDL cholesterol		
Decreased	873	76.92
Normal	262	23.08
Body mass index		
Low weight	36	3.17
Normal	405	35.68
Over weight	267	23.52
Obesity	427	37.62
Ratio of waist-hip circumference		
Normal	771	67.93
Abdominal obesity	364	32.07
Systolic blood pressure		
Normal	838	73.83
Hypertension	297	26.17
Diastolic blood pressure		
Normal	773	68.11
Hypertension	362	31.89

4.3 Experimental Results

First, we have conducted comparison experiments among Bayesian network models that have different numbers of attributes. The first model is BN with eight basic attributes, the second one is the proposed model with 11 selected attributes, and the last one is the BN model with all attributes. Table 3 indicates that the result of 11 attributes is the best.

Table 3. The comparison of accuracies by the number of attributes

Attribute size	8	**11**	18
Accuracy (%)	70.74 (± 0.0017)	**72.15 (± 0.0082)**	70.82 (± 0.0079)

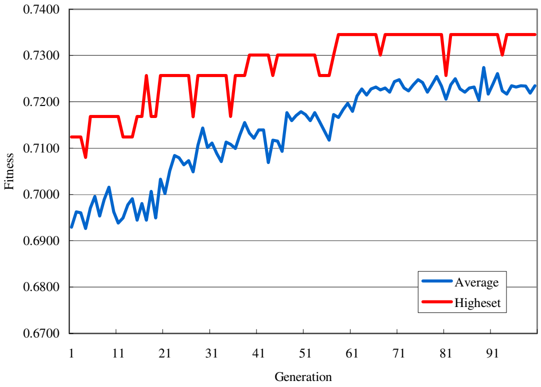

Fig. 6. The fitness transition graph during an evolution process

Table 4. The comparison of accuracies before and after optimization

Ordering	Accuracy (%)
Optimized	**72.12 (± 0)**
Random	70.09 (± 0.0513)

Table 5. The comparison of accuracies by the number of attributes

Prediction model	Accuracy (%)
BN	**72.12 (± 0)**
NN	63.19 (± 0.0152)
3NN	62.56 (± 0)

We have applied the proposed method to total 1,135 data, and Fig. 6 illustrates the evolution process. It evolves well since average and the highest fitness values get a little higher accordingly as the generation grows. The highest fitness converges after the 60th generation, and the average fitness converges after a few generations.

After that, we have compared the model of optimized ordering with the model of random ordering. Table 4 provides the comparison result in terms of accuracy. Here, the difference between two models is statistically significant (p<0.001).

Finally, we have compared the BN model after attribute optimization with two other models of neural networks and k nearest neighbors that have been frequently used in pattern recognition field in order to solve the prediction problem. Table 5 shows the comparison result with accuracy, and the differences of the proposed model and other models (BN vs. NN, BN vs. 3NN) are statistically significant (p<0.001). Generally, NN has strength in accuracy though it cannot be interpreted easily. However, BN model has provided better performance, and we can guess that the pre-processing and attribute selection processes using medical domain knowledge were effective.

5 Conclusions

This paper proposed a Bayesian network model with evolutionary algorithm in order to predict the metabolic syndrome. In processes of building the prognostic model, we applied the medical domain knowledge in order to make the model more reliable. We also adopted the genetic algorithm to optimize attribute ordering, and completed the model efficiently using the medical domain knowledge in the optimization process.

We verified that the proposed method provided better performance compared with the model before ordering optimization as well as other models such as neural networks and *k*-nearest neighbor.

Acknowledgments. This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. Mykkanen, L., Kuusisto, J., Pyorala, K., Laakso, M.: Cardiovascular Disease Risk Factors as Predictors of Type 2 (Non-Insulin-Dependent) Diabetes Mellitus in Elderly Subjects. Diabetologia, 50 (2004) 453-469
2. Mehta, N. N., Reilly, M. P.: Mechanisms of the Metabolic Syndrome. Drug Discov Today, 1(2) (2004) 187-194
3. Lee, S. M., Abbott, P. A.: Bayesian Networks for Knowledge Discovery in Large Datasets: Basics for Nurse Researchers. J Biomed Inform, 36 (2003) 389-399
4. Antal, P., Fannes, G., Timmerman, D., Moreau, Y., Moor, B. D.: Using Literature and Data to Learn Bayesian Networks as Clinical Models of Ovarian Tumors. Artif Intell Med, 30 (2004) 257-281
5. Wang, X.-H., Zheng, B., Good, W. F., King, J. L., Chang, Y.-H.: Computer-Assisted Diagnosis of Breast Cancer Using a Data-driven Bayesian Belief Network. Int J Med Inform, 54 (1999) 115-126
6. Sierra, B.: Using Bayesian Networks in the Construction of A Bi-level Multi-Classifier. A Case Study Using Intensive Care Unit Patients Data. Artif Intell Med, 22 (2001) 233-248
7. Larranaga, P., Kuijpers, C. M. H., Murga, R. H., Yurramendi, Y.: Learning Bayesian Network Structures by Searching for the Best Ordering with Genetic Algorithms. IEEE T Syst Man Cy A, 26 (4) (1996)
8. Moon, M. K., Cho, Y. M., Lim, K. S., Park S., Lee, H. K.: Metabolic Syndrome. The Korean Society of Endocrinology, 18 (2003) 105-116
9. Lindblad, U., Langer, R. D., Wingard, D. L., Thomas, R. G., Barrett-Connor, E. L.: Metabolic Syndrome and Ischemic Heart Disease in Elderly Men and Women. Am J Epidemiol, 153 (2001) 481-489

10. Girod, J. P., Brotman, D. J.: The Metabolic Syndrome as a Vicious Cycle: Does Obesity Beget Obesity. Med Hypotheses, 60(4) (2003) 584-589
11. Cooper, G. F., Herskovits, E. A.: A Bayesian Method for Induction of Probabilistic Networks from Data. Mach Learn, 9(4) (1992) 309-347
12. Park, Y.: Prevalence of Diabetes and IGT in Yonchon County, South Korea Diabetes Care, 18 (1995) 545-548